National Research University Higher School of Economics

as a manuscript

Dean Fantazzini

Modelling and forecasting univariate and multivariate time series using Google data and copulae

Dissertation
for the purpose of obtaining academic degree
Doctor of Science in Economics

Moscow - 2020

Федеральное государственное автономное образовательное
учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

На правах рукописи

Деан Фантаццини

Моделирование и прогнозирование одномерных и многомерных временных рядов с использованием данных Google и копул

ДИССЕРТАЦИЯ
на соискание ученой степени
доктора экономических наук

Москва – 2020

Contents

1 Introduction **3**
1.1 Google Trends 4
1.2 Copula models 5

2 Main results: short summary **7**

References **11**

Appendices

Appendix 1 "Copulae and operational risks."

Appendix 2 "The effects of misspecified marginals and copulas on computing the value at risk: A Monte Carlo study."

Appendix 3 "Three-stage semi-parametric estimation of t-copulas: Asymptotics, finite-sample properties and computational aspects."

Appendix 4 "Modelling and forecasting the global financial crisis: Initial findings using heterosckedastic log-periodic models."

Appendix 5 "Forecasting the global financial crisis in the years 2009-2010: Ex-post analysis."

Appendix 6 "Analysis of multidimensional probability distributions with copula functions - Part I."

Appendix 7 "Analysis of multidimensional probability distributions with copula functions - Part II."

Appendix 8 "Analysis of multidimensional probability distributions with copula functions - Part III."

Appendix 9 "Everything you always wanted to know about log-periodic power laws for bubble modeling but were afraid to ask."

Appendix 10 "Nowcasting and Forecasting the Monthly Food Stamps Data in the US Using Online Search Data."

Appendix 11 "Proposed coal power plants and coal-to-liquids plants in the US: Which ones survive and why?"

Appendix 12 "Forecasting German car sales using Google data and multivariate models."

Appendix 13 "The oil price crash in 2014/15: Was there a (negative) financial bubble?"

Appendix 14 "Everything you always wanted to know about bitcoin modelling but were afraid to ask - Part 1."

Appendix 15 "Everything you always wanted to know about bitcoin modelling but were afraid to ask - Part 2."

Appendix 16 "Big Data for computing social well-being indices of the Russian population."

Appendix 17 "The importance of being informed: forecasting market risk measures for the Russian RTS index future using online data and implied volatility over two decades."

Appendix 18 "A multivariate approach for the simultaneous modelling of market risk and credit risk for cryptocurrencies."

1 Introduction

The quick development of the internet and information technology (IT) worldwide has given access to a large amount of data, which are usually known as "big data". Even though there is not an unambiguous definition, big data can be described along these three dimensions originally introduced by Laney (2001): *volume* (the size of the data), *variety* (the type of the data) and *velocity* (the speed at which the data is created and stored). Gantz and Reinsel (2011) put forward *value* (the utility of the data) as the fourth dimension. In this regard, the process to extract valuable information from this data is now widely known as 'big data analytics'. Additional dimensions have also been proposed more recently: *veracity* (the data quality), *variability* (the value and characteristics of the data depends on the context where they are created), *scaleability*, *exhaustivity* (the entire system is recorded or not), etc, see Jin et al. (2015), Wamba et al. (2015), Gandomi and Haider (2015) and Kitchin and McArdle (2016) for more details. Big data have stimulated the development of a large field of the economic and financial literature, and they contributed to a better understanding of contemporary economic phenomenons, see the surveys by Edelman (2012), Varian (2014), Blazquez and Domenech (2018), Li et al. (2018) and references therein for a thorough discussion.

One of the main tools that can be used to analyze big data is a search engine, which is often considered the first step in the consumer decision-making process, see e.g. Xiang and Fesenmaier (2005) and Kaushik (2009). Moreover, search engines can be used to understand social dynamics and to make better predictions. In this regard, Google is the search engine with the largest market share worldwide (90% in 2018) and the analysis of its search data has been one the most important and well-known examples of the use of big data. In 2006, it launched a tool named Google Trends which shows how frequently a particular keyword or a topic are searched online in a specific region, at a specific period of time, and also in different languages. The famous paper by Ginsberg et al. (2009) published in the journal *Nature* showed that Google Trends can help to forecast the spread of influenza earlier than the Centers for Disease Control and Prevention (CDC): this seminal work spurred a hot debate and contributed to the increasing use of Google Trends in research studies of different academic fields, including IT, communications, medicine, health, business and economics, see the large survey by Jun et al. (2018) for a detailed review.

The advent of these large datasets further stimulated the interest in developing multivariate models able to consider departures from the assumption of normality and to be computationally tractable: copula models can deal with both these two issues. The theory of copulas dates back to Hoeffding (1940) and Sklar (1959), but its large scale use in empirical applications is far more recent and dates back to the first decade of the 21st century. In simple terms, a copula is a multivariate function that models the dependence structure between the variables of a random vector. When a copula is applied to marginal distributions which may not necessarily belong to the same distribution family, it delivers a proper multivariate distribution. Therefore, copulas allow for a flexible modelling of the dependence structure between different variables, as well as for the possibility to have different marginal distributions. Moreover, the separation of the dependence structure from the marginals strongly decreases the computational burden of estimating a multivariate model.

The 20 publications that constitute my dissertation investigated several cases where it is possible to model and forecast economic and financial time series using Google data

and/or copula models. Before discussing in sections 2 and 3 the main results of my publications, I will briefly discuss how Google Trends work and the theory of copulas to make this summary self-contained. The detailed bibliographic information of the publications included in the dissertation is reported in Appendix A.

1.1 Google Trends

Google Trends is a website (https://trends.google.com) that shows the standardized volume of Google searches for a keyword or a topic. The data can be filtered according to search type (for example, web, videos, images), search category (there are 25 categories and 288 subcategories), geographic location, and time range. The amount of searches is divided by the total amount of searches for the same period and region, and the resulting time series is divided by its highest value and multiplied by 100. Google Trends data are computed using a sampling method, so the results can slightly differ if the data are downloaded on different days. Moreover, only queries with a minimum volume are tracked due to privacy considerations: if the search volume is insufficient, a value of zero is reported. The data are available from an intraday time-frequency up to a monthly frequency, depending on the selected time range. Google Trends allows comparing the search volumes of up to five search terms, or up to a maximum of 30 search terms grouped in a single entry using quotation marks (to return searches that match an exact expression), and using the "+" or "-" signs between the search terms to include or exclude search terms, respectively. The data are available since 2004, for several countries and their regions, see https://support.google.com/trends for more details. An example of Google Trends data with the top five car manufacturers on the Russian market according to AEB (Association of European Businesses) sales in 2016 is reported in Figure 1.

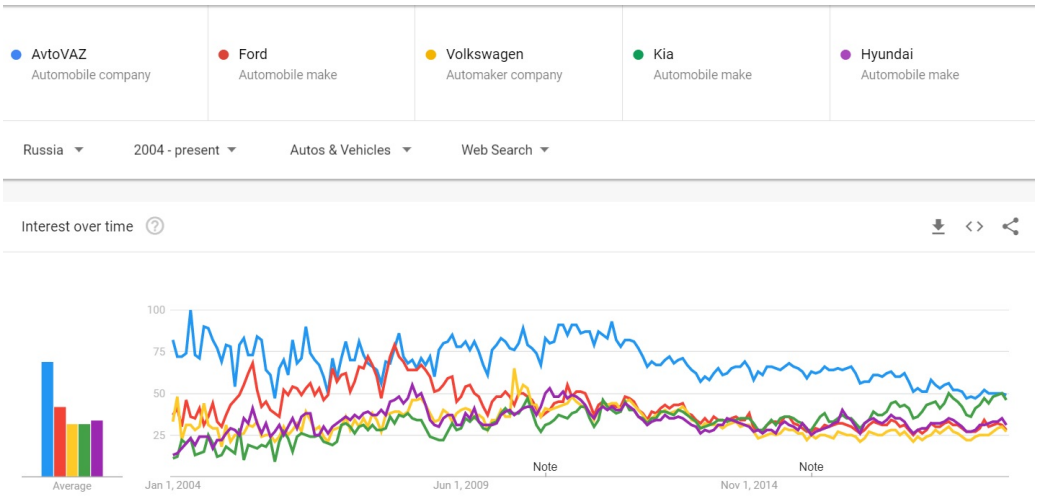


Figure 1: Google Trends search data for Russia - Autos and Vehicles category.

Google Trends has become an important source for big data research because it provides an easy and excellent platform for observing consumers' information seeking activities, and it also supplies several tools to analyze the historical search data (Jun et al. (2014)). In this regard, there is a large theoretical literature that shows that the process of searching for information allows consumers to reduce the risks associated with the purchasing process, and for that reason online searches leads to product purchasing, see Shim et al. (2001) and

To et al. (2007). Moreover, data on search queries can be used to monitor, analyze, and to predict the level of acceptance for a new product or a new technology by consumers, see Kotler and Keller (2008), Ettredge et al. (2005), Zimmer et al. (2007) and Fenn and Raskino (2008).

Since the introduction of Google Trends in 2006, several researchers in the economic and financial fields have begun to use Google search data to produce real-time forecasts when information from official sources is released with a lag (the so-called 'nowcasting'), or simply for forecasting purposes, see Askitas and Zimmermann (2009), Goel et al. (2010), Engelberg and Gao (2011), Choi and Varian (2012), D'Amuri and Marcucci (2017), Dinis et al. (2019) just to name a few. More specifically, applications range from macroeconomic forecasting (unemployment rate, national consumption, inflation rates, etc), to financial forecasting (stock prices, real estate pricing, portfolio management), to microeconomic forecasting (box-office revenues, rank of songs on the Billboard Hot 100 chart, retail sales, travel, etc), see Jun et al. (2018) for a review. There were also attempts to use Google Trends in the political field, but in this case, the results are more mixed, ranging from outstanding forecasting power in regards to the 2015 Greek Referendum results by analyzing data from Google Trends on the 'YES' and 'NO' search terms (Mavragani and Tsagarakis (2016)), to disappointing results when forecasting the victory of a candidate in US Congress elections in 2008 and 2010 (Lui et al. (2011). The latter results can be explained by the fact the Google Trends data can include both positive and negative interest, so that an increase in online searches may not be necessarily associated with an increasing positive attitude towards a candidate. In the case of sociology and political studies, other social big data should be considered and sentiment analysis must be used to disentangle positive attitudes from negative attitudes.

1.2 Copula models

The study of copulas has started with the seminal papers by Hoeffding (1940) and Sklar (1959) and has seen various applications in statistics and financial literature. For more details, I refer the interested reader to the methodological overviews by Joe (1997) and Nelsen (1999), while Cherubini et al. (2004), McNeil et al. (2015) and Fantazzini (2019) provide detailed discussions of copula techniques for financial applications.

In simple terms, an n-dimensional copula is a multivariate cumulative distribution function with uniform distributed margins in [0,1]. More formally, if we consider $X_1, \ldots X_n$ to be random variables, and H their joint distribution function, then a copula can be defined as a multivariate distribution function H of random variables $X_1 \ldots X_n$ with standard uniform marginal distributions $F_1, \ldots, Fn$, defined on the unit n-cube $[0,1]^n$, with the following properties:

1. The range of $C(u_1, u_2, ..., u_n)$ is the unit interval [0,1];
2. $C(u_1, u_2, ..., u_n) = 0$ if any $u_i = 0$, for $i = 1, 2, \ldots, n$.
3. $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$, for all $u_i \in [0, 1]$

The previous three conditions provides the lower bound on the distribution function and ensures that the marginal distributions are uniform. The Sklar's theorem justifies the role of copulas as dependence functions.

Theorem 1.1 (Sklar's theorem)**.** *Let H denote a n-dimensional distribution function with margins $F_1 \dots F_n$. Then there exists a n-copula C such that for all real $(x_1, \dots, x_n)$*

$$H(x_1, \dots, x_n) = C(F(x_1), \dots, F(x_n)) \tag{1}$$

If all the margins are continuous, then the copula is unique; otherwise C is uniquely determined on $RanF_1 \times RanF_2 \dots RanF_n$, where Ran is the range of the marginals. Conversely, if C is a copula and $F_1, \dots F_n$ are distribution functions, then the function H defined in (1) is a joint distribution function with margins $F_1, \dots, F_n$.

Proof: See Sklar (1959), Joe (1997) or Nelsen (1999). ■

The last statement is the most interesting for multivariate density modelling, because it implies that we may link together any $n \geq 2$ univariate distributions, of any type (not necessarily from the same family), with any copula to get a valid multivariate distribution. Copulas allow to separate the distribution of a random vector into individual components (the marginals) with a dependence structure (the copula) among them, thus greatly simplifying the multivariate model estimation.

If we use the Sklar's theorem and the relation between the distribution and the density functions, we can derive the multivariate copula density $c(F_1(x_1), \dots, F_n(x_n))$, associated to a copula function $C(F_1(x_1), \dots, F_n(x_n))$:

$$f(x_1, ..., x_n) = \frac{\partial^n \left[C(F_1(x_1), \dots, F_n(x_n))\right]}{\partial F_1(x_1), \dots, \partial F_n(x_n)} \cdot \prod_{i=1}^{n} f_i(x_i) = c(F_1(x_1), \dots, F_n(x_n)) \cdot \prod_{i=1}^{n} f_i(x_i) \tag{2}$$

To fully appreciate the potential of copulas when modelling multivariate time series, I briefly present below a copula-VAR-GARCH model, see Bauwens et al. (2012) and Fantazzini (2019) for a discussion at the textbook level. Let $\mathbf{Y}_t$ be a vector stochastic process of dimension $n \times 1$, then a conditional model for $\mathbf{Y}_t$ can be expressed as follows:

$$\mathbf{Y}_t = \boldsymbol{\mu}_t + \mathbf{D}_t \mathbf{z}_t$$

where $\boldsymbol{\mu}_t$ is a vector of conditional means, $\mathbf{D}_t = diag(\sigma_{1,t}, \dots, \sigma_{n,t})$ a diagonal matrix of conditional standard deviations, while $\mathbf{z}_t$ is a vector of standardized errors with a conditional multivariate distribution $H_t(z_{1,t}, \dots, z_{n,t}; \boldsymbol{\theta})$ and parameter vector $\boldsymbol{\theta}$. The conditional means are modelled with a VAR(p) model,

$$\boldsymbol{\mu}_t = \mathbf{a}_0 + \sum_{m=1}^{p} \mathbf{A}_m \mathbf{Y}_{t-m}$$

while the conditional variances with GARCH(p,q) models,

$$\sigma_{i,t}^2 = \omega_i + \sum_{m=1}^{p} \alpha_{i,m} (\sigma_{i,t-m} z_{i,t-m})^2 + \sum_{k=1}^{q} \beta_{i,k} \sigma_{i,t-k}^2$$

Other univariate GARCH models, like the Exponential-GARCH, the Threshold-GARCH, etc. can be used. If we use copula theory, the joint distribution H_t of the vector of standardized errors $\mathbf{z}_t$ can be expressed as follows:

$$\mathbf{z}_t \sim H_t(z_{1,t}, \dots, z_{n,t}; \boldsymbol{\theta}) \equiv C_t(F_{1,t}(z_1; \delta_1), \dots, F_{n,t}(z_{n,t}; \delta_n); \boldsymbol{\gamma}) \tag{3}$$

which means that the joint distribution H_t of $\mathbf{z}_t$ is the copula $C_t(\cdot;\boldsymbol{\gamma})$ of the cumulative distribution functions of the innovation marginals $F_{1,t}(z_1;\delta_1),\ldots,\ F_{n,t}(z_{n,t},;\delta_n)$, where $\boldsymbol{\gamma},\delta_1,\ldots,\delta_n$ are the copula and marginal parameters, respectively. It follows from (2)-(3) that the log-likelihood function for the joint conditional distribution $H_t(\cdot,\boldsymbol{\theta})$ is given by

$$l(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log\Big(c\big(F_{1,t}(z_1;\delta_1),\ldots,F_{n,t}(z_{n,t};\delta_n);\boldsymbol{\gamma}\big)\Big) + \sum_{t=1}^{T}\sum_{i=1}^{n} \log f_i(z_{i,t};\delta_i)$$

where c is the copula density function, whereas f_i are the marginal densities. The maximization of the previous log-likelihood with respect to the parameters $(\boldsymbol{\gamma},\delta_1,\ldots,\delta_n)$ can be made in 1 step, or in several steps by partitioning of the parameter vector into separate parameters for each margin and the parameters for the copula. This multi-step procedure is known as the method of Inference Functions for Margins (IFM), and it greatly simplify the estimation process.

It is straightforward to show that the famous multivariate normally distributed DCC model by Engle (2002) can be represented as a special case within a more general copula framework,

$$\begin{aligned}\mathbf{Y}_t &= \boldsymbol{\mu}_t + \mathbf{D}_t\mathbf{z}_t, \quad \text{where} \quad \mathbf{z}_t \sim H_t(z_{1,t},\ldots,z_{n,t};\boldsymbol{\theta}), \quad \text{and}\\ \mathbf{z}_t &\sim H_t \equiv C_t^{Normal}(F_{1,t}^{Normal}(z_1;\delta_1),\ldots,F_{n,t}^{Normal}(z_{n,t};\delta_n);\mathbf{R}_t)\end{aligned}$$

However, the copula approach allows us to consider more general cases than a multivariate normal DCC model. For example, if we consider skewed-t distributions for the marginals, then a multivariate model allowing for marginal skewness, kurtosis and normal dependence can be expressed as follows:

$$\begin{aligned}\mathbf{Y}_t &= \boldsymbol{\mu}_t + \mathbf{D}_t\mathbf{z}_t\\ \mathbf{z}_t &\sim H_t \equiv C_t^{Normal}(F_{1,t}^{Skewed-t}(z_{1,t};\delta_1),\ldots,F_{n,t}^{Skewed-t}(z_{n,t};\delta_n);\mathbf{R}_t)\end{aligned}$$

where $F_{i,t}^{Skewed-t}$ is the cumulative distribution function of the marginal Skewed-t, and $\mathbf{R}_t$ can be made constant or time-varying, as in the constant conditional correlation (CCC) model or in the DCC model, respectively. If we suppose that our assets have symmetric tail dependence, we can use a Student's t copula, instead,

$$\begin{aligned}\mathbf{Y}_t &= \boldsymbol{\mu}_t + \mathbf{D}_t\mathbf{z}_t\\ \mathbf{z}_t &\sim H_t \equiv C_t^{Student'sT}(F_{1,t}^{Skewed-t}(z_1;\delta_1),\ldots,F_{n,t}^{Skewed-t}(z_{n,t};\delta_n);\mathbf{R}_t,\nu)\end{aligned}$$

where ν denotes the degrees of freedom of the t-copula.

2 Main results: short summary

This section contains a brief overview of the results of my dissertation. The main purpose is to place these results into a historical context so that the reader can understand the full picture, while more details are provided in section 3.

My research work originally started with copula methods for finance and copula theory: Fantazzini et al. (2008) were the first to propose copulas for modeling high dimensional operational risks in a more flexible way able to takes (partial) dependence into account. This is currently one of the main approaches to modeling and measuring operational risks: the Basel Committee on Banking Supervision (2011) -paragraph 230- found that 43% of

the surveyed banks used copulas to model the dependence structure across operational risks.

The effect on the estimation of the Value at Risk when dealing with multivariate portfolios when there is a misspecification both in the marginals and in the copulas were investigated for the first time in Fantazzini (2009), while the asymptotic properties of the three-stage semi-parametric estimation of T-copulas were developed in Fantazzini (2010b), together with its finite-sample behavior, which was examined via simulations. One of the consequences of my works with copulas was the publication of one of the largest reviews of copula theory and methods in Fantazzini (2011a,b,c)[1].

After the publication of the famous paper by Ginsberg et al. (2009) in the journal Nature about Google Trends, I decided to investigate whether this type of big data could be useful for economic and financial forecasting. My first use of Google Trends was for financial bubbles modelling, which was a topic I was working with at that time (see Fantazzini (2010a, 2011d)) and culminated in the work published in Geraskin and Fantazzini (2013), which is still today one of the most downloaded papers of the European Journal of Finance published in 2013[2].

The articles published after Geraskin and Fantazzini (2013) examined the role of Google Trends in modelling and forecasting several economic and financial variables, considering both univariate and multivariate models: Fantazzini (2014) proposed to use Google search data for nowcasting and forecasting the monthly number of food stamps recipients because the administrative burden for enrolling and remaining enrolled in the food stamps program is nontrivial, and searching the web for information is one of the main strategies a potential applicant can do. Fantazzini and Maggi (2015) analyzed the main determinants (including Google data) that influenced the decision to abandon or to proceed with a coal project using a dataset of 145 coal-power plants projects and 25 CTL plants (between 2004 and 2013) and a large set of binary models, while Fantazzini and Toktamysova (2015) proposed a set of models for the long-term forecasting of car sales in Germany, which consider both economic variables and online search queries. Fantazzini (2016) suggested that there was a negative financial bubble in oil prices in 2014/15, which decreased them beyond the level justified by economic fundamentals, and he employed two sets of bubble detection strategies to corroborate this proposition (one of them employed Google data).

Fantazzini et al. (2018) showed that it is possible to use Google data to explain and predict the dynamics of the Russian social well-being indices computed by VTsIOM, while Fantazzini and Shangina (2019) investigated the predictive power of online search activity and implied volatility from option prices for forecasting the realized volatility and the Value-at-Risk at multiple confidence levels for the Russian RTS index future: forecasting risk measures for Russian assets turned out to be the only case (among all my works) where Google Trends did not significantly improve the forecasting performances of the models involved.

Fantazzini et al. (2016) and Fantazzini et al. (2017) reviewed the econometric and mathematical tools which have been proposed so far to model the bitcoin price and several related issues: these reviews paved the way to the article by Fantazzini and Zimin (2020), who proposed a set of models (employing copulas and Google data) which can be used to estimate the market risk for a portfolio of crypto-currencies, and simultaneously to

[1]Many publications of mine dealing with copulas were not included in this dissertation, see https://scholar.google.com/citations?user=M1UMUp4AAAAJ&hl=en for the full list.

[2]The three papers published as Fantazzini (2010a, 2011d) and Geraskin and Fantazzini (2013) were developed jointly, but the latter was published a couple of years later due to a longer review process and a publication queue.

estimate also their credit risk. Moreover, these research efforts resulted in the publication of a 600-page monograph with Amazon KDP titled *Quantitative Finance with R and Cryptocurrencies* (see Fantazzini (2019)), which contains a very detailed discussion of copula methods and Google data for modelling and forecasting cryptocurrencies. This textbook was officially presented at the Central Bank of Russia on the 22/10/2019.

According to Google Scholar, at the end of 2019, the articles composing my dissertation were cited nearly 400 times. The detailed bibliographic information of the 20 publications included in my dissertation is reported below:

Detailed bibliographic information of the publications included in the dissertation

⇒ **Note: the quartile of the journal was retrieved from Scopus-Scimago and refers to the year when the author's article was published in the journal, or to the latest if not available for that year**.

1. Fantazzini, D., Dalla Valle, L., Giudici, P. (2008). Copulae and operational risks.*International Journal of Risk Assessment and Management*, 9(3), 238 - 257. ***Q3 Business and International Management***

2. Fantazzini, D. (2009). The effects of misspecified marginals and copulas on computing the value at risk: A Monte Carlo study. *Computational Statistics & Data Analysis*, 53(6), 2168-2188.***Q1 Applied Mathematics / Computational Mathematics / Statistics and Probability.***

3. Fantazzini, D. (2010a). Three-stage semi-parametric estimation of t-copulas: Asymptotics, finite-sample properties and computational aspects. *Computational Statistics & Data Analysis*, 54(11), 2562-2579. ***Q1 Applied Mathematics / Computational Mathematics / Statistics and Probability.***

4. Fantazzini, D. (2010b). Modelling and forecasting the global financial crisis: Initial findings using heterosckedastic log-periodic models. *Economics Bulletin*, 30(3), 1833-1841. ***Q2 Economics, Econometrics and Finance***

5. Fantazzini, D. (2011a). Forecasting the global financial crisis in the years 2009-2010: Ex-post analysis. *Economics Bulletin*, 31(4), 3259-3267. ***Q2 Economics, Econometrics and Finance***

6. Fantazzini, D. (2011b). Analysis of multidimensional probability distributions with copula functions - Part I. *Applied Econometrics*, 22(2), 98-134. ***Q4 Economics, Econometrics and Finance***

7. Fantazzini, D. (2011c). Analysis of multidimensional probability distributions with copula functions - Part II. *Applied Econometrics*, 23(3), 98-132. ***Q4 Economics, Econometrics and Finance***

8. Fantazzini, D. (2011d). Analysis of multidimensional probability distributions with copula functions - Part III. *Applied Econometrics*, 24(4), 100-130. ***Q4 Economics, Econometrics and Finance***

9. Geraskin, P., & Fantazzini, D. (2013). Everything you always wanted to know about log-periodic power laws for bubble modeling but were afraid to ask. *The European Journal of Finance*, 19(5), 366-391. ***Q2 Economics, Econometrics and Finance***

10. Fantazzini, D. (2014). Nowcasting and Forecasting the Monthly Food Stamps Data in the US Using Online Search Data. *PloS one*, 9(11), e111894. ***Q1 Multidisciplinary***

11. Fantazzini, D., & Maggi, M. (2015). Proposed coal power plants and coal-to-liquids plants in the US: Which ones survive and why?. *Energy Strategy Reviews*, 7, 9-17. ***Q1 Energy (including Energy Economics)***

12. Fantazzini, D., & Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics*, 170, 97-135. ***Q1 Economics, Econometrics and Finance***

13. Fantazzini, D. (2016). The oil price crash in 2014/15: Was there a (negative) financial bubble? *Energy Policy*, 96, 383-396. ***Q1 Energy (including Energy Economics)***

14. Fantazzini, D., Nigmatullin, E., Sukhanovskaya, V., & Ivliev, S. (2016). Everything you always wanted to know about bitcoin modelling but were afraid to ask - Part 1. *Applied Econometrics*, 44, 5-24. ***Q4 Economics, Econometrics and Finance***

15. Fantazzini, D., Nigmatullin, E., Sukhanovskaya, V., & Ivliev, S. (2017). Everything you always wanted to know about bitcoin modelling but were afraid to ask - Part 2. *Applied Econometrics* 45, 5-28. ***Q4 Economics, Econometrics and Finance***

16. Fantazzini, D., Shakleina, M., Yuras, N. (2018). Big Data for computing social well-being indices of the Russian population. *Applied Econometrics*, 50(2), 43-66. ***Q4 Economics, Econometrics and Finance***

17. Fantazzini D., Shangina T. (2019) The importance of being informed: forecasting market risk measures for the Russian RTS index future using online data and implied volatility over two decades, *Applied Econometrics* 55, 5-31. ***Q4 Economics, Econometrics and Finance***

18. Fantazzini, D. Zimin, S. (2020) A multivariate approach for the simultaneous modelling of market risk and credit risk for cryptocurrencies, *Journal of Industrial and Business Economics,* 47, 19-69. ***Q2 Economics, Econometrics and Finance***

References

Askitas, N. and Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting.

Basel Committee on Banking Supervision (2011). Operational risk–supervisory guidelines for the advanced measurement approaches, *Technical Report* .

Bauwens, L., Hafner, C. M. and Laurent, S. (2012). *Handbook of volatility models and their applications*, Vol. 3, John Wiley & Sons.

Blazquez, D. and Domenech, J. (2018). Big data sources and methods for social and economic analyses, *Technological Forecasting and Social Change* **130**: 99–113.

Cherubini, U., Luciano, E. and Vecchiato, W. (2004). *Copula Methods in Finance*, John Wiley & Sons.

Choi, H. and Varian, H. (2012). Predicting the present with google trends, *Economic Record* **88**: 2–9.

Dinis, G., Breda, Z., Costa, C. and Pacheco, O. (2019). Google trends in tourism and hospitality research: a systematic literature review, *Journal of Hospitality and Tourism Technology* **10**(4): 747–763.

D'Amuri, F. and Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment, *International Journal of Forecasting* **33**(4): 801–816.

Edelman, B. (2012). Using internet data for economic research, *Journal of Economic Perspectives* **26**(2): 189–206.

Engelberg, J. and Gao, P. (2011). In search of attention, *The Journal of Finance* **66**(5): 1461–1499.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models, *Journal of Business & Economic Statistics* **20**(3): 339–350.

Ettredge, M., Gerdes, J. and Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics, *Communications of the ACM* **48**(11): 87–92.

Fantazzini, D. (2009). The effects of misspecified marginals and copulas on computing the value at risk: A monte carlo study, *Computational Statistics & Data Analysis* **53**(6): 2168–2188.

Fantazzini, D. (2010a). Modelling and forecasting the global financial crisis: Initial findings using heterosckedastic log-periodic models", *Economics Bulletin* **30**(3): 1833–1841.

Fantazzini, D. (2010b). Three-stage semi-parametric estimation of t-copulas: Asymptotics, finite-sample properties and computational aspects, *Computational Statistics & Data Analysis* **54**(11): 2562–2579.

Fantazzini, D. (2011a). Analysis of multidimensional probability distributions with copula functions, *Applied Econometrics* **22**(2): 98–134.
URL: *https://ideas.repec.org/a/ris/apltrx/0077.html*

Fantazzini, D. (2011b). Analysis of multidimensional probability distributions with copula functions. ii, *Applied Econometrics* **23**(3): 98–132.
URL: *https://ideas.repec.org/a/ris/apltrx/0094.html*

Fantazzini, D. (2011c). Analysis of multidimensional probability distributions with copula functions. iii, *Applied Econometrics* **24**(4): 100–130.
URL: *https://ideas.repec.org/a/ris/apltrx/0105.html*

Fantazzini, D. (2011d). Forecasting the global financial crisis in the years 2009-2010: Ex-post analysis", *Economics Bulletin* **31**(4): 3259–3267.

Fantazzini, D. (2014). Nowcasting and forecasting the monthly food stamps data in the US using online search data, *PloS one* **9**(11): e111894.

Fantazzini, D. (2016). The oil price crash in 2014/15: Was there a (negative) financial bubble?, *Energy Policy* **96**: 383–396.

Fantazzini, D. (2019). *Quantitative finance with R and cryptocurrencies*, Amazon KDP, ISBN-13: 978–1090685315.

Fantazzini, D., Dalla Valle, L. and Giudici, P. (2008). Copulae and operational risks, *International Journal of Risk Assessment and Management* **9**(3): 238–257.

Fantazzini, D. and Maggi, M. (2015). Proposed coal power plants and coal-to-liquids plants in the US: Which ones survive and why?, *Energy Strategy Reviews* **7**: 9–17.

Fantazzini, D., Nigmatullin, E., Sukhanovskaya, V. and Ivliev, S. (2016). Everything you always wanted to know about bitcoin modelling but were afraid to ask. part 1, *Applied Econometrics* **44**: 5–24.

Fantazzini, D., Nigmatullin, E., Sukhanovskaya, V. and Ivliev, S. (2017). Everything you always wanted to know about bitcoin modelling but were afraid to ask. part 2, *Applied Econometrics* **45**: 5–28.

Fantazzini, D., Shakleina, M., Yuras, N. et al. (2018). Big Data for computing social well-being indices of the Russian population, *Applied Econometrics* **50**: 43–66.

Fantazzini, D. and Shangina, T. (2019). The Importance of Being Informed: Forecasting Market Risk Measures for the Russian RTS Index Future Using Online Data and Implied Volatility Over Two Decade, *Applied Econometrics* **55**: 5–31.

Fantazzini, D. and Zimin, S. (2020). A multivariate approach for the simultaneous modelling of market risk and credit risk for cryptocurrencies, *Journal of Industrial and Business Economics* **47**: 19–69.

Fenn, J. and Raskino, M. (2008). *Mastering the hype cycle: how to choose the right innovation at the right time*, Harvard Business Press.

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics, *International journal of information management* **35**(2): 137–144.

Gantz, J. and Reinsel, D. (2011). Extracting value from chaos, *IDC iview* **1142**(2011): 1–12.

Geraskin, P. and Fantazzini, D. (2013). Everything you always wanted to know about log-periodic power laws for bubble modeling but were afraid to ask, *The European Journal of Finance* **19**(5): 366–391.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data, *Nature* **457**(7232): 1012.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M. and Watts, D. J. (2010). Predicting consumer behavior with web search, *Proceedings of the National academy of sciences* **107**(41): 17486–17490.

Hoeffding, W. (1940). Masstabinvariante korrelationstheorie, *Schriften des Mathematischen Instituts und Instituts fur Angewandte Mathematik der Universitat Berlin* **5**: 181–233.

Jin, X., Wah, B. W., Cheng, X. and Wang, Y. (2015). Significance and challenges of big data research, *Big Data Research* **2**(2): 59–64.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*, CRC Press.

Jun, S.-P., Park, D.-H. and Yeom, J. (2014). The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference, *Technological Forecasting and Social Change* **86**: 237–253.

Jun, S.-P., Yoo, H. S. and Choi, S. (2018). Ten years of research change using google trends: From the perspective of big data utilizations and applications, *Technological Forecasting and Social Change* **130**: 69–87.

Kaushik, A. (2009). *Web analytics 2.0: The art of online accountability and science of customer centricity*, John Wiley & Sons.

Kitchin, R. and McArdle, G. (2016). What makes big data, big data? exploring the ontological characteristics of 26 datasets, *Big Data & Society* **3**(1): 2053951716631130.

Kotler, P. and Keller, K. (2008). *Marketing management 13th edition*, Prentice Hall.

Laney, D. (2001). 3d data management: Controlling data volume, velocity and variety, *META group research note* **6**(70): 1.

Li, J., Xu, L., Tang, L., Wang, S. and Li, L. (2018). Big data in tourism research: A literature review, *Tourism Management* **68**: 301–323.

Lui, C., Metaxas, P. T. and Mustafaraj, E. (2011). On the predictability of the US elections through search volume activity, *Proceedings of the IADIS International Conference on e-Society*.

Mavragani, A. and Tsagarakis, K. (2016). YES or NO: Predicting the 2015 Greek referendum results using Google Trends, *Technological Forecasting and Social Change* **109**: 1–5.

McNeil, A. J., Frey, R. and Embrechts, P. (2015). *Quantitative risk management: Concepts, techniques and tools*, Princeton university press.

Nelsen, R. B. (1999). *An introduction to Copulas*, Vol. 139, Springer-Verlag, New York.

Shim, S., Eastlick, M. A., Lotz, S. L. and Warrington, P. (2001). An online prepurchase intentions model: the role of intention to search, *Journal of retailing* **77**(3): 397–416.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges, *Publ. inst. statist. univ. Paris* **8**: 229–231.

To, P.-L., Liao, C. and Lin, T.-H. (2007). Shopping motivations on internet: A study based on utilitarian and hedonic value, *Technovation* **27**(12): 774–787.

Varian, H. R. (2014). Big data: New tricks for econometrics, *Journal of Economic Perspectives* **28**(2): 3–28.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G. and Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study, *International Journal of Production Economics* **165**: 234–246.

Xiang, Z. and Fesenmaier, D. R. (2005). Assessing the initial step in the persuasion process: Meta tags on destination marketing websites, *Information Technology & Tourism* **8**(2): 91–104.

Zimmer, J. C., Henry, R. M. and Butler, B. S. (2007). Determinants of the use of relational and nonrelational information sources, *Journal of Management Information Systems* **24**(3): 297–331.

Appendix 1 "Copulae and operational risks."

Fantazzini, D., Dalla Valle, L., Giudici, P. (2008). Copulae and operational risks. *International Journal of Risk Assessment and Management*, 9(3), 238 - 257. DOI: https://doi.org/10.1504/IJRAM.2008.019743

Permisson to post the Publisher's version/PDF: according to https://www.inderscience.com/mobile/inauthors/index.php?pid=74 "*only in cases where a funding agency providing the grant for the research on which the Article is based requires this of the Author, upon condition that it shall not be accessible until after six months from Inderscience's publication date*" and/or when *"incorporating the article content in other works by the Author"*. However, according to https://www.inderscience.com/mobile/librarians/index.php?pid=89, "*PhD students requesting permission to reproduce their own work in their thesis should contact support@inderscience.com"*. I requested permission, but I have not yet received an answer.

Copulae and Operational Risks

Luciana Dalla Valle *
University of Milano-Bicocca

Dean Fantazzini†
Universityà of Pavia

Paolo Giudici ‡
University of Pavia

Abstract

The management of Operational Risks has always been difficult due to the high number of variables to work with and their complex multivariate distribution. A Copula is a statistic tool which has been recently used in finance and engineering to build flexible joint distributions in order to model a high number of variables. The goal of this paper is to propose its use to model Operational Risks, by showing its benefits with an empirical example.

JEL classification: $C13, C32, C51$

Keywords: Copulae, Two-step estimation, Operational Risks, VaR, Expected Shortfall.

PRE-PRINT VERSION OF THE FOLLOWING PUBLICATION:
Fantazzini, D., Dalla Valle, L., Giudici, P. (2008). Copulae and operational risks. *International Journal of Risk Assessment and Management*, 9(3), 238 - 257.
The published version is available at https://doi.org/10.1504/IJRAM.2008.019743

*Department of Statistics, University of Milano-Bicocca, Italy.
†Department of Economics and Quantitative Methods, University of Pavia, 27100 Pavia, Italy. Phone ++39-338-9416321, Fax ++39-0382-304226, E-mail: deanfa@eco.unipv.it
‡Department of Economics and Quantitative Methods, University of Pavia, 27100 Pavia, Italy.

1 Introduction

The term "operational risks" is used to define all financial risks that are not classified as market or credit risks . Examples include different categories: we can have the simple risk due to transactions, lack of authorizations, human errors, law suits, etc. Of course there are many more operational risks to be estimated (e.g. security risk, IT risks, etc): here we deal only with the financial risk management of operational risks.

A more precise definition of operational risks includes the direct or indirect losses caused by the inadequacy or malfunction of procedures, human resources and inner systems, or by external events. Basically, they are all losses due to human errors, technical or procedural problems or other causes not linked to the behavior of financial operators or market events.

Operational risks may present different typologies of data, qualitative and quantitative: the first ones include evaluation questionnaires compiled by experts; the second ones include direct and indirect financial losses, performance indicators, rating classes and risk scoring. The main problems when modelling operational risks are therefore the shortage of data and their complex multivariate distribution.

The Basel Committee on Banking supervision (1996,1998) allows for both a simple "top-down " approach (Basic Indicator Approach and Standardized Approach) and a more complex " bottom-up " approach, like the Advanced Measurement Approach (AMA) to estimate the required capital for operational risks. The first approach includes all the models which consider operational risks at a central level, so that local Business Lines (BLs) are not involved. The second approach measures operational risks at the BLs level, instead, and then they are aggregated, thus allowing for a better control at the local level. The methodology we propose belongs to this this second approach and is named Loss Distribution Approach (LDA). The novelty of our approach lies in taking into account the dependence structure among intersections. This is achieved by the copula function: the full Value at Risk (or Expected Shortfall) is then estimated by simulating the joint distribution function of all losses with a Monte Carlo procedure. This approach is able to reduce the required capital imposed by the Basel Committee and financial institutions can save important resources.

The rest of the paper is organized as follows: Section 2 describes the model we propose, while Section 3 presents the marginal distributions used for modelling the frequency and the severity of losses. Section 4 reviews the main points of copula theory, while Section 5 reports the results of a Monte Carlo study of the small sample properties of the marginal distributions estimators. Section 6 presents the empirical analysis, and Section 7 concludes.

2 Model Description

The actuarial approach employs two types of distributions: the one that describes the frequency of risky events and the one that describes the severity of the losses that arise for each considered event. The frequency represents the number of loss events in a time horizon, while the severity is the loss associated to the $k-th$ loss event. Formally, for each type of risk i and for a given time period, operational losses could be defined as a sum (S_i) of the random number (n_i) of the losses (X_{ij}):

$$S_i = X_{i1} + X_{i2} + \ldots + X_{in_i}. \tag{2.1}$$

A widespread statistical model is the *actuarial* model . In this model, the probability distribution of S_i could be described as follows:

$$F_i(S_i) = F_i(n_i) \cdot F_i(X_{ij}), \quad \text{where}$$

- $F_i(S_i)$ = probability distribution of the expected loss for risk i;
- $F_i(n_i)$ = probability of event (frequency) for risk i;
- $F_i(X_{ij})$ = loss given event (severity) for risk i.

The underlying assumptions for the actuarial model are:

- the losses are random variables, independent and identically distributed (i.i.d.);
- the distribution of n_i (frequency) is independent of the distribution of X_{ij} (severity).

Alternative Bayesian models were proposed by Cornalba and Giudici (2004). In the actuarial model, the frequency of a loss event in a certain time horizon could be modelled by a Poisson distribution or a Negative Binomial. For the severity, we could use an Exponential, a Pareto or a Gamma distribution. The distribution F_i of the losses S_i for each intersection i among business lines and event types, is then obtained by the convolution of the frequency and severity distributions: nevertheless, the analytic representation of this distribution is computationally difficult or impossible. For this reason we prefer to approximate this distribution by Monte Carlo simulation: we generate a great number of possible losses (i.e. 100000) with random extractions from the theoretical distributions that describe frequency and severity. We obtain in this way a loss scenario for each risky intersection i.
A risk measure like *Value at Risk (VaR)* or *Expected Shortfall (ES)* is then estimated to evaluate the capital requirement for that particular intersection i. The Value at Risk could be defined as a statistical tool that measures the worst expected loss over a specific time interval at a given confidence level. Formally,

Definition 2.1 (Value at Risk) *The VaR at the confidence level α is the α-quantile of the loss distribution for the $i-th$ risk:*

$$VaR(S_i;\alpha): \quad \Pr(S_i \geq VaR) \leq \alpha. \tag{2.2}$$

while 1 - α is the confidence level.

For example, the 1% VaR is defined as the $1-th$ percentile of the loss distribution F_i. As we've said before, the analytical representation of this distribution doesn't exist or is computationally difficult, and by thus we use a Monte Carlo simulation.
Therefore, VaR represents the maximum loss of a risky intersection i, for a given confidence level $1-\alpha$: however, when this event occurs, it doesn't give any information about the dimension of this loss. Moreover, it was shown that Value at Risk is not a "*coherent risk measure*" (Artzner et al.,1999), and it could underestimate risk when dealing with leptokurtic variables, with potential great losses . (Yamai and Yoshiba, 2002).
An alternative risk measure, which has recently received great attention is the *Expected Shortfall*, or expected loss (Acerbi and Tasche 2002). Formally,

Definition 2.2 (Expected Shortfall) *the ES at the confidence level α is defined as the expected loss for intersection i, given the loss has exceeded the VaR with probability level α :*

$$ES(S_i;\alpha) \equiv \quad \mathrm{E}\left[S_i | S_i \geq VaR(S_i;\alpha)\right] \tag{2.3}$$

The ES at the confidence level 1 - α for a given time horizon, represents the expected value of the losses that have exceeded the correspondent quantile given by the VaR: for example, the expected loss at the 99% confidence level is defined as the portfolio average loss, conditionally to the fact that the losses exceeded the $1-th$ percentile of the loss distribution, given by the 1% VaR. Therefore, differently from Value at Risk, the Expected Shortfall indicates the average loss level that could be achieved in a given time horizon, given the losses exceed the loss correspondent to a certain confidence level.

Once the risk measures for each intersection i are estimated, the global VaR is usually computed as the simple sum of these individual measures, thus assuming a perfect dependence among the different losses S_i. In this paper, instead, we want to show how copulae can be used to describe the dependence structure among the losses S_i and to reduce the required capital to allocate for the global VaR.

A copula is a function that models the dependence structure among the variables of a random vector: in our case, this vector contains the losses for each risk event i. Moreover, when the copula is applied to the marginal distributions of these variables, *not necessarily equal*, it defines their multivariate distribution.

In fact, the Sklar's theorem (1959)[1] tells us that the joint distribution H of a vector of losses S_i, $i = 1 \ldots R$, is the copula of the cumulative distribution functions of the losses' marginals :

$$H(S_1, \ldots, S_R) = C(F(S_1), \ldots, F(S_R)) \tag{2.4}$$

A copula allow us to splits the joint distribution of a random vector of losses into individual components given by the marginals, with a dependence structure among them given by a copula. Consequently, copulae allow us to model the dependence structure among different variables in a flexible way and, at the same time, to use marginal distributions not necessarily equal.

Nelsen (1999) provide an introduction to copula theory, while Cherubini et al. (2004) discusses the main financial applications of the copulae.

The analytic representation for the multivariate distribution of all losses S_i with copula functions is not possible, and an approximate solution with Monte Carlo methods is necessary.

To use copula functions, first of all we have to simulate a multivariate random vector from a specified copula C with marginals uniformly distributed in the unit interval [0,1]. Subsequently, we invert the uniform distributions with the losses cumulative distribution functions F_i, *i=1,...,R*, obtaining a loss scenario for each risky intersection i. Since F_i are discontinuous functions with jumps, previously generated with a Monte Carlo procedure, we have to use the *generalized inverse* of the functions F_i, given by $F_i^{-1}(u) = inf\{x : F_i(x) \geq u\}$. Then, we sum the losses S_i for each intersection i, obtaining a global loss scenario. Finally, we repeat the previous three steps a great number of times, and we calculate a risk measure like VaR or ES.

The procedure to obtain the required total capital is the following:

[1] More details are presented in Section 4.

1. Estimate the marginal distribution F_i of the losses S_i for each risk event i, $i = 1,\ldots,R$,
2. Estimate the multivariate distribution H of all losses S_i, $i = 1,\ldots,R$,
3. Calculate the global Value at Risk or Expected Shortfall.

3 Marginals Modelling

We are now going to show in detail the distributions that we'll use to model the losses frequency and severity for each risky intersection.

The ***frequency*** is a discrete phenomenon and the most suitable probability distributions to describe this random variable are the Poisson and the Negative Binomial. We actually want to determine the probability that a certain number of loss events occurs in a predetermined time horizon.
If we denote a random variable with X, it has a *Poisson* distribution with parameter λ if its probability distribution assumes the following form:

$$\varphi(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \text{for } x = 0,1,2,\ldots \text{with} \quad \lambda > 0. \tag{3.1}$$

This random variable enumerates or counts random phenomena that produce events which take place a random number of times in a predetermined time or space interval. This is why it constitutes a suitable distribution to describe the frequency.
If we assume the parameter of the Poisson random variable to be Gamma distributed, we obtain the *Negative Binomial*, which is a special type of mixture distribution. The probability function is given by

$$\varphi(x;\vartheta;p) = \begin{pmatrix} x+\vartheta-1 \\ x \end{pmatrix} p^{\vartheta}(1-p)^x \quad x = 0,1,2,\ldots \quad \text{with} \quad 0 < p < 1 \quad \vartheta > 0 \tag{3.2}$$

where p indicates the probability of success, $(1-p)$ the probability of failure and x the number of failures before obtaining the ϑ-th success. The Negative Binomial random variable is sometimes called "waiting time", since it counts the failures we ought to wait for to have exactly ϑ successes.
The ***severity*** is a continuous phenomenon, instead, which can be described by a density function belonging to the Gamma family.
The random variable X is *Gamma* distributed if it has the following density function:

$$\varphi(x;\beta_1,\beta_2,\beta_3) = \frac{\beta_2^{\beta_1}}{\Gamma(\beta_1)}(x-\beta_3)^{\beta_1-1}e^{-\beta_2(x-\beta_3)} \quad \text{with} \quad x > \beta_3, \quad (\beta_1\beta_2 > 0, \beta_3 \geq 0), \tag{3.3}$$

where

$$\Gamma(\beta) = \int_0^\infty z^{\beta-1}e^{-z}dz.$$

This is the most general form. However, If we put $\beta_1 = \alpha$, $\beta_2 = \frac{1}{\vartheta}$ and $\beta_3 = 0$, the density function is modified to its standard form as follows:

$$\varphi(x;\alpha,\vartheta) = \frac{\left(\frac{1}{\vartheta}\right)^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-x/\vartheta} \quad \text{and} \quad x > 0 \tag{3.4}$$

Another distribution which is suitable to model the loss S_i associated to a certain event for a given intersection i, is represented by the *Exponential*. This random variable could be derived by the Gamma with $\beta_1 = 1$, $\beta_2 = 1/\lambda$, $\beta_3 = 0$, thus obtaining

$$\varphi(x;\lambda) = \frac{1}{\lambda} e^{-\frac{1}{\lambda}x} \quad \lambda > 0. \tag{3.5}$$

Finally, it is possible to model the severity using the *Pareto* distribution, whose density function is given by:

$$\varphi(\alpha;\vartheta) = \frac{\alpha\vartheta^{\alpha}}{(x+\vartheta)^{(\alpha+1)}} \quad x > \vartheta, \quad \alpha > 0, \quad \vartheta > 0. \tag{3.6}$$

The parameters of the previous distribution can be estimated with empirical data by *method of moments* and *method of maximum likelihood*. See Gourieroux and Monfort (1995) for more details.

4 Copula Theory

An n-dimensional copula is a multivariate cumulative distribution function with uniform distributed margins in [0,1]. We now recall its definition, following Joe(1997) and Nelsen (1999).
Let consider $X_1, \dots X_n$ to be random variables, and H their joint distribution function, then we have:

Definition 4.1 (Copula) *A copula is a multivariate distribution function H of random variables $X_1 \dots X_n$ with standard uniform marginal distributions $F_1, \dots, Fn$, defined on the unit n-cube [0,1]^n, with the following properties:*

1. *The range of C $(u_1, u_2, ..., u_n)$ is the unit interval [0,1];*
2. *C $(u_1, u_2, ..., u_n)$ = 0 if any u_i = 0, for i = 1, 2, ..., n.*
3. *C (1, ..., 1, u_i, 1, ..., 1) = u_i , for all $u_i \in$ [0, 1]*

The previous three conditions provides the lower bound on the distribution function and ensures that the marginal distributions are uniform.
The Sklar's theorem justifies the role of copulas as dependence functions.

Theorem 4.1 (Sklar's theorem) *Let H denote a n-dimensional distribution function with margins $F_1 \dots F_n$. Then there exists a n-copula C such that for all real (x_1,..., x_n)*

$$H(x_1, \dots, x_n) = C(F(x_1), \dots, F(x_n)) \tag{4.1}$$

If all the margins are continuous, then the copula is unique; otherwise C is uniquely determined on $RanF_1 \times RanF_2 \dots RanF_n$, where Ran is the range of the marginals. Conversely, if C is a copula and $F_1, \dots F_n$ are distribution functions, then the function H defined in (2.2) is a joint distribution function with margins $F_1, \dots F_n$.

Proof: See Sklar (1959), Joe(1997) or Nelsen (1999). ■

The last statement is the most interesting for multivariate density modelling, since it implies that we may link together any $n \geq 2$ univariate distributions, of any type (not necessarily from the same family), with any copula in order to get a valid bivariate or multivariate distribution.
From Sklar's Theorem, Nelsen (1999) derives the following corollary:

Corollary 4.1 *Let $F_1^{(-1)}$, ..., $F_n^{(-1)}$ denote the generalized inverses of the marginal distribution functions, then for every (u_1 , ..., u_n) in the unit n-cube, exists a unique copula C : [0,1] ×...× [0,1] → [0,1] such that*

$$C(u_1, ..., u_n) = H(F_1^{(-1)}(u_1), \ldots, F_n^{(-1)}(u_n)) \tag{4.2}$$

Proof: See Nelsen (1999), Theorem 2.10.9 and the references given therein. ■

From this corollary we know that given any two marginal distributions and any copula we have a joint distribution. A copula is thus a function that, when applied to univariate marginals, results in a proper multivariate pdf: since this pdf embodies all the information about the random vector, it contains all the information about the dependence structure of its components. Using copulas in this way splits the distribution of a random vector into individual components (marginals) with a dependence structure (the copula) among them without losing any information.
By applying Sklar's theorem and using the relation between the distribution and the density function, we can derive the multivariate copula density $c(F_1(x_1),, \ldots, F_n(x_n))$, associated to a copula function $C(F_1(x_1),, \ldots, F_n(x_n))$:

$$f(x_1, ..., x_n) = \frac{\partial^n \left[C(F_1(x_1), \ldots, F_n(x_n))\right]}{\partial F_1(x_1), \ldots, \partial F_n(x_n)} \cdot \prod_{i=1}^{n} f_i(x_i) = c(F_1(x_1), \ldots, F_n(x_n)) \cdot \prod_{i=1}^{n} f_i(x_i)$$

where

$$c(F_1(x_1), \ ..., \ F_n(x_n)) = \frac{f(x_1, ..., x_n)}{\prod_{i=1}^{n} f_i(x_i)} \cdot \ , \tag{4.3}$$

By using this procedure, we can derive the Normal and the T-copula:

1. The copula of the multivariate Normal distribution is the *Normal-copula*, whose probability density function is:

$$\begin{aligned} c(\Phi(x_1), \ ..., \ \Phi(x_n)) &= \frac{f^{Gaussian}(x_1, ..., x_n)}{\prod_{i=1}^{n} f_i^{Gaussian}(x_i)} = \frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}x'\Sigma^{-1}x\right)}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_i^2\right)} = \\ &= \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\zeta'(\Sigma^{-1} - I)\zeta\right) \end{aligned} \tag{4.4}$$

   where $\zeta = (\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n))'$ is the vector of univariate Gaussian inverse distribution functions, $u_i = \Phi\ (x_i)$, while Σ is the correlation matrix.

2. On the other hand, the copula of the multivariate Student's T-distribution is the Student's

T-copula, whose density function is:

$$c(t_v(x_1),\ ...,\ t_v(x_n)) = \frac{f^{Student}(x_1,...,x_n)}{\prod_{i=1}^{n} f_i^{Student}(x_i)} = \ |\Sigma|^{-1/2}\frac{\Gamma\left(\frac{v+n}{2}\right)}{\Gamma\left(\frac{v}{2}\right)}\left[\frac{\Gamma\left(\frac{v}{2}\right)}{\Gamma\left(\frac{v+1}{2}\right)}\right]^n \frac{\left(1+\frac{\zeta'\Sigma^{-1}\zeta}{v}\right)^{-\frac{v+n}{2}}}{\prod_{i=1}^{n}\left(1+\frac{\zeta_i^2}{2}\right)^{-\frac{v+1}{2}}}, \tag{4.5}$$

where $\zeta = (t_v^{-1}(u_1),...,t_v^{-1}(u_n))'$ is the vector of univariate Student's T inverse distribution functions, ν are the degrees of freedom, $u_i = t_\nu(x_i)$, while Σ is the correlation matrix.

Both these copulae belong to the class of Elliptical copulae [2]. An alternative to Elliptical copulae is given by *Archimedean copulae*: however, they present the serious limitation to model only positive dependence (or only partial negative dependence), while their multivariate extension involve strict restrictions on bivariate dependence parameters. This is why we do not consider them here.
These copula densities can then be used to fit operational risks data with maximum likelihood methods. When we use the *Normal Copula*, the log-likelihood is given by

$$l^{gaussian}(\theta) = -\frac{T}{2}\ln|\Sigma| - \frac{1}{2}\sum_{t=1}^{T}\zeta_t^{'}(\Sigma^{-1}-I)\zeta_t \tag{4.6}$$

If the log-likelihood function is differentiable in θ and the solution of the equation $\partial_\theta\ l(\theta) = 0$ defines a global maximum, we can recover the $\hat{\theta}_{ML} = \hat{\Sigma}$ for the Gaussian copula:

$$\frac{\partial}{\partial\Sigma^{-1}}\ l^{gaussian}(\theta) = \frac{T}{2}\Sigma - \frac{1}{2}\sum_{t=1}^{T}\zeta_t^{'}\zeta_t = 0 \tag{4.7}$$

and therefore

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}\zeta_t^{'}\zeta_t \tag{4.8}$$

When we use the *T-copula*, the log-likelihood is defined as follows:

$l^{Student}\ (\theta\ ) =$

$$-T\ln\frac{\Gamma\left(\frac{\nu+N}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} - NT\ln\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} - \frac{T}{2}\ln|R| - \frac{\nu+N}{2}\sum_{t=1}^{T}\ln\left(1+\frac{\varsigma_t^{'}R^{-1}\varsigma_t}{\nu}\right) + \frac{\nu+1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\ln\left(1+\frac{\varsigma_{it}^2}{\nu}\right)$$

In this case, we don't have any analytical formula for the ML estimator and a numerical maximization of the likelihood is required. However, this can become computationally cumbersome, if not impossible, when the number of operational risks is very large. This is why multistep parametric or semi-parametric approaches have been proposed. Three methods are the most used: the first one, suggested by Bouyé et al. (2000), is based on a recursive optimization procedure for the correlation matrix. However, this procedure can be computationally intensive when dealing with several risky intersections i and large data sets; moreover it can present numerical instability due the inversion of close to singular matrices.
The second method proposed by Marshal and Zeevi (2002) is based on the rank correlation given

[2] See Cherubini et al. (2004) for more details

by the Kendall's Tau. Even though it is faster and more stable then the previous method, it can become computationally cumbersome when the number of considered operational risks is high.
We follow here a third approach proposed by Chen, Patton, et al.(2004), which is a mixed parametric approach, based on Method of Moments and Maximum Likelihood estimates. The estimation steps are the following:

1. Transform the dataset $(x_{1t}, x_{2t}, \ldots, x_{Nt})$, $t = 1 \ldots T$ into uniform variates $(\hat{u}_{1t}, \hat{u}_{2t}, \ldots \hat{u}_{Nt})$, using a parametric distribution function, or the empirical distribution defined as follows:

$$\hat{u}_i(\cdot) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{\{x_{it} \leq\, x_i)}, \;\; i = 1 \ldots N$$

   where $\mathbb{1}_{\{x \leq \bullet\}}$represent the indicator function.;

2. Let $\hat{\Sigma}$ the correlation matrix for the Gaussian copula, estimated using equation (4.8), and then set $\hat{\Sigma}_{Ga} = \hat{\Sigma}$

3. Estimate ν maximizing the log-likelihood function of the Student's T copula density:

$$\hat{\nu} = \arg\max \sum_{t=1}^{T} \log(c^{student}(\hat{u}_{1,\mathrm{t}}, \ldots, \hat{u}_{N,t}; \hat{\Sigma}_{Ga}, \nu)$$

4. Let $\zeta_{\hat{\nu}t} = (t_{\hat{\nu}}^{-1}(\hat{u}_{1t}), \ldots, t_{\hat{\nu}}^{-1}(\hat{u}_{Nt}))$'. Finally, get $\hat{\Sigma}_{T-copula}$ using equation (4.8) again:

$$\hat{\Sigma}_{student} = \frac{1}{T} \sum_{t=1}^{T} \varsigma_{\hat{\nu}t}^{'} \varsigma_{\hat{\nu}t}$$

An iterative procedure can be implemented as well (but Chen, Patton et al. don't do that). However, after the first step the differences are rather minimal. For a review of copula estimation methods and their asymptotic properties, see Fantazzini (2005).

5 Simulation Studies

In this section we present the results of a Monte Carlo study of the small sample properties of the estimators discussed in Section 3 for the parameters of the Frequency and Severity distributions[3]. The simulation Data Generating Processes (DGPs) are designed to reflect the stylized facts about real operational risks: we chose the parameters of the DGPs among the ones estimated in the following empirical section.
We consider two DGPs for the *Frequency*:

$$F_i(n_i) \;\; \sim \;\; Poisson(0.08) \tag{5.1}$$

$$F_i(n_i) \;\; \sim \;\; Negative\,Binomial(0.33; 0.80) \tag{5.2}$$

[3]For extensive Monte Carlo studies of copulas estimators, instead, we refer to Bouye'et al.(2000), Patton (2001), Chen et al.(2004) and Cherubini et al. (2004)

and three DGPs for the *Severity*:

$$F_i(X_{ij}) \sim Exponential(153304) \tag{5.3}$$

$$F_i(X_{ij}) \sim Gamma(0.2; 759717) \tag{5.4}$$

$$F_i(X_{ij}) \sim Pareto(2.51; 230817) \tag{5.5}$$

In addition to the five DGPs, we consider four possible data situations: 1)T = 72; 2)T = 500; 3) T = 1000; 4) T = 2000. The first situation correspond to the size of our empirical dataset, since we have 72 monthly observations ranging from January 1999 till December 2004.
We will look at the *mean* of the N = 10000 replications

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i\right) \tag{5.6}$$

where $\hat{\theta}_i$ is the estimate based on the ith Monte Carlo replication. We'll compare the estimators by looking at their *mean squared error (MSE)*,

$$MSE(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_0\right)^2 \tag{5.7}$$

where θ_0 is the true parameter, and we'll look at their *Variation Coefficient (VC)*, which is an adimensional indicator used to compare the dispersion of different sample distributions:

$$VC(\theta) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \hat{\theta}\right)^2}}{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i\right)} \tag{5.8}$$

Finally, we'll report the *percentage of times out of the N=10000 replications when the parameters are smaller than zero*, that is when the distribution is not defined. The results are reported below in Tables 1-5.

Table 1: **Small sample properties: POISSON distribution**

	POISSON (λ=0.08)			
	Mean	*MSE*	*VC*	*% negative par.*
T = 72	0.0801	0.0011	0.4200	0.0000
T = 500	0.0799	0.0002	0.1575	0.0000
T = 1000	0.0800	0.0001	0.1121	0.0000
T = 2000	0.0800	0.0000	0.0791	0.0000

Table 2: **Small sample properties: NEGATIVE BINOMIAL distribution**

	NEGATIVE BINOMIAL (p = 0.8; θ = 0.33)							
	p = 0.8				θ = 0.33			
	Mean	*MSE*	*VC*	*% negative par.*	*Mean*	*MSE*	*VC*	*% negative par.*
T = 72	0.8605	0.0310	0.1923	0.0000	0.0904	5.3818	25.5136	40.8600
T = 500	0.8194	0.0099	0.1193	0.0000	0.3807	11.3144	8.5772	2.8600
T = 1000	0.8085	0.0052	0.0887	0.0000	0.4439	2.5711	3.6794	0.3100
T = 2000	0.8047	0.0027	0.0645	0.0000	0.3738	0.0240	0.3979	0.0000

The previous tables show some interesting results:

As for ***Frequency*** distributions, the *Poisson* distribution gives already consistent estimates with 72 observations, with zero probability to get negative estimates. Moreover, the Variation Coefficient is already below one with 72 observations.

The *Negative Binomial* shows dramatic results, instead: 40 % of cases showed negative estimates for θ, and very high MSE and Variation Coefficients. Moreover, even with a dataset of 2000 observations, the estimates of θ are not yet stable, and the adimensional VC is still well above 0.1. Unreported simulation results show that the estimated values stabilize around the true values only with datasets of 5000 observations of higher.

This simulation evidence highlights that Negative Binomial estimates can be completely unreliable in small samples, so caution must be taken.

Table 3: **Small sample properties: EXPONENTIAL distribution**

	EXPONENTIAL (λ= 153304)			
	Mean	*MSE*	*VC*	*% negative par.*
T = 72	153369	3.27E+08	0.1179	0.0000
T = 500	153278	4.61E+07	0.0443	0.0000
T = 1000	153352	2.31E+07	0.0313	0.0000
T = 2000	153308	1.17E+07	0.0223	0.0000

Table 4: **Small sample properties: GAMMA distribution**

	GAMMA (α=0.2; θ = 759717)							
	α = 0.2				θ = 759717			
	Mean	*MSE*	*VC*	*% negative par.*	*Mean*	*MSE*	*VC*	*% negative par.*
T = 72	0.2408	0.0068	0.2968	0.0000	698559	1.02E+11	0.44913	0.0000
T = 500	0.2071	0.0009	0.1410	0.0000	749400	1.85E+10	0.18104	0.0000
T = 1000	0.2033	0.0005	0.1035	0.0000	756137	9.46E+09	0.12855	0.0000
T = 2000	0.2014	0.0002	0.0756	0.0000	758652	4.87E+09	0.09194	0.0000

Table 5: **Small sample properties: PARETO distribution**

	PARETO(α=2.51; θ = 230817)							
	α=2.51				θ = 230817			
	Mean	*MSE*	*VC*	*% negative par.*	*Mean*	*MSE*	*VC*	*% negative par.*
T = 72	6.0871	23954	25.4089	2.1200	662616	2.6E+14	24.2447	2.1200
T = 500	3.1439	9.0503	0.1971	0.0000	322315	2.0E+11	0.2405	0.0000
T = 1000	2.9676	7.8541	0.1485	0.0000	297831	2.2E+11	0.1912	0.0000
T = 2000	2.8534	7.1532	0.1177	0.0000	281508	2.3E+11	0.1575	0.0000

As for ***Severity*** distributions, we have again mixed results.

The *Exponential* and *Gamma* distributions give already consistent estimates with 72 observations, and they quickly stabilize around the true values when T increases. The Exponential shows slightly better properties than the Gamma, but this was an expected outcome, since the Exponential is a special case of the Gamma with some parameters restrictions.

The *Pareto* have problems in small samples instead, with 2% of cases of negative coefficients and very high MSE and VC. Similar to the Negative Binomial, the estimates do not reach the true values even with a dataset of 2000 observations, and a size of, at least, T=5000 is required.

Therefore, the previous results suggest to use the Exponential or the Gamma distributions in small samples, where the latter is a better choice when more flexibility is required. This is surely the case for operational risks, where extreme events are very important when estimating risk measures such as Value at Risk or Expected Shortfall.

6 Empirical Analysis

The model we described in Section 2 was applied to an (anonymous) banking loss dataset, ranging from January 1999 till December 2004, for a total of 72 monthly observations. The overall loss events in this dataset are 407, organized in 2 business lines and 4 event types, so that we have 8 possible risky combinations (or intersections) to deal with. For privacy law reasons, the bank assigned a random code to the business lines and the event types in order to hide their identification: however, the direct association between these latter codes and the real ones were preserved.
The overall average monthly loss was equal to 202.158 euro, the minimum to 0 (for September 2001), while the maximum to 4.570.852 euro (which took place on July 2003). Table 6 reports a piece of the dataset we used for the empirical analysis.

Table 6: **Pieces of the banking losses dataset**

Frequency	**1999 January**	**1999 February**	**1999 March**	**1999 April**		**2004 November**	**2004 December**
Intersection 1	2	0	0	0	...	5	0
Intersection 2	6	1	1	1	...	3	1
Intersection 3	0	2	0	0	...	0	0
Intersection 4	0	1	0	0	...	0	0
Intersection 5	0	0	0	0	...	0	1
Intersection 6	0	0	0	0	...	2	4
Intersection 7	0	0	0	0	...	1	0
Intersection 8	0	0	0	0	...	0	0

Severity	**1999 January**	**1999 February**	**1999 March**	**1999 April**		**2004 November**	**2004 December**
Intersection 1	35753	0	0	0	...	27538	0
Intersection 2	121999	1550	3457	5297	...	61026	6666
Intersection 3	0	33495	0	0	...	0	0
Intersection 4	0	6637	0	0	...	0	0
Intersection 5	0	0	0	0	...	0	11280
Intersection 6	0	0	0	0	...	57113	11039
Intersection 7	0	0	0	0	...	2336	0
Intersection 8	0	0	0	0	...	0	0

We estimated the parameters of the frequency and severity distributions by method of moments, for every risky intersection. Table 7 reports the parameters of the frequency distributions n_i, while Table 8 shows the ones of the severity distributions X_{ij}.

Table 7: **Estimated parameters of the *Frequency* distributions**

	Poisson	**Negative Binomial**	
	λ	**p**	θ
Intersection 1	1.40	0.59	2.01
Intersection 2	2.19	0.40	1.49
Intersection 3	0.08	0.80	0.33
Intersection 4	0.46	0.92	5.26
Intersection 5	0.10	0.84	0.52
Intersection 6	0.63	0.33	0.31
Intersection 7	0.68	0.42	0.49
Intersection 8	0.11	0.88	0.80

Table 8: **Estimated parameters of the *Severity* distributions**

	Gamma		Exponential	Pareto	
	α	θ	λ	α	θ
Intersection 1	0.15	64848	9844	2.36	13368
Intersection 2	0.20	109321	21721	2.50	32494
Intersection 3	0.20	759717	153304	2.51	230817
Intersection 4	0.11	1827627	206162	2.25	258588
Intersection 5	0.20	495701	96873	2.49	143933
Intersection 6	0.38	19734	7596	3.25	17105
Intersection 7	0.06	211098	12623	2.13	14229
Intersection 8	0.26	135643	35678	2.71	61146

We obtained the marginal distribution of the losses S_i for every intersection between business lines and event type thanks to the convolution of the frequency and severity distributions, and we approximated it by Monte Carlo simulations. We then estimated the Value at Risk and Expected Shortfall with a 95% and 99% confidence levels, and their sum for all intersections i gave us a measure of the global VaR and ES for the case of perfect dependence.

Besides, we got also the global VaR by using copulas, in order to model the dependence structure among the marginal losses S_i within a more realistic framework than the previous perfect dependence case. Table 9 presents the correlation matrix of the risky intersections i estimated with the Normal Copula, while Table 10 reports the global VaR and ES relative to different frequency and severity distributions, as well as different copulas. Figure 1 shows, as an example, the Global Loss Distribution used to estimate the VaR and ES values, when using the Negative Binomial for the frequency distribution, the Pareto for the Severity distribution and a Normal Copula for the dependence structure.

First of all, it is possible to observe that the hypothesis of perfect dependence is not realistic since all correlations are rather small and all around zero.

Secondly, one can notice that copulae allow for a remarkable saving of money for the bank. If we compare the case of perfect dependence to that of copulas, we see that in the latter case the required capital is always lower, with savings ranging between 30 and 50 % with respect to the former case. This is particularly clear when comparing the Expected Shortfall values.

The choice of the Normal or T-copula (with 9 degrees of freedom) doesn't modify the results substantially, since it is more important a proper choice of the marginal distributions, instead, and this is particularly true for the severity.

Figure 1: **Global Loss Distribution (*Negative Binomial - Pareto - Normal copula*)**

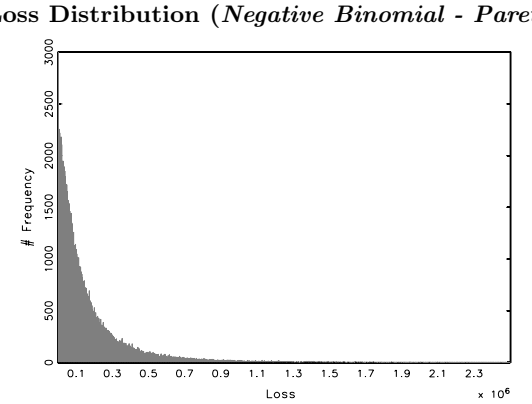


Table 9: **Correlation Matrix of the risky Intersections (Normal Copula)**

	Inters. 1	**Inters. 2**	**Inters. 3**	**Inters. 4**	**Inters. 5**	**Inters. 6**	**Inters. 7**	**Inters. 8**
Inters. 1	1	-0.050	-0.142	0.051	-0.204	0.252	0.140	-0.155
Inters. 2	-0.050	1	-0.009	0.055	0.023	0.115	0.061	0.048
Inters. 3	-0.142	-0.009	1	0.139	-0.082	-0.187	-0.193	-0.090
Inters. 4	0.051	0.055	0.139	1	-0.008	0.004	-0.073	-0.045
Inters. 5	-0.204	0.023	-0.082	-0.008	1	0.118	-0.102	-0.099
Inters. 6	0.252	0.115	-0.187	0.004	0.118	1	-0.043	0.078
Inters. 7	0.140	0.061	-0.193	-0.073	-0.102	-0.043	1	-0.035
Inters. 8	-0.155	0.048	-0.090	-0.045	-0.099	0.078	-0.035	1

Table 10: **Global VaR and ES for different marginals convolutions, dependence structures, and confidence levels**

		VaR 95%	VaR 99%	ES 95%	ES 99%
Poisson Exponential	*Perfect Dep.*	925,218	1,940,229	1,557,315	2,577,085
	Normal Copula	656,068	1,086,725	920,446	1,340,626
	T copula (9 d.o.f.)	673,896	1,124,606	955,371	1,414,868
Poisson Gamma	*Perfect Dep.*	861,342	3,694,768	2,640,874	6,253,221
	Normal Copula	767,074	2,246,150	1,719,463	3,522,009
	T copula (9 d.o.f.)	789,160	2,366,876	1,810,302	3,798,321
Poisson Pareto	*Perfect Dep.*	860,066	2,388,649	2,016,241	4,661,986
	Normal Copula	663,600	1,506,466	1,294,654	2,785,706
	T copula (9 d.o.f.)	672,942	1,591,337	1,329,130	2,814,176
Negative Bin. Exponential	*Perfect Dep.*	965,401	2,120,145	1,676,324	2,810,394
	Normal Copula	672,356	1,109,768	942,311	1,359,876
	T copula (9 d.o.f.)	686,724	1,136,445	975,721	1,458,298
Negative Bin. Gamma	*Perfect Dep.*	907,066	3,832,311	2,766,384	6,506,154
	Normal Copula	784,175	2,338,642	1,769,653	3,643,691
	T copula (9 d.o.f.)	805,747	2,451,994	1,848,483	3,845,292
Negative Bin. Pareto	*Perfect Dep.*	859,507	2,486,971	2,027,962	4,540,441
	Normal Copula	672,826	1,547,267	1,311,610	2,732,197
	T copula (9 d.o.f.)	694,038	1,567,208	1,329,281	2,750,097

Despite the Basel agreements require a backtesting procedure involving at least 250 observations, we anyway use this methodology to compare the different models (Table 5). Our decision is justified by the fact that record-keeping of operational risks losses is a very recent procedure, and older datasets that starts before 1999 are very rare and/or not reliable. For greater details about alternative backtesting methods and distributions, see Giudici(2004).

Table 11: **Backtesting results with different marginals distributions and different copulae**

		VaR	**Exceedances N / T**			**VaR**	**Exceedances N / T**
Poisson Exponential	**Perfect Dependence**	99.00%	**1.39%**	**Negative Bin. Exponential**	**Perfect Dependence**	99.00%	**1.39%**
		95.00%	**4.17%**			95.00%	**4.17%**
	Normal Copula	99.00%	**2.78%**		**Normal Copula**	99.00%	**2.78%**
		95.00%	**6.94%**			95.00%	**6.94%**
	T Copula (9 d.o.f.)	99.00%	**2.78%**		**T Copula (9 d.o.f.)**	99.00%	**2.78%**
		95.00%	**6.94%**			95.00%	**6.94%**
Poisson Gamma	**Perfect Dependence**	99.00%	**1.39%**	**Negative Bin. Gamma**	**Perfect Dependence**	99.00%	**1.39%**
		95.00%	**6.94%**			95.00%	**4.17%**
	Normal Copula	99.00%	**1.39%**		**Normal Copula**	99.00%	**1.39%**
		95.00%	**6.94%**			95.00%	**6.94%**
	T Copula (9 d.o.f.)	99.00%	**1.39%**		**T Copula (9 d.o.f.)**	99.00%	**1.39%**
		95.00%	**6.94%**			95.00%	**6.94%**
Poisson Pareto	**Perfect Dependence**	99.00%	**1.39%**	**Negative Bin. Pareto**	**Perfect Dependence**	99.00%	**1.39%**
		95.00%	**6.94%**			95.00%	**6.94%**
	Normal Copula	99.00%	**1.39%**		**Normal Copula**	99.00%	**1.39%**
		95.00%	**6.94%**			95.00%	**6.94%**
	T Copula (9 d.o.f.)	99.00%	**1.39%**		**T Copula (9 d.o.f.)**	99.00%	**1.39%**
		95.00%	**6.94%**			95.00%	**6.94%**

Table 11 shows that the Exponential distribution for severity modelling presents the worst back-testing results, while the Gamma and Pareto have a better behavior. However, we showed in Section 5 that the Pareto distribution have problems when dealing with small samples, since it requires a high number of observations to have consistent parameters estimates (at least higher than 5.000). This is why the Gamma distribution is usually the best choice.
We finally report in Table 12 the Log-Likelihood at the estimated parameters, as well as the Schwarz Criterion, to appraise the goodness of fit of the marginals distributions. Similarly to what we've found so far, the frequency distributions n_i do not show any relevant differences, while the Gamma and Pareto distributions are the best choices to model the severities X_{ji}.

Table 12: **Log-Likelihood and Schwarz Criterion**

		FREQUENCY		SEVERITY		
		Poisson	Negative Binomial	Gamma	Exponential	Pareto
Intersection 1	LOG LIKELIHOOD	-121.62	-116.08	-1057.34	-1013.97	-976.69
	Schwarz Criterion	3.44	3.34	21.03	20.17	19.43
Intersection 2	LOG LIKELIHOOD	-154.83	-147.01	-1480.77	-1730.76	-1687.98
	Schwarz Criterion	4.36	4.20	18.57	21.67	21.16
Intersection 3	LOG LIKELIHOOD	-21.60	-20.91	-99.60	-142.34	-137.72
	Schwarz Criterion	0.66	0.70	18.54	26.10	25.48
Intersection 4	LOG LIKELIHOOD	-64.70	-64.59	-430.48	-438.92	-420.01
	Schwarz Criterion	1.86	1.91	26.30	26.71	25.67
Intersection 5	LOG LIKELIHOOD	-24.01	-23.55	-149.74	-149.77	-145.29
	Schwarz Criterion	0.73	0.77	25.37	25.17	24.63
Intersection 6	LOG LIKELIHOOD	-91.95	-75.86	-439.64	-428.50	-424.62
	Schwarz Criterion	2.61	2.23	19.71	19.13	19.04
Intersection 7	LOG LIKELIHOOD	-93.95	-79.78	-542.33	-511.72	-476.04
	Schwarz Criterion	2.67	2.33	22.29	20.97	19.59
Intersection 8	LOG LIKELIHOOD	-26.27	-25.99	-91.47	-91.86	-89.47
	Schwarz Criterion	0.79	0.84	23.39	23.22	22.89

7 Conclusions

The goal of this work was to apply to loss banking data a model for estimating the required capital when dealing with multivariate operational risks. The main contribution of our paper is the proposition of a new method (based on copula theory) to model loss data in a multivariate framework.
We compared different marginals distributions to model the losses frequency and severity, and we estimated the Value at Risk and Expected Shortfall for different confidence levels. The global risk measures were then obtained by using Normal or Student's T-Copulas, which are proper tools to model the dependence structure among losses of different risky intersections.
The empirical analysis showed that is not the choice of the copula, but that of the marginals which is important, instead, particularly the ones used to model the losses severity. The best distribution for severity modelling resulted to be the Gamma distribution, while remarkable differences between the Poisson and Negative Binomial for frequency modelling, were not found. However, we have to remind that the Poisson is much more easier to estimate, especially with small samples.
We showed that copula functions are able to model the dependence structure of risky events and therefore allow us to reduce the risk measures capital requirements. Differently from the perfect dependence case, which is far more conservative, the copula approach represents a big advantage

in terms of capital savings for any financial institutions.

Note: This article is the result of the joint work of the three authors. However, Section 1 and 3 were written by Dalla Valle L., Section 4 - 5 - 6 by Fantazzini D., while Sections 2 and 7 were written by Dalla Valle L. and Fantazzini D.

References

[1] Acerbi C., Tasche D. (2002) On the coherence of Expected Shortfall, *J. of Banking and Finance*, 26,1487-1503.

[2] Artzner, P., F. Delbaen, J. Eber, and D. Heath (1999), Coherent Measures of Risk , *Mathematical Finance*.

[3] Bilotta A., Giudici P. (2004), Modelling Operational Losses: A Bayesian Approach, *Quality and Reliability Engineering International*, 20, 407-417.

[4] Bouye' E., V. Durrleman, A. Nikeghbali, G. Riboulet and T. Roncalli (2001): Copulas: an open field for risk management, *Groupe de Recherche Oprationnelle, Credit Lyonnais*, Working Paper.

[5] Casella G., Lehmann E.L. (1998), *Theory of Point Estimation*, Springer.

[6] Chen X., Fan Y., Patton A. (2004), Simple Tests for Models of Dependence Between Multiple Financial Time Series, with Applications to U.S. Equity Returns and Exchange Rates, Financial Markets Group, London School of Economics, Discussion Paper 483.

[7] Cherubini U., Luciano E., Vecchiato W., *Copula Methods in Finance*, Wiley, 2004.

[8] Cornalba C., Giudici P. (2004), Statistical models for operational risk management, *Phisica A*, 338, 166-172.

[9] Fantazzini D. (2005), The econometric Modelling of Copulas: A Review with Extensions, University of Pavia, Department of Economics Working paper

[10] Gabbi G., Marsella M., Masacesi M.(2005), *Il rischio operativo nelle banche. Aspetti teorici ed esperienze aziendali, misurazione e gestione*, EGEA.

[11] Giudici P. , *Applied Data Mining, Statistical Methods for Business and Industry*, Wiley, 2003

[12] Gourieroux C. and Monfort A. (1995), *Statistics and Econometric Models*, Cambridge University Press

[13] Mashal R. and A. Zeevi (2002), Beyond Correlation: Extreme Co-movements Between Financial Assets, Columbia University, Working Paper.

[14] Nelsen R.B. (1999): *An Introduction to Copulas*, Lecture Notes in Statistics 139, Springer, N.Y.

[15] Patton, Andrew J., (2001), Modelling Time-Varying Exchange Rate Dependence Using the Conditional Copula, Working Paper 2001-09, Department of Economics, University of California, San Diego.

[16] Shao J. (2003), *Mathematical Statistics*, Springer.

[17] Yamai, Y., Yoshiba T.(2002), Comparative analyses of Expected Shortfall and Value-at-Risk: their validity under market stress, *Monetary and Economic Studies*, October, pp. 181-238.

Appendix

A Copula Simulation

We report below the main steps to do in order to simulate from a given copula. For a detailed review of copula simulation see Cherubini et al. (2004).

A.1 Normal Copula

In order to generate random variates from the Gaussian copula (4.4), we can use the following procedure. If the matrix Σ is positive definite, then there are some $n \times n$ matrix $\mathbf{A}$ such that $\Sigma = \mathbf{A}\mathbf{A}^T$. It is also assumed that the random variables $Z_1, ..., Z_n$ are independent standard normal. Then, the random vector $\mu + \mathbf{AZ}$ (where $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ and the vector $\mu \in \mathbb{R}$), is multi-normally distributed with mean vector μ and covariance matrix Σ.
The matrix A can be easily determined with the Cholesky decomposition of Σ. This decomposition is the unique lower-triangular matrix L such as $LL^T = \Sigma$. Since we have seen in Section 4 that the Gaussian (or Normal) copula is the copula of the multivariate normal distribution, one can generate random variates from the n-dimensional Gaussian copula by running the following algorithm:

- Find the Cholesky decomposition $\mathbf{A}$ of the matrix Σ;
- Simulate n independent standard normal random variates $z = (z_1, \ldots, z_n)^T$;
- Set $\mathbf{x} = \mathbf{Az}$;
- Determine the components $u_i = \phi(x_i), i = 1, \ldots, n$, where $\phi(\cdot)$ is standard univariate Gaussian cumulative distribution function.

The vector $(u_1, \ldots, u_n)^T$ is a random variate from the n-dimensional Gaussian copula (4.4).

A.2 Student's T Copula

We have shown in Section 4 that the copula of the multivariate Student's T-distribution is the Student's T-copula. Let $\mathbf{X}$ be a vector with an n-variate standardized Students T-distribution with ν degrees of freedom, and covariance matrix $\frac{\nu}{\nu-2}\Sigma$ (for $\nu > 2$). This vector can be represented in the following way:

$$\mathbf{X} = \frac{\sqrt{\nu}}{\sqrt{S}}\mathbf{Y}$$

where $S \sim \chi_\nu$ and the random vector $\mathbf{Y} \sim MN(0, \Sigma)$ are independent.
Hence, we can use the following algorithm to simulate random variates from the Students T-copula (4.5):

- Find the Cholesky decomposition $\mathbf{A}$ of the matrix Σ;
- Simulate n independent standard normal random variates $z = (z_1, \ldots, z_n)^T$;
- Simulate a random variate s, from the χ_ν distribution, independent of z;
- Determine the vector $\mathbf{y} = \mathbf{Az}$;
- Set $\mathbf{x} = \frac{\sqrt{\nu}}{\sqrt{s}}\mathbf{y}$;
- Determine the components $u_i = t_\nu(x_i), i = 1, \ldots, n$, where $t_\nu(\cdot)$ is standard univariate Student's T cumulative distribution function.

The vector $(u_1, \ldots, u_n)^T$ is a random variate from the n-dimensional T-copula (4.5).

Appendix 2 "The effects of misspecified marginals and copulas on computing the value at risk: A Monte Carlo study."

The effects of misspecified marginals and copulas on computing the value at risk: A Monte Carlo study. *Computational Statistics & Data Analysis*, 53(6), 2168-2188. DOI: https://doi.org/10.1016/j.csda.2008.02.002

Permisson to post the Publisher's version/PDF: according to https://www.elsevier.com/about/policies/sharing "*Theses and dissertations which contain embedded Published Journal Article (PJA) as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.*"

Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 53 (2009) 2168–2188
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

The effects of misspecified marginals and copulas on computing the value at risk: A Monte Carlo study

Dean Fantazzini*

Moscow School of Economics, Moscow State University, Russia

Available online 8 February 2008

Abstract

The effect on the estimation of the Value at Risk when dealing with multivariate portfolios when there is a misspecification both in the marginals and in the copulas is investigated. It is first shown that, when there is skewness in the data and symmetric marginals are used, the estimated elliptical (normal or t) copula correlations are negatively biased, reaching values as high as 70% of the true values. Besides, the bias almost doubles if negative correlations are considered, compared to positive correlations. As for the t copula degrees of freedom parameter, the use of wrong marginals delivers large positive biases, instead. If the dependence structure is represented by a copula which is not elliptical, e.g. the Clayton copula, the effects of marginal misspecifications on the copula parameter estimation can be rather different, depending on the sign of marginal skewness. Extensive Monte Carlo studies then show that the misspecifications in the marginal volatility equation more than offset the biases in copula parameters when VaR forecasting is of concern, small samples are considered and the data are leptokurtic. The biases in the volatility parameters are much smaller, whereas those ones in the copula parameters remain almost unchanged or even increase when the sample dimension increases. In this case, copula misspecifications do play a role for VaR estimation. However, these effects depend heavily on the sign of the dependence.
© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Value at Risk (VaR) has become one of the most popular risk measures since it was recommended and adopted by the Bank of International Settlements and USA regulatory agencies in 1988. Jorion (2000) provides an introduction to Value at Risk as well as discussing its estimation, while the www.gloriamundi.org website comprehensively cites the Value at Risk literature as well as providing other VaR resources. Recently, Ané and Kharoubi (2003), Junker and May (2005) and Fantazzini (in press) investigated the influence of different bivariate joint distribution assumptions on the Value at Risk. They found that the best results can be achieved by models allowing for heavily tailed marginals and a modelling of tail dependency and asymmetric tail dependency. However, they also found out that, for short positions, a multivariate normal model with dynamic normal marginals and constant normal copula can be a proper choice.

* Corresponding address: Moscow School of Economics, Moscow State University, 1, Building 61, Leninskie Gory, 119992 M.V. Lomonosov MSU, Moscow, Russia. Tel.: +7 495 5105267; fax: +7 495 5105256.
E-mail addresses: fantazzini@mse-msu.ru, deanfa@eco.unipv.it.

0167-9473/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.02.002

The aim of this paper is to investigate how misspecifications both in the marginals and in the copulas can affect the estimation of the Value at Risk when dealing with multivariate portfolios. Given the partitioning of the multivariate parameter vector into separate parameters for each margin and parameters for the copula, it has become usual practice to break up the optimization problem into several small optimizations, each with fewer parameters. This multi-step procedure is known as the method of Inference Functions for Margins (IFM); e.g., see Patton (2004), McNeil et al. (2005), Patton (2006a), Cherubini et al. (2004), Gunky et al. (2007) and Karlis and Nikoloulopoulos (2008). Therefore, in order to achieve our aim, we first analyze the effects of misspecified marginals on the multivariate parameter vector estimation, while we examine in a second step the effects of model misspecification on the computation of the Value at Risk. Regarding the first point, we find the interesting result that, when there is skewness in the data and symmetric marginals are used, the estimated normal copula correlations are negatively biased, and the bias increases when moving from the Student's t to the normal distribution, reaching values as high as 27% of the true values. We find that the bias almost doubles if negative correlations are considered, compared to positive correlations. When the true dependence function is represented by the t copula, the choice of the marginals tends to have much stronger effects on copula parameter estimation, with biases up to 50% of the true values for the correlations and up to 380% for the t copula degrees of freedom parameter. If the dependence structure is represented by a copula which is not elliptical, e.g. the Clayton copula, the effects of marginal misspecifications on the copula parameter estimation can be rather different, depending on the sign of the marginal skewness.

The second contribution of the paper is an extensive Monte Carlo study carried out to assess the potential impact of both misspecified margins and misspecified copulas on the estimation of multivariate VaR for equally weighted portfolios. We find that, when small samples are considered and the data are leptokurtic, the overestimation/underestimation in the marginal volatility parameters is so large as to offset the negative biases in copula parameters, thus delivering very conservative VaR estimates for all quantiles. This is true for all multivariate models considered in the analysis.

When the sample dimension increases, the biases in the volatility parameters are much smaller, whereas those in the copula parameters remain almost unchanged or even increase, like for the t copula and the Clayton copula parameters, when (wrong) symmetric marginals are used instead of skewed ones. In this case, copula misspecifications do play a role. However, these effects depend heavily on the sign of the dependence: if it is negative, the bias can be as large as 70%, like for t copula correlations, for instance. If it is positive, the bias is much smaller (10% or less for the t copula correlations), and the effects on quantile estimation are much more limited, if not completely offset by marginal misspecifications. Therefore, this Monte Carlo evidence gives some insights into why previous empirical literature found that the influence of a misspecification in the copula is given with 20% or less of the whole estimation error for the VaR; see e.g. Ané and Kharoubi (2003) and Junker and May (2005).

Finally, we perform an empirical analysis with ten trivariate portfolios, where we quantify the risk of the portfolio under different joint distribution assumptions. The rest of the paper is organized as follows. Section 2 presents the copula–VAR–GARCH models, while in Section 3 we perform simulation studies to assess the finite sample properties of these models under different DGPs. In Section 4, we conduct an empirical analysis with ten portfolios composed of stocks quoted at the NYSE, while we conclude in Section 5.

2. Copula–VAR–GARCH modelling

Consider a general copula–vector autoregression model, where the n endogenous variables $x_{i,t}$ are explained by an intercept μ_i, autoregressive terms of order p, and an error term $\sqrt{h_{i,t}}\eta_{i,t}$:

$$\begin{aligned} x_{1,t} &= \mu_1 + \sum_{i=1,}^{n}\sum_{l=1}^{p}\phi_{1,i,l}x_{i,t-l} + \sqrt{h_{1,t}}\,\eta_{1,t} \\ &\vdots \\ x_{n,t} &= \mu_n + \sum_{i=1,}^{n}\sum_{l=1}^{p}\phi_{n,i,l}x_{i,t-l} + \sqrt{h_{n,t}}\eta_{n,t}. \end{aligned} \tag{1}$$

Let the standardized innovations $\eta_{i,t}$ have mean zero and variance one, while $\sqrt{h_{i,t}}$ can be constant or time varying like in GARCH(1, 1) models; see Bollerslev et al. (1994) for a detailed survey of GARCH models:

$$\begin{aligned} h_{1,t} &= \omega_1 + \alpha_1 \left(\eta_{1,t-1}\sqrt{h_{1,t-1}}\right)^2 + \beta_1 h_{1,t-1} \\ &\vdots \\ h_{n,t} &= \omega_n + \alpha_n \left(\eta_{n,t-1}\sqrt{h_{n,t-1}}\right)^2 + \beta_n h_{n,t-1}. \end{aligned} \tag{2}$$

Furthermore, the $\eta_{i,t}$ have a conditional joint distribution $H_t\left(\eta_{1,t}, \ldots, \eta_{n,t}; \theta\right)$ with parameters vector θ. Copula theory provides an easy way to deal with the (otherwise) complex multivariate modelling. The essential idea of the copula approach is that a joint distribution can be factored into the marginals and a dependence function called a copula. The joint distribution $H_t\left(\eta_{1,t}, \ldots, \eta_{n,t}; \theta\right)$ can be expressed as follows, thanks to the so-called Sklar theorem (1959):

$$\left(\eta_{1,t}, \ldots, \eta_{n,t}\right) \sim H_t\left(\eta_{1,t}, \ldots, \eta_{n,t}; \boldsymbol{\theta}\right) = C_t(F_{1,t}(\eta_{1,t}; \delta_1), \ldots, F_{n,t}(\eta_{n,t}; \delta_n); \boldsymbol{\gamma}), \tag{3}$$

that is the joint distribution H_t of a vector of innovations $\eta_{i,t}$ is the copula $C_t\left(\cdot\,; \gamma\right)$ of the cumulative distribution functions of the innovation marginals $F_{1,t}(\eta_{1,t}; \delta_1), \ldots, F_{n,t}(\eta_{n,t}; \delta_n)$, where $\gamma, \delta_1, \ldots, \delta_n$ are the copula and marginal parameters, respectively. The study of copulas originated with the seminal papers by Sklar (1959) and has seen various applications in the statistics and financial literature. For more details, we refer the interested reader to the recent methodological overviews by Joe (1997) and Nelsen (1999), while Cherubini et al. (2004) provide a detailed discussion of copula techniques for financial applications. It follows from (3) that the log-likelihood function for the joint conditional distribution $H_t(\,\cdot\,; \theta)$ is given by

$$l(\theta) = \sum_{t=1}^{T} \log(c(F_1(x_{1,t}; \delta_1), \ldots, F_n(x_{n,t}; \delta_n); \boldsymbol{\gamma})) + \sum_{t=1}^{T}\sum_{i=1}^{n} \log f_i(x_{i,t}; \delta_{i,t}), \tag{4}$$

where c is the copula density function, whereas f_i are the marginal densities. Hence, the log likelihood of the joint distribution is just the sum of the log likelihoods of the margins and the log likelihood of the copula. Standard ML estimates may be obtained by maximizing the above expression with respect to the parameters $(\delta_1, \ldots, \delta_n; \gamma)$. In practice this can involve a large numerical optimization problem with many parameters which can be difficult to solve. However, given the partitioning of the parameter vector into separate parameters for each margin and parameters for the copula, one may use (4) to break up the optimization problem into several small optimizations, each with fewer parameters. This multi-step procedure is known as the method of Inference Functions for Margins (*IFM*); see Joe and Xu (1996) and Joe (1997). Joe (1997) compares the efficiency of the IFM method relative to full maximum likelihood for a number of multivariate models and finds the IFM method to be highly efficient. Therefore, we think it safe to use the IFM method and benefit from the huge reduction in complexity that it implies for the numerical optimization.

3. Simulation studies

In this section we present the results of the simulation studies concerning the following specification for a trivariate copula–VAR(1)–GARCH(1, 1) model:

$$Y_t = \mu + \Phi_Y^{(1)} Y_{t-1} + \sqrt{h_t}\eta_t, \tag{5}$$

where the matrix h is diagonal and contains the variances:

$$h_{i,t} = \omega_i + \alpha_i(\eta_{i,t-1}\sqrt{h_{i,t-1}})^2 + \beta_i h_{i,t-1}, \quad i = 1, 2, 3,$$

and where we consider the following possible Data Generating Processes (DGPs):

(1) We examine six different types of marginals for the innovations η_t:
- A Generalized t (see Hansen (1994)) with skewness parameter $\lambda = -0.5$ and degrees of freedom $\nu = 3$.

- A Generalized t with skewness parameter $\lambda = 0$ and degrees of freedom $\nu = 3$, that is a symmetric Student's t distribution.
- A Generalized t with skewness parameter $\lambda = 0.5$ and degrees of freedom $\nu = 3$.
- A Generalized t with skewness parameter $\lambda = -0.5$ and degrees of freedom $\nu = 10$.
- A Generalized t with skewness parameter $\lambda = 0$ and degrees of freedom $\nu = 10$, that is a symmetric Student's t distribution close to a standard normal distribution.
- A Generalized t with skewness parameter $\lambda = 0.5$ and degrees of freedom $\nu = 10$.

(2) We examine three types of copulas to model the joint dependence (3) of the innovations η_t:
- A normal copula with correlation matrix
$$\Sigma = \begin{pmatrix} 1 & -0.5 & 0 \\ -0.5 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix}.$$
- A Student's t copula with correlation matrix
$$\Sigma = \begin{pmatrix} 1 & -0.5 & 0 \\ -0.5 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix},$$
and degrees of freedom $\nu_c = 3$.
- A Clayton copula with parameter $\alpha_c = 3$.

We chose the normal and t copulas given their widespread use in financial applications (see e.g. Cherubini et al. (2004) and McNeil et al. (2005)), whereas we chose the Clayton copula because it has positive lower tail dependence and a simulation algorithm that can be easily implemented even for highly dimensional portfolios, unlike all other Archimedean copulas. We do not consider time-varying copulas in this work to keep the number of simulated DGPs tractable. Besides, recent literature (see Chen et al. (2004) and Patton (2004)) has shown that a simple normal copula with no dynamics is sufficient for describing the joint dependence structure for financial data in most cases. Only when the number of variables considered is higher than 20 do statistically significant differences start to emerge and the (constant) Student's t copula become required. Fantazzini et al. (in press) found similar evidence with monthly operational risk data, too.

(3) We consider two possible data situations: $T = 500$, $T = 2000$.

(4) We consider the same conditional mean and variance specification for all the three variables:
$$\mu = \begin{bmatrix} 0.15 \\ 0.15 \\ 0.15 \end{bmatrix}, \qquad \Phi_Y^{(1)} = \begin{bmatrix} 0.5 & -0.2 & 0.15 \\ 0.5 & -0.2 & 0.15 \\ 0.5 & -0.2 & 0.15 \end{bmatrix}, \qquad \omega = \begin{bmatrix} 0.01 \\ 0.01 \\ 0.01 \end{bmatrix},$$
$$\alpha = \begin{bmatrix} 0.05 \\ 0.05 \\ 0.05 \end{bmatrix}, \qquad \beta = \begin{bmatrix} 0.9 \\ 0.9 \\ 0.9 \end{bmatrix}.$$

We justify this choice as made for the sake of simplicity and again to keep the number of simulated DGPs tractable. However, we chose values to mimic the most common stylized facts of financial markets, such as strong persistence in conditional variances; see, for example, Tsay (2002) and references therein. We also tried other specifications for $\Phi_Y^{(1)}$, but the simulation results were robust with respect to this choice.

We generated 1000 Monte Carlo samples for each marginal and copula specification previously described and we then estimated the nine models reported in Table 1. We first analyze in the next two subsections the effects of misspecified marginals on marginal and copula parameter estimation, since they will later help us explain the effects of model misspecification on the computation of the Value at Risk.

3.1. *Effects of marginal misspecifications on marginal parameter estimation*

We report here only the main findings, while the full set of results are reported in the technical report by Fantazzini (2007) in Tables 3–8. The simulation studies show some interesting results:

- ***Normal*** **[Model (1)–Model (4)–Model (7)]**:
  . *Conditional mean*: The estimated parameters do not show any particular bias or trend.

Table 1
Multivariate distribution specifications

MODEL	**Marg.**	**Copula**	MODEL	**Marg.s**	**Copula**	MODEL	**Marg.**	**Copula**
Model (1)	Normal	Normal	**Model (4)**	Normal	t copula	**Model (7)**	Normal	Clayton
Model (2)	St. t	Normal	**Model (5)**	St. t	t copula	**Model (8)**	St. t	Clayton
Model (3)	Gen. t	Normal	**Model (6)**	Gen. t	t copula	**Model (9)**	Gen. t	Clayton

. *Conditional variance*: The lack of a parameter modelling the fat tails induces a large positive bias in the parameters ω and negative ones in α and β. The mean biases of these parameters can be larger than 110%, 50% and 10%, respectively, when $T = 500$, $\nu = 3$ and the data are skewed, while they are slightly smaller when the data are symmetric. Instead, if we consider the median estimates, the biases are much smaller: these differences will later explain why mean and median VaR estimates are very different when small samples are considered. These results indicate the difficulties of estimating GARCH models with small samples, thus confirming previous simulation evidence from Hwang and Valls Pereira (2006) which, however, considered only univariate models with normally distributed errors and did not examine the effect of different joint distributions. Moreover, the strong biases were expected, since it has been shown by Newey and Steigerwald (1997) in a theoretical work that a QML estimator can be biased when data are not symmetric.

. *Sample dimension*: When the dimension T increases, the biases and the RMSE decrease for all parameters.

. *Copula dependence structure*: No qualitative differences are found across different types of dependence functions.

- ***Student's t*** **[Model (2)–Model (5)–Model (8)]**:

. *Conditional mean*: If the true DGP has skewed marginals but we use the Student's t, the biases in the conditional mean parameters can be as high as 45% of the true values for the parameter μ, when $T = 500$ and $\nu = 3$. Instead, no differences are found between left skewed and right skewed data.

. *Conditional variance*: The use of this symmetric distribution causes severe biases in the ω and α estimates, when $T = 500$ and $\nu = 3$ and the data are skewed: up to 90% and 25% of the true values for the former two parameters, respectively. Instead, the parameter β is estimated much more precisely, with biases smaller than 7%. However, like for the normal distribution, the median estimates are more precise: up to 30% and 8% for the ω and α, whereas the negative bias for β decreases to 2%.

. *Degrees of freedom*: When the data are skewed, there is a negative bias that increases both with the sample dimension and the magnitude of ν. It can be higher than 25% of the true value when $T = 2000$.

. *Sample dimension*: Except for the degrees of freedom previously discussed, when the dimension T increases, the biases and the RMSE decrease as well.

. *Copula dependence structure*: No qualitative differences are found across different types of dependence functions.

- ***Generalized t*** **[Model (3)–Model (6)–Model (9)]**:

. *Conditional mean*: The estimated parameters of the conditional mean do not show any particular bias or trend.

. *Conditional variance*: When there is skewness in the data, the estimated parameters are estimated more precisely than when the observations are symmetric: in the latter case, the constant ω shows a strong positive bias which can be higher than 20% (median) when $T = 500$, while in the former case the bias is around 10% (median). These differences are even larger if mean estimates are considered. This difference can be explained considering that the Generalized t is not the most efficient model when the data are symmetric. However, these positive biases are present for all the distributions considered, and clearly highlight how difficult estimating a GARCH model for small samples is.

. *Degrees of freedom*: As expected, the degrees of freedom ν are estimated much more precisely when they are low than vice versa (positive mean bias around 35% when $T = 500$), given that this parameter is much more difficult to identify when it is high in magnitude. If the sample dimension increases, the biases and the RMSE decrease as expected.

. *Skewness parameter*: The estimates show a negative mean bias around 15% for small samples with fat tails, i.e. $\nu = 3$, which disappears as the sample dimension increases. Small samples and thinner tails, i.e. $\nu = 10$, mitigate the bias to 10%. Similar but smaller biases hold when the median is considered.

. *Sample dimension*: When the dimension T increases, the biases and the RMSE decrease for all parameters.

. *Copula dependence structure*: No difference is found across different types of dependence functions: this result was expected since the Generalized t is the true marginal specification and, in this case, its parameters are variation free with respect to the copula parameters (see Patton (2006a,b)).

3.2. Effects of marginal misspecifications on copula parameter estimation

The full set of results for this part are reported in the technical report by Fantazzini (2007) in Tables 9–14. The main findings can be summarized as follows:

- ***Normal copula***:
  . *Correlation matrix*: When there is skewness in the data and symmetric marginals are used, the estimated correlations are negatively biased, and the bias increases when moving from the Student's t to the normal distribution, reaching values as high as 27% of the true values. It is interesting to note that the bias almost doubles if negative correlations (-0.5) are considered, compared to the positive correlation case. This result remains the same both with positively skewed and with negatively skewed data, while no biases are found if the variables are uncorrelated. If the true Generalized t is used, the estimation biases are equal to or smaller than 1% even with $T = 500$. No differences between the mean and the median estimates are found. When the degrees of freedom of the true marginals increase from 3 to 10, the biases in the correlations decrease as expected, but the previous conclusions remain unchanged: the biases for negative correlations are double compared to those for positive correlations, and increase when the used marginals are normally distributed.
  . *Sample dimension*: No major differences in the biases are found when moving from $T = 500$ to $T = 2000$, even though the RMSE are smaller.
- ***T copula***:
  . *Correlation matrix*: When the true dependence function is represented by the t copula, the choice of the marginals tends to have much stronger effects on copula parameter estimation. If the true DGP has skewed marginals but we use symmetric ones, instead, the biases in the correlations can be as high as 15% of the true values for the Student's t and 50% when using the normal distribution and $T = 500$. The difference between negative (-0.5) and positive (0.5) correlations is even more striking here: the bias in the former case is, in general, five times the bias in the latter case. Interestingly, if the true correlations are negative and normal marginals are used, then the bias tends to increase with the sample dimension, while this effect is weaker or negligible with positive correlations. Instead, if Student's t distributions are used, the biases tend to decrease in all cases as the sample dimension increases.

    Differently from what we saw with the normal copula, the use of normal marginals when the true ones are skewed induces positive correlation between variables that are uncorrelated. This effect is much smaller if Student's t marginals are used, instead. Furthermore, even if we use the true Generalized t, nevertheless the correlations can be negatively biased in small samples.

    Similarly to the case for the normal copula, when the degrees of freedom of the true marginals increase from 3 to 10, the biases in the correlations decrease, but the previous qualitative conclusions remain unchanged. However, we also note that the biases tend to remain as high as 50% when using the normal distribution for the marginals, if negative correlations are of concern.
  . *t copula degrees of freedom* ν_c: The use of wrong marginals can have dramatic effects, with positive biases ranging from 60% for the Student's t distribution up to 380% for the normal distribution, when the true DGPs are skewed and $T = 500$. Besides, we note that the estimates can be poor for small samples also when using the Generalized t, with biases around 5% of the true values. Instead, when the sample dimension increases and the true marginal degrees of freedom increase to 10, the estimation bias decreases in all cases. However, it still remains as high as 40% for the symmetric Student's t and over 100% for the normal distributions.
- ***Clayton copula***: This copula is not elliptical and the effects of marginal misspecifications on the copula parameter estimation can be rather different, depending on the sign of marginal skewness.
  . If the *true marginals show negative skewness*, then using symmetric marginals causes the Clayton parameter α_c to be negatively biased, where the larger bias is caused by the normal distribution. Increasing the sample dimension to $T = 2000$ slightly decreases these biases, but they still remain higher than 20% for the normal distribution and 17% for the symmetric student's t. Remarkably, if we increase the marginal degrees of freedom of the true DGPs to 10, the biases still remain higher than 15% for all the symmetric distributions.

. If the *true marginals are symmetric*, the estimation biases are smaller than 5% for all distributions and sample dimensions.
. If the *true marginals show positive skewness*, then using symmetric marginals causes the Clayton parameter α_c to be positively biased, where the larger bias is again caused by the normal distribution. Interestingly, increasing the sample dimension to $T = 2000$ increases the estimation bias, which can reach 35% of the true value for the normal distribution. If we increase the degrees of freedom of the true DGPs to 10, the biases still remain higher than 5% when $T = 500$, and 10% when $T = 2000$, for all the symmetric distributions.

3.3. *Effects of marginal and copula misspecifications on multivariate VaR estimation*

Due to the large growth of trading activity and the well known trading loss of financial institutions, the regulators and supervisory committees of banks have developed and supported quantitative techniques in order to evaluate the possible losses that these institutions can incur. The most well known risk measure is the Value at Risk (VaR), which is defined as the maximum loss which can be incurred by a portfolio, at a given time horizon and at a given confidence level $(1 - p)$. While we remark that Value at Risk as a risk measure is criticized for not being sub-additive and hence may fail to stimulate diversification (see Embrechts (2000) for an overview of the criticism), still, the final Basel Capital Accord that came into force in 2007 focuses on VaR only. Therefore, we explore here the consequences of marginal and copula misspecifications on VaR estimation, by using the same DGPs as were discussed at the beginning of Section 3. For the sake of simplicity, we suppose investment of an amount $M_i = 1, i = 1, \ldots, n = 3$, in every asset, and we consider an equally weighted portfolio. We make this choice for the sake of interest, since it represents the most common case in the financial literature (see, e.g., Junker and May (2005)), and because DeMiguel et al. (2007) have recently shown that a wide range of models are not consistently better than the $1/n$ rule in terms of Sharpe ratio, certainty-equivalent return, or turnover, indicating that, out of the sample, the gain from optimal diversification is more than offset by the estimation error.

We consider eight different quantiles to better highlight the overall effects of the misspecifications on the joint distribution of the losses: 0.25%, 0.50%, 1.00%, 5.00%, 95.00%, 99.00%, 99.50%, 99.75%, that is we consider both the "loss tail" and the "win tail". The full set of results for this part are reported in Fantazzini (2007), Tables 15–32. In general, the estimated quantiles show a very large degree of underestimation/overestimation, which depends heavily on both the sample dimension and the underlying joint distribution. The major findings are reported below.

- **True dependence function: Normal copula**.
  . *True marginals: Skewed marginals*
  Fat tails: $\nu = 3$. If we consider the mean VaR across Monte Carlo samples, all empirical distributions used in the analysis show overestimated VaRs, ranging from 5%–100% of the true value for the Generalized t/normal copula, to 150%–2800% for the models which use the normal distribution for the marginals, whose RMSE are usually three times the RMSE of the models which use the Generalized t and two times that of the models which employ the Student's t. Instead, if we consider the median VaR, the use of symmetric distributions now involves negative biases in the skewed tail and positive ones the other way round, ranging between 20% and 30% for the symmetric t and between 40% and 60% for the normal distribution (central quantiles 5%–95% excluded). Instead, the Generalized t marginal distribution is conservative for the skewed tail and aggressive for the other tail, with biases ranging between 5% and 20%.

  These results can be explained by looking at Tables 3–8 in Fantazzini (2007), which shows that the marginals' parameters of the conditional variance parameters ω and α are overestimated and β is underestimated for all marginal distributions, and the biases increase when moving from the Generalized t to the normal distribution. Moreover, as previously discussed in Section 3.1, when $T = 500$, there is a large difference between the estimated mean and median marginal parameters due to the poor estimates of GARCH models for within small samples. It is interesting to note that these biases more than offset the negative biases in the t copula or normal copula estimated correlations which would decrease the estimated VaR, and which do not show any substantial difference between mean and median estimates, instead, as reported in Tables 9–14 in Fantazzini (2007).

  When $T = 2000$, the biases and especially the RMSE are much smaller compared to the case for $T = 500$ for all models considered, (1)–(9). The improvement is more evident for extreme quantiles than for central quantiles. The higher precision is due to less noisy marginal estimates, since the biases in the correlations remain almost unchanged, as one can see by looking at Tables 9–14 in Fantazzini (2007). Even though the Generalized

t/normal copula is the most efficient model (as expected), the differences among copulas are not relevant when the correct marginal distribution are chosen. In particular, if the extreme quantiles in left tail are of concern, the VaR estimates delivered by the Clayton copula are very close to those of the normal copula.

Thin tails: $\nu = 10$. The biases and the RMSE are much smaller than in the case with fat tails. This result is due to the much smaller bias (and RMSE as well) in the conditional variance parameter α, which now ranges over 10%–18% instead of 20%–50% for the case with fatter tails (while the biases in the other parameters are almost unchanged). Besides, there is no major differences between the mean and median VaR estimates.

When $T = 2000$, the biases and the RMSE for correct joint distributions decrease (but the sign remains unchanged), while for misspecified models this is true for the RMSE only: the biases can actually increase. All the previous qualitative comments regarding the type of copula remain valid, even though the differences are now smaller.

. *True marginals: Symmetric marginals*

Fat tails: $\nu = 3$. As expected, the results for the Generalized t and the symmetric t are very similar, but the latter shows smaller RMSE since it is the most efficient model. As for normal marginals, the biases range over $-10\%/-40\%$ (central quantiles excluded) and they are stable across different samples.

Thin tails: $\nu = 10$. Similarly to the case with skewed marginals, the biases and the RMSE are much smaller than in the case with fat tails, due to the much smaller bias in the conditional variance parameter α, which now ranges over 12%–14% instead of 25%–40% as for the case with fatter tails.

- **True dependence function: *t* copula**.

. *True marginals: Skewed marginals*

Fat tails: $\nu = 3$. If we consider the mean VaR across Monte Carlo samples together with $T = 500$, all empirical distributions used in the analysis show overestimated VaRs. They range from 7%–80% of the Generalized t/t copula to 70%–4300% for the models which use the normal for the marginal distributions, whose RMSE are usually three/four times the RMSE of the models which use the Generalized t and two times the RMSE of the models which employ the Student's t. Instead, if we consider the median VaR, the use of symmetric distributions again involves negative biases in the skewed tail and positive ones vice versa, ranging between 20% and 30% for the symmetric t and between 40% and 60% for the normal distribution (central quantiles 5%–95% excluded). Furthermore, the Generalized t marginal distribution is conservative for the skewed tail, 0%–10%, and aggressive for the other tail, with negative biases ranging between 5% and 25%. Similarly to the case with the normal copula, these results follow from the overestimation of the conditional variance parameters ω and α and the underestimation of β, and their biases increase when moving from the Generalized t to the normal distribution. As expected, for symmetric distributions, the biases are smaller when considering the skewed tail, and larger vice versa. Furthermore, the more extreme quantiles are much better estimated than the central quantiles.

For when $T = 2000$, two results are worth noting: model (9), that is the Generalized t/Clayton copula, does much better than the Generalized t/normal copula or the Generalized t/t copula when extreme quantiles are of concern, thus confirming similar evidence with $T = 500$. However, differently from the case for $T = 500$, the mean and the median VaR estimates are much more closer and they clearly highlight that the use of symmetric distributions involves negative biases in the skewed tail and positive ones vice versa. In particular, we observe that the median VaR biases remain the same as with $T = 500$ even though the marginal parameters are now much more precisely estimated: This result is due to the severe underestimation of the t copula correlation parameters (-70% with normal marginals) and overestimation of the degrees of freedom parameter ($+300\%$ with normal marginals) when symmetric marginals are used, as reported in Table 11 in Fantazzini (2007).

Thin tails: $\nu = 10$. The biases and the RMSE are much smaller than in the case with fat tails. This result is due to the much smaller bias (and RMSE as well) in the conditional variance parameter α, which now ranges over 7%–18% instead of 20%–65% as for the case with fatter tails. Furthermore, the mean and median VaR estimates are now very close. Moreover, the use of the normal copula when the true one is the t copula delivers negative biases for almost all quantiles. The results that emerged with $T = 500$ hold also with $T = 2000$ with no major qualitative differences.

. *True marginals: Symmetric marginals*

Fat tails: $\nu = 3$. As expected, the results for the Generalized t/t copula and the symmetric t/t copula are very similar, but the latter shows smaller RMSE since it is the most efficient model. This is not true when the Clayton or the normal copula is used: in this case, the use of the Generalized t marginal delivers more precise estimates

of the quantiles than the symmetric t or the normal distributions. Besides, if we look at the median VaR, the use of the normal copula delivers underestimated VaR for almost all quantiles.

Thin tails: $\nu = 10$. Similarly to the case with skewed marginals, the biases and the RMSE are much smaller than in the case with fat tails, due to the much smaller bias in the conditional variance parameter α, which now ranges over 6%–15% instead of 23%–43% as for the case with fatter tails. Again, the use of the normal copula involves underestimated VaR.

When $T = 2000$, the previous comments about the negative biases delivered by the normal copula and the normal distributions for extreme quantiles remain valid here. Furthermore, we observe that the Student's t/Clayton copula model delivers the most precise estimates when $\nu = 3$, while the Student's t/t copula model delivers the most precise estimates when $\nu = 10$.

- **True dependence function: Clayton copula**.

. *True marginals: Skewed marginals*

Fat tails: $\nu = 3$. If we consider the mean VaR across the Monte Carlo samples together with $T = 500$, almost all the empirical distributions used in the analysis show overestimated VaRs. They range from 3%–100% for the Generalized t/Clayton copula, up to 30%–600% for the models which use the normal for the marginal distributions, whose RMSE are usually three times the RMSE of the models which use the Generalized t and two times the RMSE of the models which employ the Student's t. A notable exception to this pattern is represented by the models which include the Generalized t as marginal distribution (i.e. models (3), (6) and (9)): when the data are positively skewed and the left tail is of concern, these models show a negative bias ranging between −5% and −20%. Instead, if we consider the median VaR, the use of symmetric distributions involves negative biases in the skewed tail and positive ones the other way round, similarly to what we observed when the true dependence function was represented by the normal or the t copula. Furthermore, the Generalized t marginal distribution is conservative for the skewed tail, and aggressive for the other tail. The previous comments regarding the importance of the conditional variance parameters ω and α and the minor role played by the type of copula remain unchanged. However, a major difference that we can observe when the true dependence function is modelled by a Clayton copula is the much smaller biases, RMSE and RRMSE for all models considered, (1)–(9), the normal distribution included. This result probably follows from the simpler dependence structure implied by this copula compared to the heavily parameterized normal and t copula.

When $T = 2000$, the Generalized t/t copula and Generalized t/Clayton copula models show the most precise estimates when the data are negatively skewed and left quantiles are of concern. The Generalized t/normal copula or the Generalized t/t copula is a better choice if the data are positively skewed and the right quantiles are of concern. Furthermore, we observe again that if we use symmetric marginals, we get more aggressive VaR estimates for the left quantiles if the data are negatively skewed, whereas we get more conservative estimates if the data are positively skewed. This result follows from the different bias in the Clayton copula parameter when misspecified marginals are used: a negative bias around 20% in the former case, a positive bias around 30% in the latter case; see Table 13 in Fantazzini (2007).

Thin tails: $\nu = 10$. The biases and the RMSE are much smaller than in the case with fat tails. This result is due to the much smaller bias (and RMSE as well) in the conditional variance parameter α, which now ranges over 8%–19% instead of 20%–55% as for the case with fatter tails. Again, the biases in the correlations and, in general, the choice of the copula play a residual role when $T = 500$.

Finally, when $T = 2000$, the Generalized t/Clayton copula and the Generalized t/t copula are the models with lowest RMSE, but the latter show biases around zero, while the former show positive biases between 5% and 10%, that is the VaR estimates are more conservative.

. *True marginals: Symmetric marginals*

Fat tails: $\nu = 3$. As expected, the results for the Generalized t/Clayton copula and the symmetric t/Clayton copula are very similar, but the latter shows smaller RMSE since it is the most efficient model. This is not true when the t copula or the normal copula (for extreme quantiles) is used: the use of the Generalized t marginal delivers more precise estimates of the quantiles.

Thin tails: $\nu = 10$. Similarly to the case with skewed marginals, the biases and the RMSE are much smaller than in the case with fat tails, due to the smaller bias in the conditional variance parameter α, which now ranges over 11%–15% instead of 19%–27% as for the case with fatter tails. The use of the true Clayton copula results in VaR estimates that are almost unbiased (biases smaller than 1%), while the biases for the other two copulas

range between 10% and 20%. This is in sharp contrast with what we saw when the true dependence was given by the t or normal copula, where the differences among copulas were minor. Instead, similarly to the previous cases, the mean and median VaR estimates are now very close.

When $T = 2000$, as expected, the Student's t/Clayton model shows the lowest RMSE and its biases are around zero, differently from the Student's t/normal copula and Student's t/t copula which show negative biases between 5% and 10%, instead. Moreover, the mean and median VaR estimates are close when $\nu = 3$, too, similarly to what we observed with the normal and the t copula.

3.4. Discussion

In order to summarize the previous findings, we report in Table 2 the Min and the Max biases in the mean VaR estimates for all the nine multivariate specifications, across all quantiles. We consider the case for $T = 2000$ for the sake of interest, since this sample dimension will be used in the following empirical analysis. Moreover, we have just shown that the estimated mean and median VaR are very close when using large samples.

The simulation results presented in the previous section indicate that, when small samples are considered and the data are leptokurtic, the overestimation/underestimation in the marginal GARCH parameters is so large as to deliver conservative VaR estimates even with a simple multivariate normal distribution (model (1)). In general, the VaR estimates are very poor and suffer from computational problems when estimating GARCH models for small samples, as also discussed in Hwang and Valls Pereira (2006). When the sample dimension increases, the biases in the volatility parameters are much smaller, whereas those in the copula parameters remain almost unchanged or even increase: see, for example, the t copula and the Clayton copula parameters, when symmetric marginals are used instead of skewed ones. In this case, copula misspecifications do play a role. However, these effects depend heavily on the sign of the dependence: if it is negative, the bias can be as large as 70%, like for t copula correlations, for example. If it is positive, the bias is much smaller (10% or less for the t copula correlations), and the effects on quantile estimation are much more limited, if not completely offset by marginal misspecifications. Therefore, this Monte Carlo evidence gives some insights into why previous empirical literature found that the influence of a misspecification in the copula is given with 20% or less of the whole estimation error for the VaR; see e.g. Ané and Kharoubi (2003) and Junker and May (2005).

4. Empirical analysis

In order to compare the nine multivariate distribution specifications proposed in Table 1 and analyzed in detail in Section 3, we measured the Value at Risk of ten trivariate portfolios. These portfolios are composed of assets from the Dow Jones Industrial Index, with daily data taking into consideration the very volatile period between March 1996 and November 2003. Following Giacomini and Komunjer (2005) and González-Rivera et al. (2006), we use a rolling forecasting scheme of 2000 observations, because it may be more robust against a possible parameter variation. Furthermore, we chose this time dimension because the previous simulation studies in Section 3 clearly highlighted that the biases and the RMSE of the VaR estimates are much smaller and they are close to zero for correctly specified joint distributions. In our case we have 3000 observations so we split the sample in this way: 2000 observations for the estimation window and 1000 for the out-of-sample evaluation. Besides, Christoffersen and Diebold (2000) and Giot and Laurent (2003) showed that volatility forecastability decays quickly with the time horizon of the forecasts. An immediate consequence is that volatility forecastability is relevant for short time horizons (such as daily trading) but not for long time horizons. Therefore, we focused on daily returns and VaR performances for daily trading portfolios, only. A general algorithm for estimating the 0.25%, 0.5%, 1%, 5%, 95%, 99%, 99.5% and 99.75% VaR over a one-day holding period for a portfolio P of n assets with invested positions equal to M_i, $i = 1, \ldots, n$, is the following:

(1) Given a set of estimated parameter values at time $t-1$, simulate 100,000 scenarios for each asset log-return, $\{y_{1,t}, \ldots, y_{n,t}\}$, over the time horizon $[t-1, t]$, using a general multivariate distribution as in Table 1, by using this procedure:
   (a) Firstly, *generate n random variates* $(u_{1,t}, \ldots, u_{n,t})$ *from the copula* $\hat{C}_t$ forecast at time t, which can be normal, Student t, or Clayton.
   (b) Secondly, get a vector $n \times 1\mathbf{Q_t}$ of *standardized asset log-returns* $\eta_{i,t}$ by using the inverse functions of the forecast marginals at time t, which can be symmetric or skewed:

$$\mathbf{Q_t} = \left(\eta_{1,t}, \ldots, \eta_{n,t}\right) = \left(F_1^{-1}(u_{1,t}; \hat{\delta}_1), \ldots, F_n^{-1}(u_{n,t}; \hat{\delta}_n)\right).$$

Table 2
VaR Bias: Min and Max in % when $T = 2000$ across all quantiles

	True copula: Normal copula				**True copula:$t$copula**				**True copula: Clayton**			
	Skewed m.		**Sym. m.**		**Skewed m.**		**Sym. m.**		**Skewed m.**		**Sym. m.**	
	$\nu = 3$	$\nu = 10$	$\nu = 3$	$\nu = 10$	$\nu = 3$	$\nu = 10$	$\nu = 3$	$\nu = 10$	$\nu = 3$	$\nu = 10$	$\nu = 3$	$\nu = 10$
Model (1)	−10/891	−25/31	−9/208	−10/3	12/1454	−39/48	−20/263	−30/10	−25/231	−32/31	−30/77	−23/12
Model (2)	−111/333	−54/76	2/12	0/1	−106/583	−52/95	−4/35	−21/7	−50/103	−32/55	−10/31	−11/23
Model (3)	−39/24	−25/13	6/35	0/1	−47/42	−41/14	−5/33	−21/8	−45/46	−35/36	−9/37	−11/24
Model (4)	−1/992	−24/30	−7/217	−9/2	−42/1496	−22/77	−10/289	−16/5	−32/184	−30/33	−41/32	−23/12
Model (5)	−115/384	−55/75	4/23	0/2	−112/528	−50/98	3/18	−2/0	−52/106	−31/60	−7/38	−11/26
Model (6)	−39/36	−24/14	1/6	0/2	−56/35	−25/14	2/15	−1/0	−44/50	−34/39	−9/32	−11/26
Model (7)	3/1130	−19/46	−9/208	−5/12	0/1918	−28/77	1/375	−16/34	−27/243	−28/23	−30/72	−15/3
Model (8)	−99/438	−46/97	0/26	−1/11	−98/739	−36/125	−3/75	−11/31	−47/100	−28/45	1/3	0/1
Model (9)	−29/34	−17/16	4/33	−1/12	−35/59	−29/32	1/86	−11/31	−36/24	−28/11	1/4	0/0

(c) Thirdly, *rescale the standardized asset log-returns* by using the forecast means and variances, estimated with AR–GARCH models:

$$\{y_{1,t}, \ldots, y_{n,t}\} = \left(\hat{\mu}_{1,t} + \eta_{1,t} \cdot \sqrt{\hat{h}_{1,t}}, \quad \ldots, \quad \hat{\mu}_{n_t} + \eta_{n,t} \cdot \sqrt{\hat{h}_{n,t}}\right). \tag{6}$$

(d) Finally, repeat this procedure $j = 1, \ldots, 100{,}000$ times.

(2) By using these 100,000 scenarios, the portfolio P is re-evaluated at time t; that is,

$$P_t^j = M_{1,t-1} \cdot \exp(y_{1,t}) + \ldots + M_{n,t-1} \cdot \exp(y_{n,t}), \quad j = 1, \ldots, 100{,}000. \tag{7}$$

(3) Portfolio *losses* in each scenario j are then computed (possible profits are considered as negative losses):

$$\text{Loss}_j = P_{t-1} - P_t^j, \quad j = 1, \ldots, 100{,}000. \tag{8}$$

(4) The calculation of the 0.25%, 0.5%, 1%, 5%, 95%, 99%, 99.5% and 99.75% VaR is now straightforward:
   (a) Order the 100,000 Loss_j in increasing order.
   (b) The pth VaR is the $(1-p)\cdot 100{,}000$th ordered scenario, where $p = \{0.25\%, 0.5\%, 1\%, 5\%, 95\%, 99\%, 99.5\%, 99.75\%\}$. For example, the 0.25% VaR is the 99,750th ordered scenario.

The predicted one-step-ahead VaR forecasts were then compared with the observed portfolio losses and both results were recorded for later assessment.

4.1. VaR evaluation

We compare the different multivariate models by looking at their VaR forecasts, using Hansen's (2005) Superior Predictive Ability (SPA) test together with the Giacomini and Komunjer (2005) asymmetric loss function

$$\mathcal{T}_p(e_{t+1}) \equiv (p - \mathbb{1}(e_{t+1} < 0))e_{t+1}, \tag{9}$$

where $e_{t+1} = L_{t+1} - \widehat{\text{VaR}}_{t+1|t}$, $\mathbb{1}$ is the indicator function, L_{t+1} is the realized loss, while $\widehat{\text{VaR}}_{t+1|t}$ is the VaR forecast at time $t+1$ on information available at time t. We also employ Kupiec's (1995) unconditional coverage test and Christoffersen's (1998) conditional coverage test, given their importance in the empirical literature, even though their power can be very low.

4.2. VaR out-of-sample results

The actual VaR exceedances N/T, the p-values p_{UC} of Kupiec's Unconditional Coverage test, and the p-values p_{CC} of Christoffersen's Conditional Coverage test, for the VaR forecasts at all probability levels are reported in Tables 3 and 4. Instead, Tables 5 and 6 report the asymmetric loss functions (9) and the p-values of the SPA test, for all the quantiles and for long and short positions. The previous tables show that the ten portfolios present a wide range of joint distributions. The major insights are the following ones:

- 1st *portfolio: 3M, AT&T, Alcoa*: The symmetric Student's t/Clayton copula model represents the most appropriate choice according to all tests and the asymmetric loss distributions. However, the Generalized t/Clayton copula model cannot be outperformed for all quantiles, too, according to Hansen's tests. The simulation results reported in Table 32 in Fantazzini (2007) show that the use of symmetric marginals together with the true Clayton copula, when the true margins are positively skewed and the marginal degrees of freedom are not too low, results in positive biases for the left quantiles and negative ones for the right quantiles. Instead, the same table shows that the true Generalized t marginals produce smaller biases but with the opposite signs. Therefore, this MC evidence explains why the empirical ratio of VaR exceedances N/T is smaller for the Student's t compared to the Generalized t when dealing with long positions, while the reverse is true for short positions. Besides, the empirical exception frequencies N/T in Table 3 show that the symmetric Student's T/t copula and the symmetric Student's t/normal copula deliver aggressive VaR estimates when dealing with long positions, whereas conservative if short positions are of concern. This evidence can be explained by looking at Tables 20 and 25 in Fantazzini (2007), where simulation results show that when the true dependence structure is represented by the Clayton copula but elliptical copulas are used, the resulting VaR are underestimated for long positions and overestimated for short positions.

- 2nd *portfolio: Altria Group, American Express, Boeing*. The Generalized t/normal copula model turns out to be the best joint distribution, but the Student's t/normal copula performs rather well, too. It is interesting to note that the models using the t copula and the Clayton copula produce conservative VaR estimates for the extreme left quantiles: this is line with the simulation evidence reported in Tables 24 and 30 in Fantazzini (2007), where the use of these copulas when the true one is the normal copula results in VaR estimates which are 5%–10% higher than the true ones, given that the true marginals are symmetric with low degrees of freedom.
- 3rd *portfolio: Caterpillar, Citigroup, Coca Cola*. Like for the previous portfolio, the Student's t/normal copula model is already a proper choice, but the models with the t copula and Clayton copula follow closely. The same comments about left quantiles discussed for the 2nd portfolio apply here, too.
- 4th *portfolio: DuPont, Eastman Kodak, Exxon Mobil*. The Generalized t/t copula, i.e. model (6), seems to be the preferred model according to loss functions and Hansen tests. However, the Kupiec and Christoffersen tests highlight that it is slightly too aggressive when central left quantiles are of concern: such evidence can be explained by looking at Table 25 in Fantazzini (2007), which shows that when there is positive marginal skewness and the true dependence structure is the t copula, model (6) shows negative biases for left quantiles, particularly for central quantiles, and positive ones for right quantiles (even though the RRMSE are very small). For the same D.G.P., the previous table shows that a symmetric Student's t/t copula model delivers higher biases but with opposite signs, instead: this is exactly what we observe in Table 3, where model (5) performs slightly better for left quantiles, but much worse for right quantiles.
- 5th *portfolio: General Electric, General Motors, Hewlett-Packard*. The Student's t/t copula is the model that is able to pass all tests and it shows the lowest loss functions for many quantiles, too. Instead, the Student's t/Clayton performs better if left quantiles are of concern: the Monte Carlo results reported in Table 31 in Fantazzini (2007) show that if the Clayton copula is used instead of the true t copula (and the true marginals show no asymmetry), the former copula delivers overestimated VaR estimates for the left quantiles, whereas it delivers underestimated ones for the right quantiles.
- 6th *portfolio: Home Depot, Honeywell, Intel*. The Student's t/Clayton copula is the best model for long positions, while the Student's t/t copula is preferred for short positions, instead. However, both models cannot be outperformed for all quantiles, according to the Hansen's test. Furthermore, the Generalized t/Clayton model delivers close results, too. This case is very similar to that of the first portfolio and the same comments apply.
- 7th *portfolio: IBM, Intl. Paper, JPMorgan Chase & Co*. Again, the Student's t/Clayton copula is the best model for long positions, while the Student's t/t copula is preferred for short positions, instead. This case is very similar to those of the first and the sixth portfolios and the same comments apply.
- 8th *portfolio: Johnson & Johnson, McDonald's, Merck*. The simple multivariate normal model is already a proper choice, as it passes all tests and its loss functions are very low, if not the lowest. Therefore, it does not come as a surprise that the models using the t copula and the Clayton copula tend to be too conservative, particularly for the left quantiles, thus confirming the MC evidence reported in tables 24 and 30 in Fantazzini (2007).
- 9th *portfolio: Microsoft, Procter & Gamble, SBC*. The Student's t/normal copula model results to be the best choice. This portfolio presents a joint distribution which is very similar to the one estimated for the previous second and third portfolios and the same comments apply.
- 10th *portfolio: United Technologies, Wal-Mart, Walt Disney*. The symmetric Student's t/t copula, i.e. model (5), is the preferred model according to loss functions and statistical tests, whereas the Student's t/Clayton copula model is more conservative for the 0.25% left quantile: this portfolio is therefore similar to the fifth portfolio and the same comments apply.

The previous results indicate that a Generalized t or a symmetric t distribution with a GARCH(1, 1) specification for the variance, together with a Clayton copula, can be a good choice for precise VaR estimates for long positions. A symmetric t/normal copula model is a better choice if short positions are of concern, instead. Finally, a Generalized t/t copula model can be a good compromise if both long and short positions are of interest.

5. Conclusions

This paper investigated how misspecification both in the marginals and in the copulas may affect the estimation of the Value at Risk when dealing with multivariate portfolios. We first performed a Monte Carlo study to assess the potential impact of misspecified margins on the estimation of the copula parameters under different hypotheses for

Table 3
VaR exceedances N/T, Kupiec's and Christoffersen's tests: 1st–5th portfolios

	N/T 0.25%	p_{UC}	p_{CC}	N/T 0.50%	p_{UC}	p_{CC}	N/T 1%	p_{UC}	p_{CC}	N/T 5%	p_{UC}	p_{CC}	N/T 95%	p_{UC}	p_{CC}	N/T 99%	p_{UC}	p_{CC}	N/T 99.5%	p_{UC}	p_{CC}	N/T 99.75%	p_{UC}	p_{CC}
P1M1	1.60%	**0.00**	**0.00**	2.40%	**0.00**	**0.00**	3.40%	**0.00**	**0.00**	8.30%	**0.00**	**0.00**	1.10%	**0.00**	**0.00**	1.60%	**0.00**	**0.00**	2.20%	**0.00**	**0.00**	6.50%	**0.04**	0.08
P1M2	0.40%	0.38	0.67	0.80%	0.22	0.44	1.70%	**0.04**	0.07	6.60%	**0.03**	0.08	0.20%	0.74	0.94	0.50%	1.00	0.07	1.00%	**0.03**	0.90	5.60%	0.39	0.61
P1M3	0.50%	0.16	0.37	1.20%	**0.01**	**0.03**	2.30%	**0.00**	**0.00**	7.40%	**0.00**	**0.00**	0.10%	0.28	0.56	0.40%	0.64	0.07	0.90%	0.75	0.87	4.80%	0.77	0.54
P1M4	1.50%	**0.00**	**0.00**	2.10%	**0.00**	**0.00**	3.30%	**0.00**	**0.00**	8.30%	**0.00**	**0.00**	1.00%	**0.00**	**0.00**	1.50%	**0.00**	0.08	2.10%	**0.00**	**0.00**	6.80%	**0.01**	**0.04**
P1M5	0.40%	0.38	0.67	0.70%	0.40	0.67	1.60%	0.08	0.17	6.70%	**0.02**	0.06	0.10%	0.28	0.56	0.50%	1.00	0.07	1.20%	0.54	0.71	5.60%	0.39	0.61
P1M6	0.40%	0.38	0.67	1.10%	**0.02**	0.06	1.90%	**0.01**	**0.03**	8.00%	**0.00**	**0.00**	0.10%	0.28	0.56	0.20%	0.13	0.08	0.80%	0.51	0.75	5.00%	1.00	0.65
P1M7	1.20%	**0.00**	**0.00**	1.80%	**0.00**	**0.00**	3.30%	0.00	0.00	8.50%	0.00	0.00	1.70%	**0.00**	**0.00**	2.30%	**0.00**	0.09	3.30%	**0.00**	**0.00**	9.00%	**0.00**	**0.00**
P1M8	0.20%	0.74	0.94	0.40%	0.64	0.88	1.10%	0.75	0.84	5.70%	0.32	0.60	0.50%	0.16	0.37	0.70%	0.40	0.06	2.00%	**0.01**	**0.01**	6.30%	0.07	0.03
P1M9	0.20%	0.74	0.94	0.50%	1.00	0.98	1.30%	0.36	0.56	6.80%	**0.01**	0.05	0.40%	0.38	0.67	0.50%	1.00	0.07	1.40%	0.23	0.40	5.90%	0.20	0.19
P2M1	0.70%	**0.02**	0.06	1.00%	0.05	0.13	1.50%	0.14	0.27	6.90%	**0.01**	**0.03**	1.10%	**0.00**	**0.00**	1.20%	**0.01**	**0.03**	1.80%	**0.02**	0.05	5.70%	0.32	0.56
P2M2	0.10%	0.28	0.56	0.20%	0.13	0.31	0.90%	0.75	0.87	6.70%	**0.02**	0.05	0.50%	0.16	0.37	0.50%	1.00	0.07	1.30%	**0.02**	0.24	5.60%	0.39	0.41
P2M3	0.10%	0.28	0.56	0.30%	0.33	0.62	0.90%	0.75	0.87	6.60%	**0.03**	0.07	0.40%	0.38	0.67	0.50%	1.00	0.07	1.10%	0.75	0.84	5.30%	0.67	0.41
P2M4	0.60%	0.06	0.17	1.00%	0.05	0.13	1.60%	0.08	0.17	7.20%	**0.00**	**0.01**	0.90%	**0.00**	**0.01**	1.20%	**0.01**	0.07	1.70%	**0.04**	0.07	5.70%	0.32	0.56
P2M5	0.10%	0.28	0.56	0.10%	**0.03**	0.09	0.90%	0.75	0.87	6.70%	**0.02**	0.05	0.40%	0.38	0.67	0.50%	1.00	0.07	1.20%	0.54	0.71	5.80%	0.26	0.35
P2M6	0.10%	0.28	0.56	0.20%	0.13	0.31	0.90%	0.75	0.87	6.70%	**0.02**	0.05	0.30%	0.76	0.95	0.50%	1.00	0.07	0.80%	0.51	0.75	5.30%	0.67	0.41
P2M7	0.60%	0.06	0.17	1.10%	**0.02**	0.06	1.80%	0.02	0.05	7.90%	**0.00**	**0.00**	1.50%	**0.00**	**0.00**	2.00%	**0.00**	0.08	2.60%	**0.00**	**0.00**	7.30%	**0.00**	**0.01**
P2M8	0.10%	0.28	0.56	0.10%	**0.03**	0.09	0.60%	0.17	0.38	6.40%	0.05	0.12	0.50%	0.16	0.37	1.00%	0.05	0.06	1.50%	0.14	0.16	6.40%	0.05	0.13
P2M9	0.10%	0.28	0.56	0.10%	**0.03**	0.09	0.80%	0.51	0.75	6.60%	**0.03**	0.07	0.50%	0.16	0.37	0.80%	0.22	0.07	1.60%	0.08	0.11	6.10%	0.12	0.24
P3M1	0.60%	0.06	**0.01**	1.30%	**0.00**	**0.00**	2.10%	**0.00**	**0.01**	5.80%	0.26	0.15	0.70%	**0.02**	**0.01**	1.00%	0.05	**0.03**	1.40%	0.23	0.20	5.40%	0.57	0.20
P3M2	0.40%	0.38	**0.01**	0.50%	1.00	0.05	1.40%	0.23	0.20	4.90%	0.88	0.56	0.10%	0.28	0.56	0.30%	0.33	0.05	1.00%	0.88	0.23	4.80%	0.77	0.86
P3M3	0.40%	0.38	**0.01**	0.60%	0.66	0.07	1.40%	0.23	0.20	5.40%	0.57	0.71	0.10%	0.28	0.56	0.40%	0.64	0.05	0.90%	0.75	0.17	4.40%	0.37	0.51
P3M4	0.50%	0.16	**0.01**	1.40%	0.00	**0.00**	2.00%	**0.01**	**0.01**	5.90%	0.20	0.12	0.60%	0.06	**0.01**	1.00%	0.05	0.06	1.40%	0.23	0.20	5.40%	0.57	0.20
P3M5	0.30%	0.76	**0.01**	0.40%	0.64	**0.03**	1.10%	0.75	0.26	5.20%	0.77	0.86	0.10%	0.28	0.56	0.30%	0.33	0.05	1.00%	1.00	0.23	4.80%	0.77	0.86
P3M6	0.40%	0.38	**0.01**	0.50%	1.00	0.05	1.20%	0.54	0.26	5.70%	0.32	0.45	0.10%	0.28	0.56	0.20%	0.13	0.06	0.70%	0.31	0.57	4.40%	0.37	0.51
P3M7	0.50%	0.16	**0.01**	0.90%	0.11	0.05	1.90%	0.01	**0.03**	6.00%	0.16	0.23	1.40%	**0.00**	**0.00**	2.10%	**0.00**	0.06	2.90%	**0.00**	**0.00**	7.20%	**0.00**	**0.01**
P3M8	0.10%	0.28	0.56	0.40%	0.64	**0.03**	0.60%	0.17	**0.03**	4.60%	0.56	0.57	0.60%	0.06	0.17	1.00%	0.05	0.05	2.00%	**0.01**	**0.01**	6.10%	0.12	0.24
P3M9	0.20%	0.74	**0.01**	0.40%	0.64	**0.03**	0.80%	0.51	0.12	5.30%	0.67	0.79	0.40%	0.38	0.67	0.90%	0.11	0.05	1.50%	0.14	0.16	5.70%	0.32	0.38
P4M1	1.00%	**0.00**	**0.00**	1.90%	**0.00**	**0.00**	2.50%	**0.00**	**0.00**	6.10%	0.12	0.18	1.00%	**0.00**	**0.00**	1.20%	**0.01**	**0.03**	1.70%	**0.04**	0.10	6.20%	0.09	0.07
P4M2	0.40%	0.38	0.67	0.70%	0.40	0.67	2.10%	0.00	0.01	6.60%	**0.03**	0.08	0.80%	**0.01**	**0.02**	1.10%	**0.02**	0.07	1.20%	**0.03**	0.71	6.30%	0.07	0.11
P4M3	0.60%	0.06	0.17	0.90%	0.11	0.25	2.10%	0.00	0.01	6.90%	**0.01**	0.03	0.60%	0.06	0.17	1.00%	0.05	0.07	1.20%	0.54	0.71	6.20%	0.09	0.13
P4M4	1.10%	**0.00**	**0.00**	2.00%	**0.00**	**0.00**	2.40%	**0.00**	**0.00**	6.20%	0.09	0.13	1.00%	**0.00**	**0.00**	1.10%	0.02	0.06	1.80%	**0.02**	0.05	6.50%	**0.04**	**0.02**
P4M5	0.20%	0.74	0.94	0.70%	0.40	0.67	2.00%	**0.01**	**0.01**	6.80%	**0.01**	0.04	0.70%	**0.02**	0.06	1.10%	0.02	0.07	1.20%	0.54	0.71	6.80%	**0.01**	**0.02**
P4M6	0.40%	0.38	0.67	0.80%	0.22	0.44	2.10%	**0.00**	**0.01**	7.00%	**0.01**	0.02	0.50%	0.16	0.37	0.90%	0.11	0.07	1.10%	0.75	0.84	6.20%	0.09	0.13
P4M7	1.20%	**0.00**	**0.00**	2.00%	**0.00**	**0.00**	2.90%	**0.00**	**0.00**	7.60%	**0.00**	**0.00**	1.30%	**0.00**	**0.00**	2.40%	**0.00**	0.08	3.10%	**0.00**	**0.00**	8.10%	**0.00**	**0.00**

(*continued on next page*)

Table 3 (*continued*)

	N/T 0.25%	p_{UC}	p_{CC}	N/T 0.50%	p_{UC}	p_{CC}	N/T 1%	p_{UC}	p_{CC}	N/T 5%	p_{UC}	p_{CC}	N/T 95%	p_{UC}	p_{CC}	N/T 99%	p_{UC}	p_{CC}	N/T 99.5%	p_{UC}	p_{CC}	N/T 99.75%	p_{UC}	p_{CC}
P4M8	0.20%	0.74	0.94	0.50%	1.00	0.98	1.50%	0.14	0.27	6.40%	0.05	0.13	0.90%	**0.00**	**0.01**	1.10%	**0.02**	0.06	2.00%	**0.01**	**0.01**	7.90%	**0.00**	**0.00**
P4M9	0.20%	0.74	0.94	0.60%	0.66	0.88	1.90%	0.01	0.03	6.80%	**0.01**	**0.04**	0.90%	**0.00**	**0.01**	1.10%	**0.02**	0.07	1.70%	**0.04**	0.10	7.40%	**0.00**	**0.00**
P5M1	1.40%	**0.00**	**0.00**	2.00%	**0.00**	**0.00**	2.50%	**0.00**	**0.00**	7.80%	**0.00**	**0.00**	1.00%	**0.00**	**0.00**	1.70%	**0.00**	**0.00**	2.50%	**0.00**	**0.00**	6.70%	**0.02**	0.05
P5M2	0.70%	**0.02**	0.06	1.30%	**0.00**	**0.01**	1.70%	**0.04**	0.07	6.40%	0.05	0.07	0.30%	0.76	0.95	0.60%	0.66	0.06	1.00%	0.05	0.90	6.30%	0.07	0.17
P5M3	0.70%	**0.02**	0.06	1.30%	**0.00**	**0.01**	1.70%	**0.04**	0.07	6.70%	**0.02**	**0.02**	0.30%	0.76	0.95	0.40%	0.64	0.07	1.00%	1.00	0.90	5.90%	0.20	0.43
P5M4	1.40%	**0.00**	**0.00**	2.00%	**0.00**	**0.00**	2.60%	**0.00**	**0.00**	7.90%	**0.00**	**0.00**	0.90%	**0.00**	**0.01**	1.50%	**0.00**	0.08	2.50%	**0.00**	**0.00**	6.80%	**0.01**	**0.04**
P5M5	0.50%	0.16	0.37	1.00%	0.05	0.13	1.50%	0.14	0.27	6.40%	0.05	0.07	0.30%	0.76	0.95	0.50%	1.00	0.06	1.00%	1.00	0.90	6.30%	0.07	0.17
P5M6	0.60%	0.06	0.17	1.10%	**0.02**	0.06	1.60%	0.08	0.17	6.70%	**0.02**	**0.02**	0.30%	0.76	0.95	0.40%	0.64	0.07	0.90%	0.75	0.87	6.00%	0.16	0.36
P5M7	1.40%	**0.00**	**0.00**	1.70%	**0.00**	**0.00**	2.70%	**0.00**	**0.00**	8.50%	**0.00**	**0.00**	2.00%	**0.00**	**0.00**	3.20%	**0.00**	0.09	4.30%	**0.00**	**0.00**	9.50%	**0.00**	**0.00**
P5M8	0.30%	0.76	0.95	0.60%	0.66	0.88	1.30%	0.36	0.56	5.90%	0.20	0.29	0.60%	0.06	0.17	1.00%	0.05	0.06	2.00%	**0.01**	**0.01**	7.20%	**0.00**	**0.01**
P5M9	0.30%	0.76	0.95	0.90%	0.11	0.25	1.40%	0.23	0.40	6.40%	0.05	0.07	0.50%	0.16	0.37	1.00%	0.05	0.06	1.70%	**0.04**	0.10	7.00%	**0.01**	**0.02**

Table 4
VaR exceedances N/T, Kupiec's and Christoffersen's tests: 6th–10th portfolios

	N/T 0.25%	p_{UC}	p_{CC}	N/T 0.50%	p_{UC}	p_{CC}	N/T 1%	p_{UC}	p_{CC}	N/T 5%	p_{UC}	p_{CC}	N/T 95%	p_{UC}	p_{CC}	N/T 99%	p_{UC}	p_{CC}	N/T 99.5%	p_{UC}	p_{CC}	N/T 99.75%	p_{UC}	p_{CC}
P6M1	1.80%	**0.00**	**0.00**	2.20%	**0.00**	**0.00**	3.30%	**0.00**	**0.00**	7.90%	**0.00**	**0.00**	0.90%	**0.00**	**0.01**	1.50%	**0.00**	**0.00**	2.20%	**0.00**	**0.00**	7.30%	**0.00**	**0.00**
P6M2	0.30%	0.76	0.95	1.20%	**0.01**	**0.03**	2.40%	**0.00**	**0.00**	6.70%	**0.02**	0.05	0.50%	0.16	0.37	0.80%	0.22	0.07	1.10%	**0.02**	0.84	5.30%	0.67	0.41
P6M3	0.40%	0.38	0.67	1.20%	**0.01**	**0.03**	2.50%	**0.00**	**0.00**	7.00%	**0.01**	**0.02**	0.40%	0.38	0.67	0.60%	0.66	0.07	1.10%	0.75	0.84	5.20%	0.77	0.17
P6M4	1.70%	**0.00**	**0.00**	2.20%	**0.00**	**0.00**	3.10%	**0.00**	**0.00**	8.30%	**0.00**	**0.00**	0.90%	**0.00**	**0.01**	1.50%	**0.00**	0.08	2.10%	**0.00**	**0.01**	7.50%	**0.00**	**0.00**
P6M5	0.30%	0.76	0.95	1.00%	0.05	0.13	2.20%	**0.00**	**0.00**	6.80%	**0.01**	**0.04**	0.30%	0.76	0.95	0.60%	0.66	0.07	1.10%	0.75	0.84	5.50%	0.47	0.21
P6M6	0.30%	0.76	0.95	1.10%	**0.02**	0.06	2.50%	**0.00**	**0.00**	7.30%	**0.00**	**0.01**	0.30%	0.76	0.95	0.60%	0.66	0.07	1.00%	1.00	0.90	5.10%	0.88	0.15
P6M7	2.00%	**0.00**	**0.00**	2.60%	**0.00**	**0.00**	3.90%	**0.00**	**0.00**	9.10%	**0.00**	**0.00**	1.90%	**0.00**	**0.00**	2.90%	**0.00**	0.09	3.80%	**0.00**	**0.00**	10.30%	**0.00**	**0.00**
P6M8	0.30%	0.76	0.95	0.50%	1.00	0.98	1.80%	**0.02**	0.05	6.80%	**0.01**	0.05	0.80%	**0.01**	**0.02**	1.20%	**0.01**	0.07	2.00%	**0.01**	**0.01**	7.50%	**0.00**	**0.00**
P6M9	0.30%	0.76	0.95	0.60%	0.66	0.88	1.80%	**0.02**	0.05	6.80%	**0.01**	**0.04**	0.80%	**0.01**	**0.02**	1.00%	0.05	0.07	1.90%	**0.01**	**0.03**	7.20%	**0.00**	**0.01**
P7M1	1.90%	**0.00**	**0.00**	2.50%	**0.00**	**0.00**	3.30%	**0.00**	**0.00**	7.10%	**0.00**	**0.01**	1.00%	**0.00**	**0.00**	1.70%	**0.00**	**0.00**	2.50%	**0.00**	**0.00**	6.20%	0.09	0.24
P7M2	0.50%	0.16	0.37	0.80%	0.22	0.44	2.20%	**0.00**	**0.00**	7.10%	**0.00**	**0.01**	0.20%	0.74	0.94	0.60%	0.66	0.07	1.00%	**0.00**	0.90	6.30%	0.07	0.19
P7M3	0.50%	0.16	0.37	1.00%	0.05	0.13	2.10%	**0.00**	**0.01**	7.10%	**0.00**	**0.01**	0.20%	0.74	0.94	0.20%	0.13	0.07	1.00%	1.00	0.90	6.10%	0.12	0.30
P7M4	1.80%	**0.00**	**0.00**	2.40%	**0.00**	**0.00**	3.20%	**0.00**	**0.00**	7.50%	**0.00**	**0.00**	1.00%	**0.00**	**0.00**	1.60%	**0.00**	0.08	2.70%	**0.00**	**0.00**	6.20%	0.09	0.24
P7M5	0.40%	0.38	0.67	0.60%	0.66	0.88	2.10%	**0.00**	**0.01**	7.20%	**0.00**	**0.01**	0.20%	0.74	0.94	0.40%	0.64	0.07	1.00%	1.00	0.90	6.30%	0.07	0.19
P7M6	0.40%	0.38	0.67	0.60%	0.66	0.88	2.00%	**0.01**	**0.01**	7.20%	**0.00**	**0.01**	0.10%	0.28	0.56	0.20%	0.13	0.07	0.90%	0.75	0.87	6.20%	0.09	0.24
P7M7	1.50%	**0.00**	**0.00**	2.20%	**0.00**	**0.00**	3.20%	**0.00**	**0.00**	7.60%	**0.00**	**0.00**	2.40%	**0.00**	**0.00**	2.90%	**0.00**	0.08	3.70%	**0.00**	**0.00**	8.00%	**0.00**	**0.00**
P7M8	0.20%	0.74	0.94	0.40%	0.64	0.88	1.40%	0.23	0.40	6.50%	**0.04**	0.11	0.60%	0.06	0.17	0.90%	0.11	0.07	2.10%	**0.00**	**0.00**	6.90%	**0.01**	**0.03**
P7M9	0.20%	0.74	0.94	0.60%	0.66	0.88	1.50%	0.14	0.27	6.60%	**0.03**	0.08	0.30%	0.76	0.95	0.70%	0.40	0.07	2.10%	**0.00**	**0.00**	6.90%	**0.01**	**0.03**
P8M1	0.30%	0.76	**0.01**	0.60%	0.66	0.07	1.00%	1.00	0.23	3.70%	0.05	0.02	0.40%	0.38	0.67	1.00%	0.05	0.13	1.60%	0.08	0.11	4.80%	0.77	0.54
P8M2	0.20%	0.74	0.94	0.20%	0.13	0.31	0.90%	0.75	0.17	4.10%	0.18	0.11	0.30%	0.76	0.95	0.50%	1.00	**0.04**	1.20%	0.18	0.26	4.50%	0.46	0.33
P8M3	0.20%	0.74	0.94	0.30%	0.33	0.01	0.90%	0.75	0.17	4.30%	0.30	0.07	0.30%	0.76	0.95	0.30%	0.33	**0.04**	1.10%	0.75	0.26	4.50%	0.46	0.33
P8M4	0.20%	0.74	0.94	0.60%	0.66	0.07	1.00%	1.00	0.23	3.80%	0.07	0.03	0.40%	0.38	0.67	1.00%	0.05	**0.04**	1.50%	0.14	0.16	4.80%	0.77	0.54
P8M5	0.00%	**0.03**	0.08	0.20%	0.13	0.31	0.70%	0.31	0.07	4.10%	0.18	**0.04**	0.30%	0.76	0.95	0.30%	0.33	**0.04**	1.20%	0.54	0.26	4.50%	0.46	0.33
P8M6	0.10%	0.28	0.56	0.30%	0.33	**0.01**	0.90%	0.75	0.17	4.50%	0.46	0.01	0.30%	0.76	0.95	0.30%	0.33	0.05	1.10%	0.75	0.26	4.50%	0.46	0.33
P8M7	0.10%	0.28	0.56	0.30%	0.33	0.62	0.90%	0.75	0.17	4.00%	0.13	0.08	1.50%	**0.00**	**0.00**	1.90%	**0.00**	**0.04**	2.70%	**0.00**	**0.00**	6.20%	0.09	0.07
P8M8	0.00%	**0.03**	0.08	0.10%	**0.03**	0.09	0.30%	**0.01**	**0.00**	3.60%	**0.03**	**0.01**	0.80%	**0.01**	**0.02**	1.40%	**0.00**	**0.04**	2.00%	**0.01**	**0.00**	5.90%	0.20	0.09
P8M9	0.00%	**0.03**	0.08	0.10%	**0.03**	0.09	0.60%	0.17	0.03	3.90%	0.10	0.05	0.60%	0.06	0.17	1.30%	**0.00**	**0.04**	1.80%	**0.02**	**0.01**	5.50%	0.47	0.08
P9M1	0.90%	**0.00**	**0.01**	1.50%	**0.00**	**0.00**	2.40%	**0.00**	**0.00**	6.40%	0.05	0.15	0.90%	**0.00**	**0.01**	1.00%	0.05	0.13	1.10%	0.75	0.84	4.40%	0.37	0.51
P9M2	0.50%	0.16	0.37	1.00%	0.05	0.13	1.60%	0.08	0.17	6.40%	0.05	0.12	0.10%	0.28	0.56	0.40%	0.64	0.06	0.90%	0.05	0.87	4.10%	0.18	0.39
P9M3	0.60%	0.06	0.17	0.90%	0.11	0.25	1.60%	0.08	0.17	6.60%	**0.03**	0.08	0.10%	0.28	0.56	0.20%	0.13	0.07	0.90%	0.75	0.87	3.70%	0.05	0.12
P9M4	0.90%	**0.00**	**0.01**	1.40%	**0.00**	**0.00**	2.40%	**0.00**	**0.00**	6.50%	**0.04**	0.11	0.90%	**0.00**	**0.01**	1.00%	0.05	0.07	1.20%	0.54	0.26	4.70%	0.66	0.79
P9M5	0.50%	0.16	0.37	0.80%	0.22	0.44	1.50%	0.14	0.27	6.60%	**0.03**	0.08	0.10%	0.28	0.56	0.20%	0.13	0.07	0.90%	0.75	0.87	4.40%	0.37	0.67
P9M6	0.40%	0.38	0.67	0.80%	0.22	0.44	1.60%	0.08	0.17	6.70%	**0.02**	0.06	0.10%	0.28	0.56	0.10%	**0.03**	0.07	0.90%	0.75	0.87	3.70%	0.05	0.12
P9M7	1.00%	**0.00**	**0.00**	2.20%	**0.00**	**0.00**	2.60%	**0.00**	**0.00**	7.30%	**0.00**	**0.01**	1.10%	**0.00**	**0.00**	1.40%	**0.00**	0.07	2.00%	**0.01**	**0.01**	6.70%	**0.02**	0.05

(*continued on next page*)

Table 4 (*continued*)

	N/T 0.25%	p_{UC}	p_{CC}	N/T 0.50%	p_{UC}	p_{CC}	N/T 1%	p_{UC}	p_{CC}	N/T 5%	p_{UC}	p_{CC}	N/T 95%	p_{UC}	p_{CC}	N/T 99%	p_{UC}	p_{CC}	N/T 99.5%	p_{UC}	p_{CC}	N/T 99.75%	p_{UC}	p_{CC}
P9M8	0.20%	0.74	0.94	0.50%	1.00	0.98	1.10%	0.75	0.84	6.30%	0.07	0.17	0.30%	0.76	0.95	0.90%	0.11	0.06	1.30%	0.36	0.56	5.60%	0.39	0.69
P9M9	0.30%	0.76	0.95	0.60%	0.66	0.88	1.30%	0.36	0.56	6.50%	**0.04**	0.11	0.10%	0.28	0.56	0.70%	0.40	0.07	1.20%	0.54	0.71	5.30%	0.67	0.90
P10M1	1.00%	**0.00**	**0.00**	1.60%	**0.00**	**0.00**	2.20%	**0.00**	**0.00**	6.70%	**0.02**	0.06	0.50%	0.16	0.37	0.70%	0.40	0.67	1.30%	0.36	0.56	5.10%	0.88	0.96
P10M2	0.50%	0.16	0.37	0.60%	0.66	0.88	1.20%	0.54	0.26	5.70%	0.32	0.60	0.30%	0.76	0.95	0.50%	1.00	0.06	0.70%	0.32	0.57	4.70%	0.66	0.07
P10M3	0.50%	0.16	0.37	0.60%	0.66	0.88	1.20%	0.54	0.26	5.80%	0.26	0.49	0.30%	0.76	0.95	0.50%	1.00	0.06	0.80%	0.51	0.75	4.70%	0.66	0.07
P10M4	0.70%	**0.02**	0.06	1.20%	**0.01**	**0.01**	2.10%	**0.00**	**0.01**	6.90%	**0.01**	**0.03**	0.50%	0.16	0.37	0.50%	1.00	0.07	1.20%	0.54	0.71	4.90%	0.88	0.60
P10M5	0.40%	0.38	0.67	0.60%	0.66	0.88	0.80%	0.51	0.75	5.90%	0.20	0.43	0.30%	0.76	0.95	0.50%	1.00	0.06	0.60%	0.17	0.38	4.70%	0.66	0.07
P10M6	0.50%	0.16	0.37	0.60%	0.66	0.88	1.20%	0.54	0.26	5.80%	0.26	0.49	0.30%	0.76	0.95	0.30%	0.33	0.06	0.70%	0.31	0.57	4.70%	0.66	0.07
P10M7	0.90%	**0.00**	**0.01**	1.20%	**0.01**	**0.03**	2.50%	**0.00**	**0.00**	8.10%	**0.00**	**0.00**	1.10%	**0.00**	**0.00**	1.60%	**0.00**	0.08	2.10%	**0.00**	**0.01**	7.60%	0.00	**0.00**
P10M8	0.00%	**0.03**	0.08	0.50%	1.00	0.98	0.80%	0.51	0.75	5.50%	0.47	0.77	0.50%	0.16	0.37	0.60%	0.66	0.06	1.20%	0.54	0.71	5.90%	0.20	0.01
P10M9	0.00%	**0.03**	0.08	0.50%	1.00	0.98	0.80%	0.51	0.75	5.40%	0.57	0.85	0.50%	0.16	0.37	0.70%	0.40	0.05	1.10%	0.75	0.84	5.60%	0.39	0.09

Table 5
Asymmetric loss functions (14) and SPA tests: 1st–5th portfolios

LOSS	**0.25%**	**0.50%**	**1%**	**5%**	**0.25%**	**0.50%**	**1%**	**5%**	**SPA**	**0.25%**	**0.50%**	**1%**	**5%**	**0.25%**	**0.50%**	**1%**	**5%**
P1M1	0.7223	1.1616	1.9615	6.2138	0.6374	1.0496	1.7239	5.6260	M.1	**0.016**	**0.006**	**0.003**	**0.028**	**0.04**	**0.024**	**0.024**	0.481
P1M2	0.4927	0.8991	1.6568	6.0100	**0.4890**	0.8958	1.5798	5.6061	M.2	0.915	0.586	0.257	0.141	0.862	0.777	0.822	0.34
P1M3	0.5115	0.9447	1.7218	6.0836	0.5096	0.8951	1.5702	5.5661	M.3	0.378	0.093	**0.047**	**0.021**	**0.000**	0.804	0.824	0.505
P1M4	0.7024	1.1511	1.9537	6.2371	0.6192	1.0326	1.7109	5.6358	M.4	**0.011**	**0.003**	**0.002**	**0.013**	0.082	**0.042**	0.079	0.217
P1M5	0.4926	0.8937	1.6417	6.0134	0.5026	0.8936	1.5747	5.6096	M.5	0.976	0.544	0.302	0.175	**0.000**	0.802	0.820	0.381
P1M6	0.4925	0.9153	1.6908	6.0869	0.5287	0.8922	**1.5665**	**5.5618**	M.6	0.915	0.353	0.071	0.060	**0.000**	0.686	0.815	0.995
P1M7	0.6472	1.0917	1.8885	6.2539	0.7889	1.2034	1.8885	5.7919	M.7	0.053	0.050	**0.038**	**0.003**	**0.007**	**0.008**	**0.006**	**0.002**
P1M8	0.5075	0.8940	**1.5879**	**5.9767**	0.5228	0.9046	1.6294	5.6484	M.8	0.118	0.267	0.906	0.959	0.390	0.345	0.243	0.149
P1M9	**0.4916**	**0.8765**	1.6283	6.0443	0.4969	**0.8839**	1.5878	5.5898	M.9	0.981	1.000	0.366	0.095	0.699	0.982	0.813	0.605
P2M1	0.4303	0.7683	1.3479	5.1677	0.8107	1.2097	1.8706	5.6689	M.1	0.307	0.328	0.324	0.428	0.065	0.075	0.222	0.722
P2M2	0.4046	0.7089	1.3037	5.1615	0.6804	1.0966	1.7824	5.6555	M.2	**0.000**	0.275	0.941	0.294	0.319	0.656	0.51	0.892
P2M3	**0.3984**	**0.7029**	1.3041	5.1780	**0.6604**	**1.0867**	1.7669	**5.6512**	M.3	0.994	0.998	0.913	0.218	0.974	1.000	0.800	0.870
P2M4	0.4212	0.7640	1.3453	5.1842	0.7894	1.2050	1.8656	5.6660	M.4	0.744	0.395	0.433	0.032	0.087	0.078	0.281	0.884
P2M5	0.4073	0.7192	**1.3019**	5.1760	0.6696	1.0982	1.7740	5.6671	M.5	0.156	**0.001**	0.992	0.163	0.265	0.357	0.691	0.416
P2M6	0.4034	0.7087	1.3071	5.1880	0.6646	1.0938	**1.7633**	5.6678	M.6	0.695	0.478	0.730	0.169	0.603	0.331	0.837	0.163
P2M7	0.4342	0.7787	1.3749	5.2662	0.9629	1.3607	2.0172	5.7788	M.7	0.082	0.098	**0.033**	**0.001**	**0.010**	**0.009**	**0.011**	0.053
P2M8	0.4393	0.7721	1.3254	**5.1376**	0.7175	1.1143	1.8262	5.7166	M.8	**0.000**	**0.000**	0.280	0.920	0.349	0.646	0.397	0.226
P2M9	0.4287	0.7562	1.3205	5.1615	0.7109	1.1174	1.8258	5.7070	M.9	**0.000**	**0.000**	0.385	0.319	0.38	0.649	0.389	0.268
P3M1	0.5151	0.8730	1.5034	5.1700	0.4402	0.8067	1.4386	5.2595	M.1	0.214	0.074	0.071	0.503	0.223	0.244	0.315	0.198
P3M2	0.4382	0.7910	1.4019	**5.1329**	0.3988	**0.7220**	1.3738	**5.1925**	M.2	0.469	0.888	0.689	0.981	0.679	0.897	0.976	0.940
P3M3	0.4485	0.7998	1.4146	5.1353	0.4185	0.7590	1.3902	5.2041	M.3	0.466	0.676	0.530	0.938	**0.000**	**0.008**	0.482	0.180
P3M4	0.5084	0.8597	1.5085	5.1818	0.4344	0.7994	1.4334	5.2557	M.4	0.262	0.116	0.072	0.368	0.305	0.323	0.402	0.319
P3M5	0.4284	**0.7855**	**1.3934**	5.1338	0.4155	0.7407	**1.3734**	5.1928	M.5	0.839	0.992	0.998	0.935	**0.000**	**0.000**	0.887	0.972
P3M6	0.4298	0.7942	1.4065	5.1457	0.4339	0.7661	1.3949	5.1941	M.6	0.980	0.723	0.678	0.206	**0.000**	**0.000**	0.290	0.884
P3M7	0.4830	0.8207	1.4681	5.1917	0.5799	0.9449	1.6143	5.3441	M.7	0.318	0.507	0.170	0.199	**0.025**	**0.024**	**0.009**	0.025
P3M8	0.4330	0.7930	1.4093	5.1369	**0.3943**	0.7800	1.4277	5.2290	M.8	0.115	0.589	0.449	0.636	0.896	0.330	0.171	0.292
P3M9	**0.4231**	0.7904	1.4004	5.1373	0.3996	0.7663	1.4010	5.2144	M.9	0.999	0.861	0.788	0.863	0.731	0.466	0.687	0.677
P4M1	0.5039	0.8687	1.5060	**4.9842**	0.8837	1.2236	1.7860	**5.1306**	M.1	0.134	0.069	0.055	0.895	0.079	0.247	0.512	0.909
P4M2	0.4243	0.7747	1.4347	5.0192	0.7598	1.1843	1.7985	5.2469	M.2	0.553	0.53	0.071	0.165	0.129	0.112	0.436	0.085
P4M3	0.4310	0.7855	1.4661	5.0431	0.7164	1.1399	1.7908	5.2257	M.3	0.386	0.511	0.044	0.074	0.651	0.513	0.881	0.135
P4M4	0.4934	0.8680	1.5116	4.9975	0.8845	1.2261	**1.7829**	5.1360	M.4	0.176	0.063	0.051	0.179	0.064	0.283	0.946	0.490
P4M5	0.4144	0.7774	1.4314	5.0309	0.7375	1.1550	1.7881	5.2555	M.5	0.411	0.558	0.018	0.147	0.049	0.228	0.968	0.075
P4M6	0.4086	0.7765	1.4501	5.0597	**0.7079**	**1.1281**	1.7908	5.2343	M.6	0.993	0.670	0.020	0.039	0.875	0.915	0.692	0.141
P4M7	0.5240	0.9196	1.5752	5.0721	1.0137	1.3647	1.9650	5.3033	M.7	0.072	**0.003**	**0.002**	0.018	**0.011**	**0.008**	**0.004**	**0.010**
P4M8	**0.4063**	**0.7482**	**1.3705**	5.0025	0.8526	1.2538	1.8577	5.3547	M.8	0.761	0.877	0.838	0.643	0.077	0.144	0.191	**0.017**
P4M9	0.4100	0.7597	1.3790	5.0298	0.8306	1.2390	1.8361	5.3123	M.9	0.821	0.560	0.252	0.216	0.081	0.123	0.265	**0.033**
P5M1	1.0656	1.5936	2.4240	7.1341	0.8157	1.2786	2.0923	6.8045	M.1	**0.013**	**0.022**	**0.047**	**0.014**	0.097	0.114	0.229	0.442
P5M2	0.6988	1.2606	2.1437	6.9422	0.7208	**1.1810**	1.9905	6.7665	M.2	0.110	0.099	0.449	0.158	0.984	0.993	0.776	0.115
P5M3	0.6941	1.2962	2.1814	6.9700	**0.7200**	1.1887	1.9913	6.7354	M.3	0.145	0.058	0.081	0.051	0.915	0.521	0.565	0.614
P5M4	1.0401	1.5567	2.4034	7.1668	0.8027	1.2648	2.0790	6.8139	M.4	**0.012**	**0.024**	**0.032**	**0.003**	0.266	0.287	0.256	0.255
P5M5	0.6608	1.2219	2.1320	6.9592	0.7251	1.1834	**1.9846**	6.7703	M.5	0.411	0.176	0.468	0.095	0.788	0.901	0.998	0.163
P5M6	0.6475	1.2377	2.1479	6.9802	0.7280	1.1921	1.9878	**6.7319**	M.6	0.726	0.154	0.397	0.059	0.534	0.505	0.648	0.979
P5M7	0.9774	1.5086	2.3794	7.2616	1.0734	1.6146	2.4905	7.1181	M.7	**0.021**	**0.032**	0.056	**0.000**	**0.002**	**0.000**	**0.000**	**0.001**
P5M8	0.6358	**1.1317**	**2.0866**	**6.9043**	0.7584	1.2620	2.0629	6.9233	M.8	0.575	0.951	0.900	0.943	0.272	0.114	0.128	**0.007**
P5M9	**0.6316**	1.1718	2.1113	6.9363	0.7310	1.2220	2.0268	6.8597	M.9	0.954	0.200	0.410	0.203	0.796	0.439	0.513	0.055

the Data Generating Process (DGP). We found that, when the true copula is represented by the normal copula, there is marginal skewness in the data and symmetric marginals are used, the estimated correlations are negatively biased, and the bias increases when moving from the Student's t to the normal distribution, reaching values as high as 27% of the true correlations. Besides, we found that the bias almost doubles if negative correlations are considered, as compared to the case for positive correlations. When the true dependence function is represented by the t copula, the choice of

Table 6
Asymmetric loss functions (14) and SPA tests: 6th–10th portfolios

LOSS	**0.25%**	**0.50%**	**1%**	**5%**	**0.25%**	**0.50%**	**1%**	**5%**	**SPA**	**0.25%**	**0.50%**	**1%**	**5%**	**0.25%**	**0.50%**	**1%**	**5%**
P6M1	1.1366	1.7290	2.6792	8.0874	1.0944	1.6373	2.5408	7.6675	M.1	**0.010**	**0.003**	**0.003**	**0.01**	0.164	0.16	0.178	0.286
P6M2	0.6775	1.2231	2.2715	7.8308	0.9540	1.5040	2.4201	**7.5295**	M.2	0.616	0.057	**0.025**	0.404	0.573	0.195	0.497	0.94
P6M3	0.6918	1.2475	2.3144	7.8571	0.9519	1.4929	2.4251	7.5323	M.3	0.252	0.059	**0.012**	0.130	0.595	0.315	0.324	0.777
P6M4	1.1166	1.7205	2.6949	8.1196	1.0863	1.6273	2.5322	7.6778	M.4	**0.013**	**0.003**	**0.007**	**0.005**	0.184	0.211	0.342	0.230
P6M5	0.6669	1.1830	2.2371	7.8436	**0.9359**	**1.4698**	**2.4073**	7.5313	M.5	0.793	0.557	**0.042**	0.189	0.923	0.997	0.997	0.926
P6M6	0.6734	1.2191	2.2842	7.8766	0.9382	1.4873	2.4212	7.5328	M.6	0.713	0.092	0.025	0.045	0.734	0.170	0.404	0.894
P6M7	1.2359	1.8773	2.8897	8.3879	1.4000	1.9713	2.9047	8.1426	M.7	**0.005**	**0.000**	**0.000**	**0.000**	**0.005**	**0.005**	**0.002**	**0.000**
P6M8	**0.6526**	**1.1599**	**2.1399**	**7.8199**	1.0201	1.5926	2.4901	7.6361	M.8	0.776	0.834	0.949	0.957	0.277	0.195	0.350	0.143
P6M9	0.6662	1.1840	2.1900	7.8532	1.0001	1.5634	2.4740	7.6345	M.9	0.498	0.313	0.056	0.152	0.366	0.323	0.476	0.141
P7M1	0.8557	1.3811	2.2256	6.6212	0.6339	1.0828	1.8691	6.0986	M.1	**0.008**	**0.003**	**0.002**	**0.007**	0.066	**0.031**	**0.009**	0.125
P7M2	0.5912	1.0096	1.8297	6.4211	0.5269	0.9246	1.6256	5.9973	M.2	0.299	0.396	**0.034**	0.259	0.007	0.430	0.722	0.267
P7M3	0.5978	1.0262	1.8414	6.4295	0.5168	**0.8963**	**1.6143**	**5.9535**	M.3	0.198	0.115	**0.034**	0.403	0.590	0.991	0.999	0.789
P7M4	0.8244	1.3625	2.2282	6.6499	0.6189	1.0672	1.8682	6.1060	M.4	**0.008**	**0.006**	**0.001**	**0.007**	0.069	**0.023**	**0.005**	0.097
P7M5	0.5620	0.9920	1.7857	6.4394	0.5301	0.9218	1.6356	5.9959	M.5	0.823	0.819	0.214	0.087	0.130	0.191	0.242	0.397
P7M6	0.5789	1.0039	1.8012	6.4338	0.5273	0.9079	1.6170	5.9545	M.6	0.398	0.443	0.171	0.304	0.435	0.317	0.562	0.912
P7M7	0.7629	1.3053	2.1846	6.7023	0.8778	1.3782	2.1576	6.2408	M.7	**0.043**	**0.007**	**0.001**	**0.003**	**0.003**	**0.001**	**0.001**	**0.027**
P7M8	**0.5470**	**0.9795**	**1.7326**	**6.3982**	0.5302	0.9493	1.6856	6.0714	M.8	0.786	0.768	0.955	0.923	0.231	0.164	0.086	0.115
P7M9	0.5500	0.9851	1.7569	6.4134	**0.4948**	0.9200	1.6441	6.0494	M.9	0.822	0.695	0.255	0.47	1.000	0.688	0.587	0.104
P8M1	**0.3185**	0.6064	1.1324	4.2107	0.4576	0.7816	1.3793	**4.6894**	M.1	0.702	0.297	0.369	0.324	0.399	0.144	0.488	0.899
P8M2	0.3373	0.6221	1.1396	4.2154	0.4548	0.7790	1.3635	4.7131	M.2	**0.000**	**0.047**	**0.016**	0.449	0.865	0.854	0.381	0.285
P8M3	0.3346	0.6184	1.1410	4.2145	0.4549	0.7822	1.3675	4.7205	M.3	0.052	**0.026**	0.279	0.510	0.721	0.442	0.108	0.156
P8M4	0.3187	0.5980	1.1266	4.2055	0.4515	**0.7727**	1.3737	4.6901	M.4	0.783	0.880	0.716	0.656	0.939	0.988	0.615	0.884
P8M5	0.3422	0.6275	1.1339	4.2094	0.4516	0.7763	**1.3495**	4.7058	M.5	**0.000**	**0.002**	0.178	0.832	0.956	0.939	0.985	0.746
P8M6	0.3358	0.6184	1.1321	4.2101	**0.4511**	0.7851	1.3543	4.7120	M.6	**0.000**	**0.040**	0.388	0.613	0.708	0.182	0.604	0.534
P8M7	0.3208	**0.5948**	**1.1151**	**4.1982**	0.6237	0.9625	1.5292	4.7335	M.7	0.513	0.757	0.904	0.945	**0.010**	**0.014**	**0.041**	0.418
P8M8	0.3863	0.6669	1.1764	4.2274	0.4957	0.8546	1.4572	4.7189	M.8	**0.000**	**0.000**	**0.000**	0.213	0.162	0.059	0.119	0.656
P8M9	0.3712	0.6486	1.1590	4.2186	0.4847	0.8369	1.4438	4.7149	M.9	**0.000**	**0.000**	**0.006**	0.411	0.279	0.089	0.125	0.766
P9M1	0.5375	0.8865	1.5502	**5.2787**	0.7029	1.0908	1.7098	5.2699	M.1	**0.049**	0.110	0.117	0.912	**0.036**	0.110	0.237	0.448
P9M2	0.4355	0.8185	1.4745	5.2926	0.5676	**0.9080**	1.5924	5.2456	M.2	0.489	0.304	0.356	0.859	0.156	0.993	0.648	0.213
P9M3	0.4442	0.8189	1.4805	5.2956	0.5681	0.9179	1.5940	5.2657	M.3	0.279	0.468	0.376	0.858	0.529	0.427	0.370	0.069
P9M4	0.5187	0.8757	1.5413	5.2853	0.6805	1.0837	1.7116	5.2671	M.4	0.135	0.240	0.164	0.396	0.109	0.129	0.226	0.525
P9M5	0.4225	**0.8004**	1.4620	5.3014	0.5855	0.9217	**1.5832**	5.2333	M.5	0.932	0.993	0.958	0.554	**0.000**	**0.018**	0.977	0.771
P9M6	**0.4196**	0.8144	1.4700	5.3041	0.5834	0.9362	1.5962	5.2531	M.6	0.974	0.454	0.707	0.407	**0.000**	**0.000**	0.321	0.221
P9M7	0.5643	0.9526	1.6345	5.3626	0.8230	1.1866	1.7907	5.3030	M.7	**0.043**	**0.008**	**0.007**	**0.044**	**0.028**	0.052	0.077	0.231
P9M8	0.4274	0.8022	1.4628	5.2943	**0.5477**	0.9456	1.6081	5.2284	M.8	0.612	0.708	0.686	0.740	0.880	0.302	0.631	0.521
P9M9	0.4244	0.8060	**1.4611**	5.2937	0.5503	0.9246	1.5987	**5.2178**	M.9	0.716	0.694	0.923	0.886	0.825	0.74	0.887	1.000
P10M1	0.5748	0.9654	1.6482	5.6243	0.9933	**1.3357**	**1.9693**	**5.7744**	M.1	0.162	0.084	0.126	0.197	0.549	0.936	0.983	0.982
P10M2	0.4804	0.8794	1.5468	5.5822	0.9411	1.3438	1.9976	5.8839	M.2	0.433	0.702	0.394	0.080	0.899	0.95	0.757	0.145
P10M3	0.4819	0.8857	1.5435	5.5719	0.9385	1.3511	2.0292	5.8956	M.3	0.420	0.459	0.615	0.353	0.895	0.707	0.063	0.073
P10M4	0.5512	0.9315	1.6247	5.6266	0.9892	1.3428	1.9775	5.7810	M.4	0.246	0.277	0.149	0.172	0.625	0.393	0.493	0.636
P10M5	**0.4542**	0.8706	1.5321	5.6013	**0.9366**	1.3425	2.0122	5.8835	M.5	0.995	0.825	0.941	**0.020**	0.970	0.901	0.088	0.167
P10M6	0.4558	0.8697	**1.5311**	5.5853	0.9456	1.3523	2.0299	5.8878	M.6	0.882	0.893	0.991	0.204	0.024	0.300	0.061	0.118
P10M7	0.5612	0.9628	1.6598	5.7464	1.0626	1.4410	2.0683	5.9398	M.7	0.179	0.171	0.070	**0.012**	0.101	0.067	0.118	0.054
P10M8	0.4596	0.8580	1.5451	5.5690	0.9651	1.3460	2.0015	5.8849	M.8	**0.002**	0.711	0.558	0.238	0.694	0.816	0.686	0.258
P10M9	0.4569	**0.8569**	1.5446	**5.5497**	0.9696	1.3469	2.0082	5.8846	M.9	0.597	0.892	0.737	1.000	0.623	0.875	0.658	0.243

the marginals tends to have much stronger effects on copula parameter estimation, with biases up to 50% of the true values for the correlations and up to 380% for the t copula degrees of freedom parameter. If the dependence structure is represented by a copula which is not elliptical, e.g. the Clayton copula, the effects of marginal misspecifications on the copula parameter estimation can be rather different, depending on the sign of the marginal skewness.

We then implemented an extensive Monte Carlo study to assess the potential impact of both misspecified margins and misspecified copulas on the estimation of multivariate VaR for equally weighted portfolios. We found that, when small samples are considered and the data are leptokurtic and skewed, the overestimation/underestimation in the GARCH parameters is so large as to deliver conservative VaR estimates, even with a simple multivariate normal distribution. In general, the VaR estimates are very poor and suffer from computational problems when estimating GARCH models for small samples, as also discussed in Hwang and Valls Pereira (2006). When the sample dimension increases, the biases in the volatility parameters are much smaller, whereas those in the copula parameters remain almost unchanged or even increase. In this case, copula misspecifications do play a role for VaR computation. However, these effects depend heavily on the sign of the dependence: if it is negative, the bias can be as large as 70%, like for t copula correlations; if it is positive, the bias is much smaller (10% or less for the t copula correlations), and the effects on quantile estimation are much more limited, if not completely offset by marginal misspecifications. Therefore, this Monte Carlo evidence gives some insights into why previous empirical literature found that the influence of a misspecification in the copula is given with 20% or less of the whole estimation error for the VaR; see e.g. Ané and Kharoubi (2003) and Junker and May (2005). Finally, we performed an empirical analysis with ten trivariate portfolios, where we quantified the risk of the portfolio under different joint distribution assumptions.

An avenue for further research is performing a Monte Carlo analysis with high dimensional portfolios, where dynamic dependence and other forms of nonlinearities play an important role.

References

Ané, T., Kharoubi, C., 2003. Dependence structure and risk measure. The Journal of Business 76 (3), 411–438.

Bollerslev, T., Engle, R., Nelson, D., 1994. ARCH models. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, Vol. IV. Elsevier, Amsterdam, The Netherlands.

Chen, X., Fan, Y., Patton, A., 2004. Simple tests for models of dependence between multiple financial time series with applications to US equity returns and exchange rates. Financial Markets Group technical report n. 483, London School of Economics.

Cherubini, U., Luciano, E., Vecchiato, W., 2004. Copula Methods in Finance. Wiley.

Christoffersen, P., 1998. Evaluating interval forecasts. International Economic Review 39, 841–862.

Christoffersen, P., Diebold, F.X., 2000. How relevant is volatility forecasting for financial risk management? Review of Economics and Statistics 82, 1–11.

DeMiguel, V., Garlappi, L., Uppal, R., 2007. Optimal versus naive diversification: How inefficient is the 1/N Portfolio Strategy? Review of Financial Studies. doi:10.1093/rfs/hhm075.

Embrechts, P., 2000. Extreme value theory: Potential and limitations as an integrated risk management tool. Derivatives Use, Trading and Regulation 6 (1), 449–456.

Fantazzini, D., 2007. Using misspecified marginals and misspecified copulas to compute the value at risk: When do we have to care? Moscow School of Economics Working Paper. Available at SSRN http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1070641.

Fantazzini, D., 2008. Dynamic copula modelling for value at risk. Frontiers in Finance and Economics (in press).

Fantazzini, D., Dalla Valle, L., Giudici, P., 2008. Copulae and Operational Risks. International Journal of Risk Assessment and Management (in press).

Giacomini, R., Komunjer, I., 2005. Evaluation and combination of conditional quantile forecasts. Journal of Business and Economic Statistics 23, 416–431.

Giot, P., Laurent, S., 2003. Value at risk for long and short positions. Journal of Applied Econometrics 18, 641–664.

González-Rivera, G., Lee, T., Santosh, M., 2006. Forecasting volatility: A reality check based on option pricing, utility function, value at Risk, and predictive likelihood. International Journal of Forecasting 20, 629–645.

Gunky, K., Silvapulle, M.J., Silvapulle, P., 2007. Comparison of semiparametric and parametric methods for estimating copulas. Computational Statistics and Data Analysis 51 (6), 2836–2850.

Hansen, B., 1994. Autoregressive conditional density estimation. International Economic Review 35 (3), 705–730.

Hansen, P., 2005. A test for superior predictive ability. Journal of Business and Economic Statistics 23 (4), 365–380.

Hwang, S., Valls Pereira, P., 2006. Small sample properties of GARCH estimates and persistence. The European Journal of Finance 12, 473–494.

Karlis, D., Nikoloulopoulos, A.K., 2008. Copula model evaluation based on parametric bootstrap. Computational Statistics and Data Analysis. doi:10.1016/j.csda.2007.10.028.

Kupiec, P., 1995. Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives 2, 173–184.

Joe, H., Xu, J., 1996. The estimation method of inference functions for margins for multivariate models. Department of Statistics, University of British Columbia, Technical Report n. 166.

Joe, H., 1997. Multivariate Models and Dependence Concepts. Chapman Hall.

Jorion, P., 2000. Value at risk: The new benchmark for managing financial risk, 2nd edition. McGraw-Hill.

Junker, M., May, A., 2005. Measurement of aggregate risk with copulas. Econometrics Journal 8, 428–454.

McNeil, A., Frey, R., Embrechts, P., 2005. Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press.

Nelsen, R.B., 1999. An Introduction to Copulas. In: Lecture Notes in Statistics, vol. 139. Springer.

Newey, W., Steigerwald, D., 1997. Asymptotic bias for quasi maximum likelihood estimators in conditional heteroskedasticity models. Econometrica 3, 587–599.

Patton, A., 2004. On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. Journal of Financial Econometrics 2 (1), 130–168.

Patton, A., 2006a. Estimation of copula models for time series of possibly different lengths. Journal of Applied Econometrics 21, 147–173.

Patton, A., 2006b. Modelling asymmetric exchange rate dependence. International Economic Review 47 (2), 527–556.

Sklar, A., 1959. Fonctions de répartition á n dimensions et leurs marges. Publ. Inst. Statis. Univ. Paris, 8, pp. 229–231.

Tsay, R., 2002. Analysis of Financial Time Series. Wiley.

Appendix 3 "Three-stage semi-parametric estimation of t-copulas: Asymptotics, finite-sample properties and computational aspects."

Fantazzini, D. (2010a). Three-stage semi-parametric estimation of t-copulas: Asymptotics, finite-sample properties and computational aspects. *Computational Statistics & Data Analysis*, 54(11), 2562-2579. DOI: https://doi.org/10.1016/j.csda.2009.02.004

Permisson to post the Publisher's version/PDF: according to https://www.elsevier.com/about/policies/sharing "*Theses and dissertations which contain embedded Published Journal Article (PJA) as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.*"

Computational Statistics and Data Analysis 54 (2010) 2562–2579

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Three-stage semi-parametric estimation of T-copulas: Asymptotics, finite-sample properties and computational aspects

Dean Fantazzini *

Moscow School of Economics, Moscow State University, Russia

ARTICLE INFO

Article history:
Available online 12 February 2009

ABSTRACT

A two-stage semi-parametric estimation procedure for a broad class of copulas satisfying minimal regularity conditions has been recently proposed. In addition, a three-stage semi-parametric estimation method based on Kendall's tau in order to estimate the Student's t copula has also been designed. Its major advantage is to allow for greater computational tractability when dealing with high dimensional issues, where two-stage procedures are no more a viable choice. The asymptotic properties of this methodology are developed and its finite-sample behavior are examined via simulations. The advantages and disadvantages of this methodology are analyzed in terms of numerical convergence and positive definiteness of the estimated T-copula correlation matrix.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

The theory of copulas dates back to Sklar (1959), but its application in financial modelling is far more recent and dates back to the late 90s, instead. A copula is a function that embodies all the information about the dependence structure between the components of a random vector. When it is applied to marginal distributions which do not necessarily belong to the same distribution family, it results in a proper multivariate distribution. As a consequence, this theory enables us to incorporate a flexible modelling of the dependence structure between different variables, while allowing them to be modelled by different marginal distributions.

A semi-parametric estimation procedure for a broad class of copulas satisfying minimal regularity conditions was initially proposed by Genest et al. (1995), while a method-of-moment estimation procedure based on Kendall's tau was discussed by Genest and Rivest (1993). Genest et al. (1995) credit Oakes (1994) for the idea, although they were the first to examine its asymptotic properties. Shih and Louis (1995) also contributed to this field of the literature. However, recent empirical financial literature combined these two procedures to estimate the elliptical Student's t copula, too (see Bouyé et al. (2001), Marshall and Zeevi (2002), Cherubini et al. (2004), McNeil et al. (2005)). This new methodology entails three stages: a first stage where a non-parametric estimation of the marginal empirical distribution functions is performed; a second stage where a method-of-moment estimator based on Kendall's tau is used for the T-copula correlation matrix, and a third stage which considers Maximum Likelihood methods for the degrees of freedom ν.

This has become a common procedure among financial practitioners when dealing with high-dimensional portfolios and standard Maximum Likelihood methods cannot be used. Nevertheless, neither the asymptotic properties of such a multi-step procedure have been studied, nor its finite-sample properties. The former are important in order to build statistical tests for the structural parameters, while the latter are of interest to compare the small-sample efficiency and bias of this estimator with other estimators, such as the classical one-stage ML estimator where the margins are known. Although the

* Corresponding address: Moscow School of Economics, Moscow State University, 1, Building 61, Leninskie Gory, 119992, M.V. Lomonosov MSU, Moscow 119992, Russia. Tel.: +7 9857758816.

E-mail address: fantazzini@mse-msu.ru.

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.02.004

marginal distributions are unknown in practice, this hypothetical scenario wherein the marginals are assumed to be known, represents the ideal situation that can be used as a benchmark for comparative purposes, see, e.g., Kim et al. (2007).

What we do in this paper is to provide the asymptotic distribution for this recent semi-parametric method, and use simulations with different Data Generating Processes to examine the behavior of this estimator in small samples. The Monte Carlo study shows that this semi-parametric estimator is more efficient and less biased than the one-stage ML estimator when small samples and copulas with low degrees of freedom are of concern. We then analyze the pros and cons of this methodology in terms of numerical convergence and positive definiteness of the estimated T-copula correlation matrix. When small samples are of concern and ν is high, the number of times when the numerical maximization of the log-likelihood fails to converge is much higher for the ML method than for the three-stage Kendall's tau Moment Estimator–Canonical Maximum Likelihood (KME–CML) method. Yet, while the coverage rates at the 95% level for the ML estimates for ν do not show any particular bias or trend, the KME–CML estimates show very low coverage rates when ν becomes close to 30 and the correlations are not too strong. However, this drop in the coverage rates is large with bivariate T-copulas, only, while it is much lower when dealing with higher dimensional T-copulas, which is the usual case for real managed financial portfolios. Besides, both the ML and the KME–CML methods show high mean and median biases for the estimated correlations when the true ones are close to zero. Nevertheless, the effects on the coverage rates for the correlations are rather limited in this case.

Finally, we show that the eigenvalue method by Rousseeuw and Molenberghs (1993) has to be used to obtain a positive definite correlation matrix only when dealing with very small samples ($n < 100$) and when the true underlying process has the lowest eigenvalue close to zero. This fix induces a positive bias in the estimate of ν, but the effects on the coverage rates are rather limited. Besides, the number of times when this method has to be used quickly decreases when ν increases.

The rest of the paper is organized as follows. We introduce a recent semi-parametric estimation method of Student's t copulas based on Kendall's tau in Section 2 and we provide its asymptotics in Section 3. We present in Section 4 the results of a Monte Carlo study of the small-sample properties of this estimator, while we conclude in Section 5.

2. The three-stage KME–CML method

The study of copulas has originated with the seminal papers by Höffding (1940) and Sklar (1959) and has seen various applications in statistics and financial literature. Examples include Clayton (1978), Genest and MacKay (1986a, 1986b), Genest and Rivest (1993), Rosenberg (1998, 2003), Bouyé et al. (2001), Patton (2004, 2005), Dobric and Schmid (2006), Huarda et al. (2006), Granger et al. (2006), Fantazzini (2009, in press), Dalla Valle et al. (2008) and Karlis and Nikoloulopoulos (2008). For more details, we refer the interested reader to the recent methodological overviews by Joe (1997) and Nelsen (1999), while Cherubini et al. (2004) and McNeil et al. (2005) provide a comprehensive and detailed discussion of copula techniques for financial applications.

Genest et al. (1995) were the first to analyze a semi-parametric estimation of a bivariate copula with i.i.d. observations and to develop its asymptotic properties. Their Canonical Maximum Likelihood (CML) method differs from full Maximum Likelihood methods because no assumptions are made about the parametric form of the marginal distributions. We now briefly review this semi-parametric method since it constitutes the building block for our following analysis.

Let us consider a multivariate random sample represented by $X_i = (X_{i1}, \ldots, X_{id})$, and $i = 1, \ldots, n$, where d stands for the number of variables included and n represents the number of observations available. Let f_h be the density of the joint distribution of X. Then, by using Sklar's theorem (1959) and the relationship between the distribution and the density function we have:

$$f_h(X_i; \alpha_1, \ldots, \alpha_d, \gamma) = c(F_1(X_{1;}\alpha_1), \ldots, F_d(X_{d;}\alpha_d); \gamma) \cdot \prod_{j=1}^{d} f_j(X_j; \alpha_j) \tag{1}$$

where f_j is the univariate density of the marginal distribution F_j, c is the copula density, $\alpha_j, j = 1, \ldots, d$ is the vector of parameters of the marginal distribution F_j, while γ is the vector of the copula parameters. The CML estimation process is performed in two steps:

Definition 2.1 (*CML Copula Estimation*)**.**

1. Transform the dataset $(X_{i1}, X_{i2}, \ldots, X_{id}), i = 1, \ldots, n$ into normalized ranks $(\hat{U}_{i1}, \hat{U}_{i2}, \ldots, \hat{U}_{id})$ using the empirical distributions $\hat{U}_{ij} = F_{nj}(X_{ij})$ defined as follows:

$$F_{nj}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(X_{ij} \leq x)} \tag{2}$$

   where $\mathbb{1}_{(\cdot)}$ represents the indicator function.
2. Estimate the copula parameters by maximizing the log-likelihood:

$$\hat{\gamma}_{CML} = \arg\max \sum_{i=1}^{n} \log(c(F_{n1}(X_{i1}), \ldots, F_{nd}(X_{id})); \gamma). \tag{3}$$

Genest et al. (1995) show in Proposition A.1 that under certain regularity conditions, the semi-parametric estimator $\hat{\gamma}_{CML}$ has the following asymptotic distribution (we consider the bivariate case $d = 2$ for the sake of simplicity, see Genest et al. (1995), Section 4, for the multivariate case):

$$\sqrt{T}(\hat{\gamma}_{CML} - \gamma_0) \xrightarrow{d} \mathrm{N}\left(0, \frac{\sigma^2}{h^2}\right) \quad \text{where,} \tag{4}$$

$$\sigma^2 = \mathrm{var}[l_\gamma\,(F_1(X_1), F_2(X_2); \gamma) + W_1(X_1) + W_2(X_2)], \tag{5}$$

$$W_j(x_j) = \int \mathbb{1}_{F_j(X_j)\le u_j}\; l_{\gamma,j}\,(u_1, u_2; \gamma)\, c(u_1, u_2; \gamma)\, \mathrm{d}u_1\, \mathrm{d}u_2 \quad j = 1, 2 \quad \text{and} \tag{6}$$

$$h = -E[l_{\gamma,\gamma}\,(F_1(X_1), F_2(X_2); \gamma)] \tag{7}$$

and where $l(u_1, u_2, \gamma) = \log c(u_1, u_2; \gamma)$ and the indices $1, 2, \gamma$ denote the partial derivatives of l with respect to u_1, u_2, γ, respectively. Besides, $W_j(x_j)$ can have this alternative expression too, upon integrating by parts with respect to u_j ($j = 1, 2$):

$$W_j(x_j) = -\int \mathbb{1}_{F_j(X_j)\le u_j}\; l_\gamma\,(u_1, u_2; \gamma)\, l_j\,(u_1, u_2; \gamma)\, c(u_1, u_2; \gamma)\, \mathrm{d}u_1\, \mathrm{d}u_2. \tag{8}$$

Since the seminal work by Genest et al. (1995), it has become common practice to use semi-parametric methods with high-dimensional elliptical Student's t copulas too; see Cherubini et al. (2004) and McNeil et al. (2005) for a detailed discussion about their financial applications. Particularly, after the marginal empirical distribution functions are computed in a first stage, the correlation matrix is estimated in a second stage using a method-of-moment estimator based on Kendall's tau, while the degrees of freedom are estimated in a third stage using Maximum Likelihood methods. We remark that the fact that the margins are estimated in the first stage is not necessary to estimate the correlation matrix, since Kendall's tau is based on the number of concordances, which is unaffected by the rank transformation. However, the estimation of the margins is necessary in the last stage to estimate the degrees of freedom.

Despite the widespread use of this procedure, its asymptotic properties have not been developed yet. Before doing that and properly define this estimation method, we have to introduce Kendall's tau, its relation with linear correlation for elliptical copulas and the method-of-moment estimator based on it. A method for estimating copula parameters based on Kendall's tau has been suggested in Genest and Rivest (1993), Lindskog (2000), Lindskog et al. (2002), Cherubini et al. (2004) and McNeil et al. (2005). This dependence measure can be defined as follows:

Definition 2.2 (*Kendall's Tau*)**.** If we have $(X_1; X_2)$ and $(\tilde{X}_1; \tilde{X}_2)$ two independent and identically distributed random vectors, the population version of Kendall's tau $\tau(X_1; X_2)$ is (see Kruskal (1958)):

$$\tau(X_1, X_2) = \mathrm{E}\left[\mathrm{sign}\left((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2)\right)\right]. \tag{9}$$

The generalization of Kendall's tau to $n > 2$ dimensions can be analogous to the procedure for linear correlation, where we have a $n \times n$ matrix of pairwise correlations. However, other n-variate generalization of Kendall's tau are possible, see Clemen and Jouini (1996) and Barbe et al. (1996) for a discussion about this issue. The main properties of this dependence measure and relative proofs are reported in Embrechts et al. (2002). Besides, Kendall's tau can be expressed in terms of copulas, thus simplifying calculus, see e.g. Nelsen (1999, p. 127).

$$\tau(X_1, X_2) = 4\int_0^1 \int_0^1 C(u_1, u_2)\, \mathrm{d}C(u_1, u_2) - 1. \tag{10}$$

Evaluating Kendall's tau requires the evaluation of a double integral and for elliptical copulas like the T-copula this is not an easy task: this problem was solved by Lindskog et al. (2002) who proved that Kendall's tau for elliptical distributions is given by

$$\tau(X_1, X_2) = \frac{2}{\pi} \arcsin \rho_{X_1X_2} \tag{11}$$

where $\rho_{X_1X_2}$ is the copula correlation parameter. It follows easily from the previous equation that the correlation can be retrieved by the inversion of Kendall's tau: in the case of the T-copula, the estimated correlation matrix $\hat{\Sigma}$ has off-diagonal elements given by $\hat{\rho}_{ij} = \sin(\pi\,\hat{\tau}_{ij}/2)$, see, e.g. Wang and Wells (2000), Kowalczyka and Niewiadomska-Bugaj (2002) and Genest et al. (2006).

More formally, Kendall's tau moment estimator (KME) can be defined as follows: let us consider the population version of Kendall's tau (9) and its relationship with copula parameters (10) to build a moment function of the type

$$\mathrm{E}\,[\psi\,(X_1, X_2; \gamma_0)] = 0. \tag{12}$$

Then, we can construct an empirical estimate of Kendall's tau pairwise correlation matrix and use relationship (10) to infer an estimate of the relevant parameters of the copula. This is a method-of-moments estimate because the true moment (9) is replaced by its empirical analogue,

$$\binom{n}{2}^{-1} \sum_{1\le i<s\le n} \operatorname{sign}\left((x_{i,1}-\tilde{x}_{s,1})(x_{i,2}-\tilde{x}_{s,2})\right) \tag{13}$$

and (10) is then used to estimate the copula parameters. For example, when using elliptical copulas and Eq. (11), this moment function becomes

$$\mathrm{E}\left[\psi\left(X_1, X_2; \rho_{X_1,X_2}\right)\right] = \mathrm{E}\left[\rho_{X_1,X_2} - \sin(\pi\tau(X_1,X_2)/2)\right] = 0 \tag{14}$$

where $\rho_{X_1X_2}$ is the copula correlation parameter.

There are cases when the copula parameter vector has different kinds of parameters and only some of them can be expressed as a function of Kendall's tau. This is the case for the T-copula which is parameterized by the correlation matrix $\boldsymbol{\Sigma}$ and the degrees of freedom ν, but only the former has a direct relationship with Kendall's tau. This is the copula of the multivariate Student's t-distribution and we can derive its density function by using Sklar's theorem (1959) and the relationship between the distribution and the density function:

$$\begin{aligned} c(t_\nu(x_1),\ldots,t_\nu(x_d)) &= \frac{f^{Student}(x_1,\ldots,x_d)}{\prod_{j=1}^{d} f_j^{Student}(x_i)} \\ &= |\boldsymbol{\Sigma}|^{-1/2}\,\frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left[\frac{\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)}\right]^d \frac{\left(1+\frac{\zeta'\boldsymbol{\Sigma}^{-1}\zeta}{\nu}\right)^{-\frac{\nu+d}{2}}}{\prod_{i=1}^{d}\left(1+\frac{\zeta_i^2}{2}\right)^{-\frac{\nu+1}{2}}} \end{aligned} \tag{15}$$

where $\zeta = (t_\nu^{-1}(u_1),\ldots,t_\nu^{-1}(u_d))^\top$ is the vector of univariate Student's t inverse distribution functions, ν is the degrees of freedom, $u_j = t_\nu(x_j)$ is the univariate Student's t cumulative distribution function, while $\boldsymbol{\Sigma}$ is the correlation matrix. In this situation, Bouyé et al. (2001) and McNeil et al. (2005) have suggested the following estimation procedure:

Definition 2.3 (*Three-stage KME–CML Copula Estimation*)**.**

(1) Transform the dataset $(X_{i1}, X_{i2}, \ldots, X_{id})$ into normalized ranks $(F_{n1}(X_{i1}), F_{n2}(X_{i2}), \ldots, F_{nd}(X_{id}))$, using the empirical distribution function, and which are approximately uniform variates.

(2) Collect all pairwise estimates of the sample Kendall's tau given by (13) in an empirical Kendall's tau matrix $\hat{R}^\tau$ defined by $\hat{R}^\tau_{jk} = \tau(F_{nj}(X_j), F_{nk}(X_k))$, and then construct the correlation matrix using this relationship $\hat{\Sigma}_{j,k} = \sin(\frac{\pi}{2}\hat{R}^\tau_{j,k})$, where the estimated parameters are the $q = d\cdot(d-1)/2$ correlations $[\hat{\rho}_1,\ldots\hat{\rho}_q]'$. Since there is no guarantee that this componentwise transformation of the empirical Kendall's tau matrix is positive definite, when needed, $\hat{\boldsymbol{\Sigma}}$ can be adjusted to obtain a positive definite matrix using a procedure such as the eigenvalue method of Rousseeuw and Molenberghs (1993) or other methods.

(3) Look for the CML estimator of the degrees of freedom $\hat{\nu}_{CML}$ by maximizing the log-likelihood function of the T-copula density:

$$\hat{\nu}_{CML} = \arg\max \sum_{i=1}^{n} \log c_{T\text{-}copula}(F_{n1}(X_{i1}),\ldots,F_{nd}(X_{id});\ \hat{\Sigma},\nu). \tag{16}$$

3. Asymptotic properties of the three-stage method

The second step in the previous Definition 2.3 corresponds to a method-of-moments estimation based on q moments and Kendall tau rank correlations estimated with empirical distribution functions: we stress again that the estimation of the margins is not necessary to estimate the correlation matrix, since Kendall's tau is unaffected by the rank transformation. We can therefore generalize Eq. (14) and build a $q\times 1$ moments vector ψ for the parameter vector $\boldsymbol{\theta}_0 = [\rho_1,\ldots,\rho_q]'$ as reported below:

$$\psi\left(F_1(X_1),\ldots,F_n(X_n);\boldsymbol{\theta}_0\right) = \begin{pmatrix} \mathrm{E}\left[\psi_1\left(F_1(X_1),F_2(X_2);\ \rho_1\right)\right] \\ \vdots \\ \mathrm{E}\left[\psi_q\left(F_{d-1}(X_{d-1}),F_d(X_d);\ \rho_q\right)\right] \end{pmatrix} = 0. \tag{17}$$

Then these theorems follow (the proofs are reported in Appendix A in the technical report by Fantazzini (2009)):

Theorem 3.1 (*Consistency of* $\hat{\theta}$). *Let us assume that* $(X_{i1}, \ldots, X_{id})$ *are i.i.d random variables with dependence structure given by* $c(u_{i,1}, \ldots, u_{i,d}; \Sigma_0, \nu_0)$. *Suppose that*

(i) *the parameter space* Θ *is a compact subset of* $\mathbb{R}^q$,
(ii) *the q-variate moment vector* $\psi\ (F_1(X_1), \ldots, F_d(X_d); \boldsymbol{\theta}_0)$ *is continuous in* $\boldsymbol{\theta}_0$ *for all* X_j,
(iii) $\psi\ (F_1(X_1), \ldots, F_d(X_d); \boldsymbol{\theta})$ *is measurable in* X_j *for all* $\boldsymbol{\theta}$ *in* Θ,
(iv) $\mathrm{E}\left[\psi\ (F_1(X_1), \ldots, F_d(X_d); \boldsymbol{\theta})\right] \neq \mathbf{0}$ *for all* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ *in* Θ,
(v) $\mathrm{E}\left[\sup_{\boldsymbol{\theta}\in\Theta} \|\psi\ (F_1(X_1), \ldots, F_d(X_d); \boldsymbol{\theta})\|\right] < \infty$,

then $\hat{\boldsymbol{\theta}} \xrightarrow{d} \boldsymbol{\theta}_0$ *as* $n \to \infty$.

Theorem 3.2 (*Consistency of* $\hat{\nu}_{CML}$). *Let the assumptions of the previous theorem hold, as well as the regularity conditions reported in Proposition A.1 in Genest et al. (1995). Then* $\hat{\nu}_{CML} \xrightarrow{d} \nu_0$ *as* $n \to \infty$.

The asymptotic normality is not straightforward, since we use a three-step procedure where we perform a different kind of estimation at the second and third stage. A possible solution is to consider the CML used in the third stage as a special method-of-moment estimator. Just note that the CML estimator is defined by the derivative of the log-likelihood function with respect to the degrees of freedom:

$$\frac{\partial l(\cdot;\nu)}{\partial \nu} = \sum_{i=1}^{n} l_\nu\left(F_{n1}(X_{i1}), \ldots, F_{nd}(X_{id}); \hat{\boldsymbol{\Sigma}}, \hat{\nu}\right) = 0. \tag{18}$$

Dividing both sides by n yields the definition of the method of moments estimator:

$$\frac{1}{n}\sum_{i=1}^{n} l_\nu\left(F_{n1}(X_{i1}), \ldots, F_{nd}(X_{id}); \hat{\boldsymbol{\Sigma}}, \hat{\nu}\right) = \frac{1}{n}\sum_{i=1}^{n} \psi_\nu(F_{n1}(X_{i1}), \ldots, F_{nd}(X_{id}); \hat{\boldsymbol{\Sigma}}, \hat{\nu}) = 0.$$

Thus, the CML estimator is a simple method-of-moments (MM) estimator with the score as its moment function, where the sample mean of the score is equal to the population mean of the score. We remind that the MM estimator $\hat{\theta}$ of θ_0 based on the k moment restrictions $E[\psi(Z_i, \theta_0)] = 0$ and where $Z_i = (Y_i, X_i)$ is a vector of endogenous and explanatory variables for observation $i = 1, \ldots, n$, is the solution to the problem

$$\frac{1}{n}\sum_{i=1}^{n} \psi(Z_i, \hat{\theta}) = E[\psi(Z_i, \theta_0)] = 0.$$

The method-of-moments estimator is simply the value of θ which sets the sample moments (sample means) equal to the true population moments. Since the number of parameters is equal to k, the definition of the simple MM estimator is the solution (the root) of the k (possibly nonlinear) equation system with respect to $\hat{\theta}$, see, e.g., Greene (2008) for more details. Therefore, we can now apply the method-of-moments asymptotics together with multivariate rank statistics.

Let us define the sample moments vector $\Psi_{KME\text{-}CML}$ for the parameter vector $\hat{\boldsymbol{\Xi}} = [\hat{\rho}_1, \ldots \hat{\rho}_q, \hat{\nu}]'$ as follows:

$$\Psi_{KME-CML}\left(F_{n1}(X_{i1}), \ldots, F_{nd}(X_{id}); \hat{\boldsymbol{\Xi}}\right) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} \psi_1\left(F_{n1}(X_{i1}), F_{n2}(X_{i2});\ \hat{\rho}_1\right) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} \psi_q\left(F_{n(d-1)}(X_{i(d-1)}), F_{nd}(X_{id});\ \hat{\rho}_q\right) \\ \frac{1}{n}\sum_{i=1}^{n} \psi_\nu\left(F_{n1}(X_{i1}), \ldots, F_{nd}(X_{id}); \hat{\boldsymbol{\Sigma}}, \hat{\nu}\right) \end{pmatrix} = 0.$$

Let us also define the population moments vector with a correction to take the non-parametric estimation of the marginals into account, together with its variance (see Genest et al. (1995), § 4):

$$\Delta_0 = \begin{pmatrix} \psi_1\ (F_1(X_1), F_2(X_2);\ \rho_1) \\ \vdots \\ \psi_q\left(F_{d-1}(X_{d-1}), F_d(X_d);\ \rho_q\right) \\ \psi_\nu\ (F_1(X_1), \ldots, F_d(X_d); \boldsymbol{\Sigma_0}, \nu_0) + \sum_{j=1}^{d} W_{j,\nu}(X_j) \end{pmatrix} = 0, \tag{19}$$

$$\boldsymbol{\Upsilon}_0 \equiv \mathrm{var}[\boldsymbol{\Delta}_0] = \mathrm{E}[\boldsymbol{\Delta}_{KME-CML}\ \boldsymbol{\Delta}_{KME-CML}{}'] \tag{20}$$

where

$$W_{j,\nu}(X_j) = \int \mathbb{1}_{F_j(X_j)\le u_j} \frac{\partial^2}{\partial\nu\partial u_j} \log c(u_1,\ldots,u_d)\,\mathrm{d}C(u_1,\ldots,u_d). \tag{21}$$

Note that the population moments used to estimate the correlations are not affected by the marginals empirical distribution functions, since Kendall's tau is invariant under strictly increasing marginal transformations. Then this theorem follows:

Theorem 3.3 (*Asymptotic Distribution Three-stages KME–CML Method*)**.** *Let the assumptions of the previous theorems hold. Assume further that* $\frac{\partial \boldsymbol{\Psi}_{KME--CML}(\cdot;\boldsymbol{\Xi})}{\partial \boldsymbol{\Xi}'}$ *is* $O(1)$ *and uniformly negative definite, while* $\boldsymbol{\Upsilon}_0$ *is* $O(1)$ *and uniformly positive definite. Then, the three-stage KME–CML estimator verifies the properties of asymptotic normality:*

$$\sqrt{T}(\hat{\boldsymbol{\Xi}} - \boldsymbol{\Xi}_0) \xrightarrow{d} \mathrm{N}\left(0, \mathrm{E}\left[\frac{\partial \boldsymbol{\Psi}_{KME-CML}}{\partial \boldsymbol{\Xi}'}\right]^{-1} \boldsymbol{\Upsilon}_0\, \mathrm{E}\left[\frac{\partial \boldsymbol{\Psi}_{KME-CML}}{\partial \boldsymbol{\Xi}'}\right]^{-1\prime}\right). \tag{22}$$

The previous asymptotic properties are still valid when dealing with multivariate heteroscedastic time series models, where one first obtains consistent estimates of the parameters of each univariate marginal time-series, and computes the corresponding residuals. These are then used to estimate the joint distribution of the multivariate error terms, which is specified using a copula. Such a result is a straightforward application of Theorems 1 and 2 in Kim et al. (2008), who make use of, and build upon, recent elegant results of Koul and Ling (2006) and Koul (2002) for these models. Besides, we remark that essentially the same result of Kim et al. (2008) is also used in Chen and Fan (2006) but without proofs.

Theorem 3.4 (*Asymptotic Distribution Three-stages KME–CML Method for Multivariate Heteroscedastic Time Series Models*)**.** *Let the regularity conditions* (i)–(v) *reported in Theorem 3.1 hold, together with conditions* (A.1)–(A.9) *in Kim et al. (2008). Then, the three-stage KME–CML estimator verifies the properties of asymptotic normality defined in* (22)*.*

Conditions (A.1)–(A.4) do not involve the time-series aspects of the model and they were also used in Kim et al. (2005) for the linear regression case with iid errors. Particularly, conditions (A.1)–(A.2) require that the copula density has continuous partial derivatives up to third order, that they are finite and their second moment is finite. Condition (A.3) is a technical condition on the partial derivatives, while condition (A.4) requires that the conditions of Proposition A.1 in Genest et al. (1995) are satisfied. The conditions (A.5)–(A.8) are technical conditions taken from an earlier work by Koul (2002) and Koul and Ling (2006), while condition (A.9) is a mild one: for example, if the conditional mean is a function of past values of the time series which is strictly stationary and ergodic, then the summand forms a strictly stationary and ergodic process with mean zero, and hence (A.9) would be satisfied (see Taniguchi and Kakizawa (2000), Theorems 1.3.3–1.3.5, and Kim et al. (2008)).

Since the main results on the properties of the proposed semi-parametric method are asymptotic, a large-scale simulation study was carried out to compare its properties with other competing ones.

4. Finite-sample properties and computational aspects

In this section we present the results of a Monte Carlo study of the small-sample properties of the estimator discussed above for a representative collection of Data Generating Processes (DGPs). Furthermore, we analyze the pros and cons of this methodology in terms of numerical convergence and positive definiteness of the estimated T-copula correlation matrix.

We consider the following possible DGPs:

(1) We examine the case that two variables have a bivariate Student's t copula, with the copula linear correlation ρ ranging between -0.9 and 0.9 (step 0.1). We examine different values for the degree of freedom ν, too, ranging between 3 and 30 (step 1). The former corresponds to a case of strong tail dependence, that is there is a high probability of observing an extremely large observation on one variable, given that the other variable has yielded an extremely large observation. The latter exhibits low tail dependence, and corresponds to the case where the T-copula becomes almost indistinguishable from a Normal copula. We remark that the Student's t copula generates positive tail dependence, while the normal copula generates zero tail dependence, instead. See Cherubini et al. (2004) for more details. We consider two possible data situations: $n = 50$ and $n = 500$.

(2) We examine the case where ten variables have a multivariate Student's t copula, with the copula correlation matrix $\boldsymbol{\Sigma}$ equal to:

We choose this correlation matrix because its lowest eigenvalue is very close to zero (0.0786) and it allows us to study the effect that the eigenvalue method by Rousseeuw and Molenberghs (1993) has on the limiting distribution of

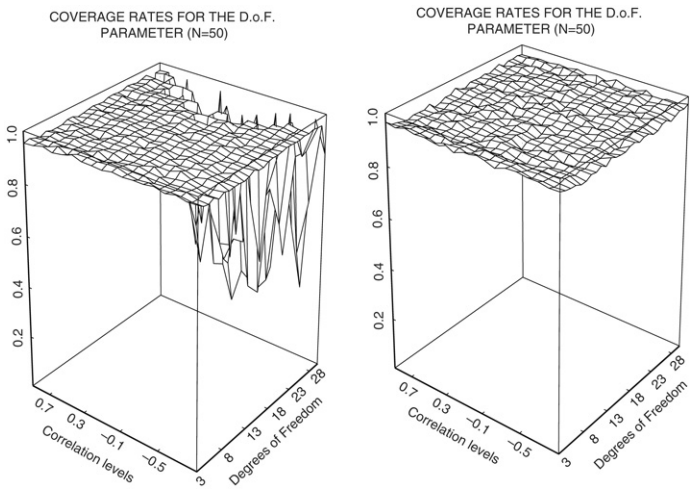


Fig. 1. Coverage rates for the 95% confidence intervals based on a normal approximation of the bivariate T-copula degrees of freedom ν, for $n = 50$. The first plot refer to the KME–CML method, while the second one to the ML method.

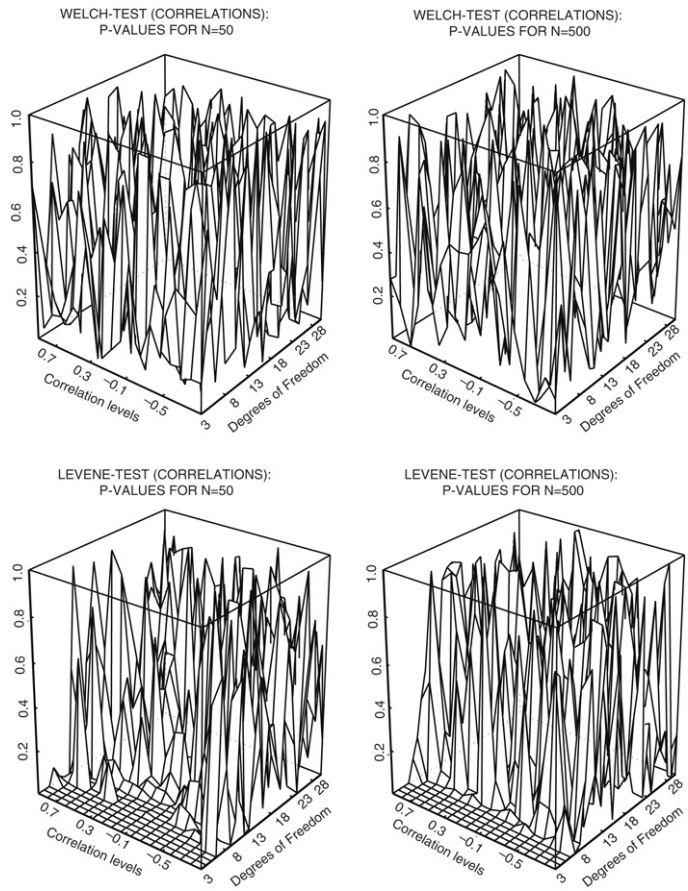


Fig. 2. The four plots report the p-values for the Welch tests (first row) and Levene tests (second row). The two samples compared in the tests are the *correlation parameters* of the bivariate T-copulas, estimated with the ML and KME–CML methods.

the KME–CML estimator. Furthermore, we examine different values for the degree of freedom ν, too, ranging between 3 and 30 (step 1), as well as two possible data situations: $n = 50$ and $n = 500$.

(3) We examine the case that ten variables have a multivariate Student's t copula, with the copula correlation matrix $\mathbf{\Sigma}$ equal to:

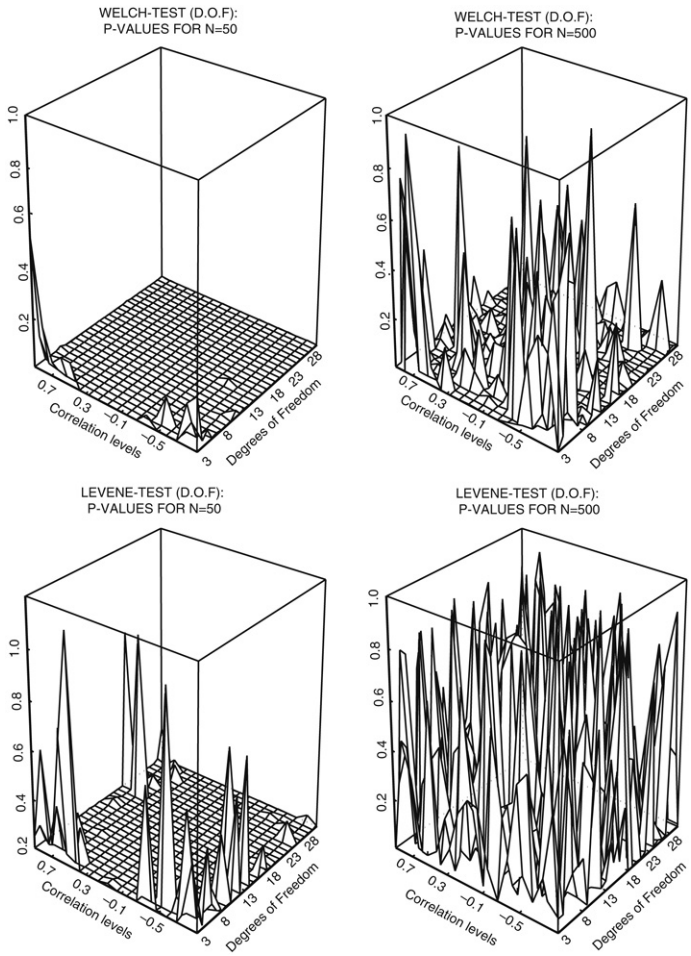


Fig. 3. The four plots report the p-values for the Welch tests (first row) and Levene tests (second row). The two samples compared in the tests are the *degrees of freedom parameters* of the bivariate T-copulas, estimated with the ML and KME–CML methods.

Table 1
Correlation matrix T-copula with lowest eigenvalue equal to 0.0768.

1	−0.15	−0.15	−0.15	−0.15	−0.14	−0.09	−0.03	0.05	0.13
−0.15	1	−0.15	−0.15	−0.15	−0.13	−0.08	−0.02	0.06	0.14
−0.15	−0.15	1	−0.15	−0.15	−0.12	−0.07	−0.01	0.07	0.15
−0.15	−0.15	−0.15	1	−0.15	−0.11	−0.06	0.01	0.08	0.15
−0.15	−0.15	−0.15	−0.15	1	−0.10	−0.05	0.02	0.09	0.15
−0.14	−0.13	−0.12	−0.11	−0.10	1	−0.04	0.03	0.10	0.15
−0.09	−0.08	−0.07	−0.06	−0.05	−0.04	1	0.04	0.11	0.15
−0.03	−0.02	−0.01	0.01	0.02	0.03	0.04	1	0.12	0.15
0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	1	0.15
0.13	0.14	0.15	0.15	0.15	0.15	0.15	0.15	0.15	1

This is the correlation matrix of the returns of the first 10 stocks belonging to the Dow Jones Industrial Index, observed between 18/11/1988 and 20/11/2003. Furthermore, we examine different values for the degree of freedom ν, too, ranging between 3 and 30, as well as two possible data situations: $n = 50$ and $n = 500$.

The estimators considered are the three-stage KME–CML method described in Section 2, and the Maximum Likelihood estimator computed with given marginals, in order to assess the loss in efficiency associated with absence of knowledge of the marginals. We also considered the two-stage CML method (where $\mathbf{\Sigma}$ and ν are estimated jointly by ML), which delivered results in-between the KME–CML and ML methods, as expected. Therefore, we do not report its results for the sake of interest and space.

We generated 1000 Monte Carlo samples for each copula specification previously described and we estimated the T-copula parameters using both the KME–CML method and the ML method. Then, we computed the following evaluation statistics:

Table 2
Correlation matrix T-copula — returns Dow Jones Industrial Index.

1	0.21	0.33	0.22	0.36	0.30	0.37	0.34	0.31	0.47
0.21	1	0.20	0.15	0.27	0.18	0.18	0.31	0.20	0.21
0.33	0.20	1	0.16	0.32	0.28	0.40	0.33	0.17	0.42
0.22	0.15	0.16	1	0.20	0.16	0.18	0.20	0.27	0.20
0.36	0.27	0.32	0.20	1	0.32	0.33	0.55	0.33	0.35
0.30	0.18	0.28	0.16	0.32	1	0.28	0.32	0.26	0.31
0.37	0.18	0.40	0.18	0.33	0.28	1	0.35	0.23	0.40
0.34	0.31	0.33	0.20	0.55	0.32	0.35	1	0.31	0.35
0.31	0.20	0.17	0.27	0.33	0.26	0.23	0.31	1	0.30
0.47	0.21	0.42	0.20	0.35	0.31	0.40	0.35	0.30	1

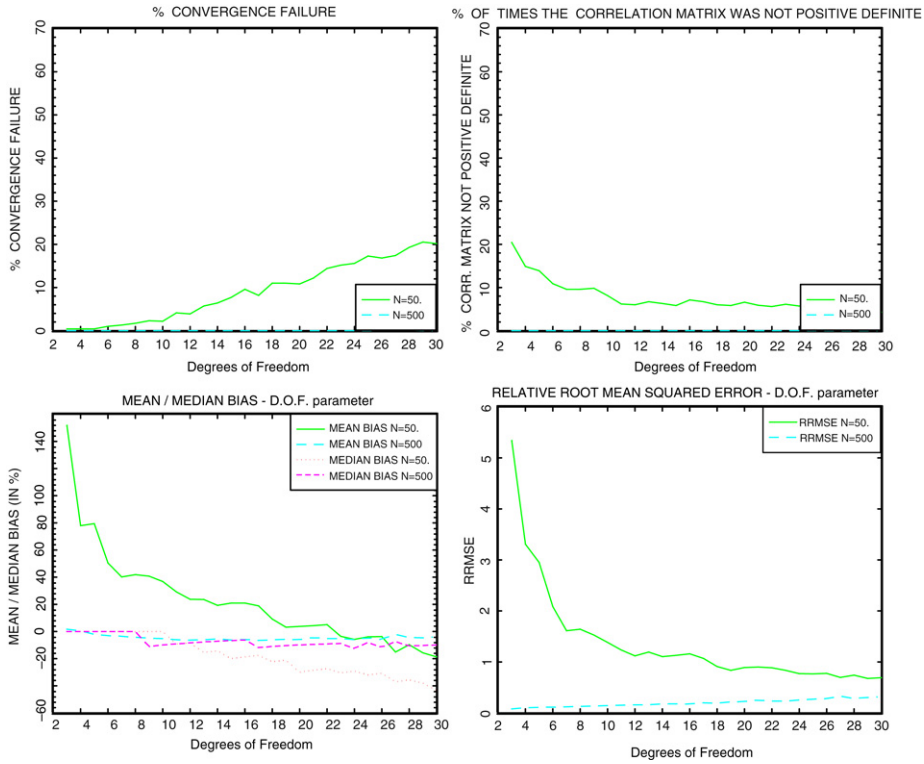


Fig. 4. The first plot reports the % of convergence failures when maximizing the log-likelihood. The second reports the % of times when the correlation matrix was not positive definite and the eigenvalue method was used. The third and the fourth plots report the Mean bias (in %), the Median bias (in %) and the Relative RMSE of the d.o.f. parameter, for the KME–CML method for the 10-variate T-copula with correlation matrix reported in Table 1.

- Mean bias (in %);
- Median bias (in %);
- Relative RMSE of the correlation and degrees of freedom parameters with respect to the true parameters;
- % of convergence failures when maximizing the log-likelihood. We used the `maxlik` library of the GAUSS software and a convergence tolerance for the gradient of the estimated parameters equal to 1e−5;
- % of times when the correlation matrix $\mathbf{\Sigma}$ was not positive definite and the eigenvalue method by Rousseeuw and Molenberghs (1993) was used;
- The coverage rate for the 95% confidence intervals of the T-copula parameters based on a normal approximation.
- t-tests under the null hypothesis that the sample mean across estimated parameters is equal to the true value.

Furthermore, we employed ANOVA based tests to test equality of the means and variances across the sample estimates delivered by the two considered methodologies. Particularly, we considered the following tests:

- The standard *F-test for the equality of variances of two groups*;
- The Levene test (1960) to assess the equality of variances in different samples. Levene test and the previous F-test are often used before a comparison of means: when Levene test is significant, modified procedures are used that do not

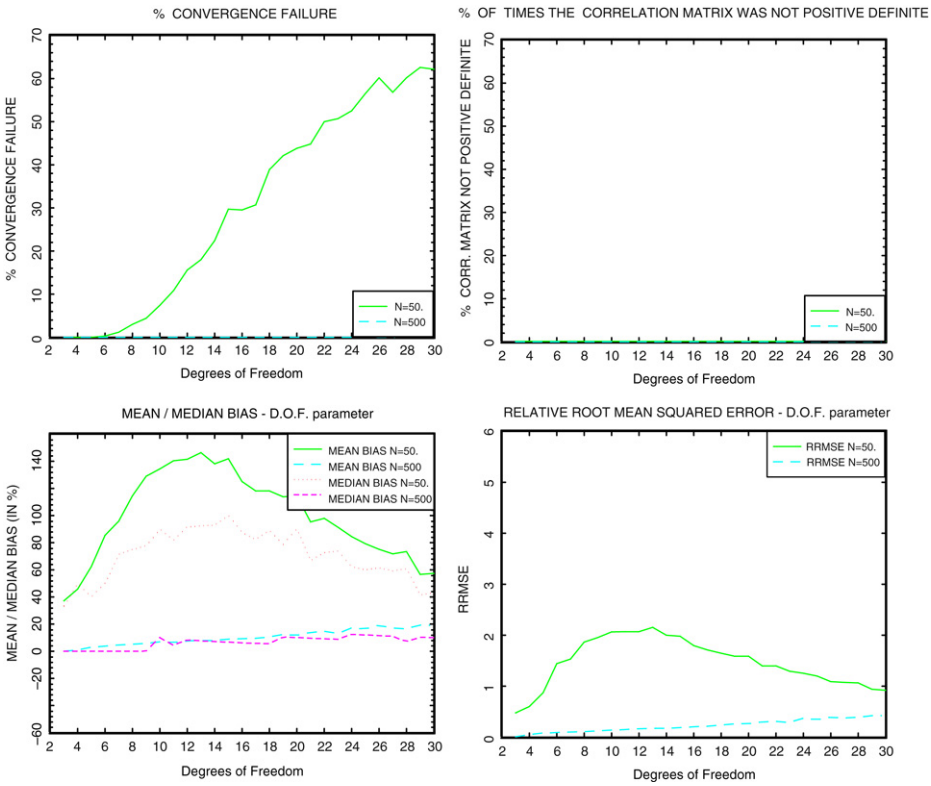


Fig. 5. The first plot reports the % of convergence failures when maximizing the log-likelihood. The second reports the % of times when the correlation matrix was not positive definite and the eigenvalue method was used. The third and the fourth plots report the Mean bias (in %), the Median bias (in %) and the Relative RMSE of the d.o.f. parameter, for the ML method for the 10-variate T-copula with correlation matrix reported in Table 1.

assume equality of variance:

$$F_{Levene} = \frac{N-G}{G-1} \cdot \frac{\sum_{g=1}^{G} n_g (\bar{Z}_g - \bar{Z})^2}{(G-1) \sum_{g=1}^{G} \sum_{j=1}^{n_g} (Z_{gj} - \bar{Z}_g)^2}$$

where $Z_{gj} = |x_{gj} - \bar{x}_g|$, $g = 1, \ldots, G$, $\bar{Z}$ is the overall mean of all Z_{gj}, $\bar{Z}_g$ is the mean of the Z_{gj} for group g, while n_g is the number of observations in group g. The F-statistic for the Levene test has an approximate F-distribution with $G = 1$ numerator degrees of freedom and $N - G$ denominator degrees of freedom under the null hypothesis of equal variances in each subgroup (Levene, 1960).

- The standard *F-test for the equality of means in two groups*;
- If the two groups' variances are heterogeneous, it is advisable to use the Welch (1951) *version of the previous test statistic*. The basic idea is to form a modified F-statistic that accounts for the unequal variances. Using the Cochran (1937) weight function,

$$w_g = n_g / s_g^2$$

where s_g^2 is the sample variance in group g, $g = 1, \ldots, G$, we may form the modified F-statistic as follows:

$$F^* = \frac{\sum_{g=1}^{G} w_g (\bar{x}_g - \bar{x}^*)^2 / (G-1)}{1 + \frac{2(G-2)}{G^2-1} \sum_{g=1}^{G} \frac{(1-h_g)^2}{n_g - 1}}$$

where $\bar{x}_g$ is the sample mean within group g, h_g is a normalized weight and $\bar{x}^*$ is the weighted grand mean,

$$h_g = w_g / \left(\sum_{g=1}^{G} w_k \right),$$

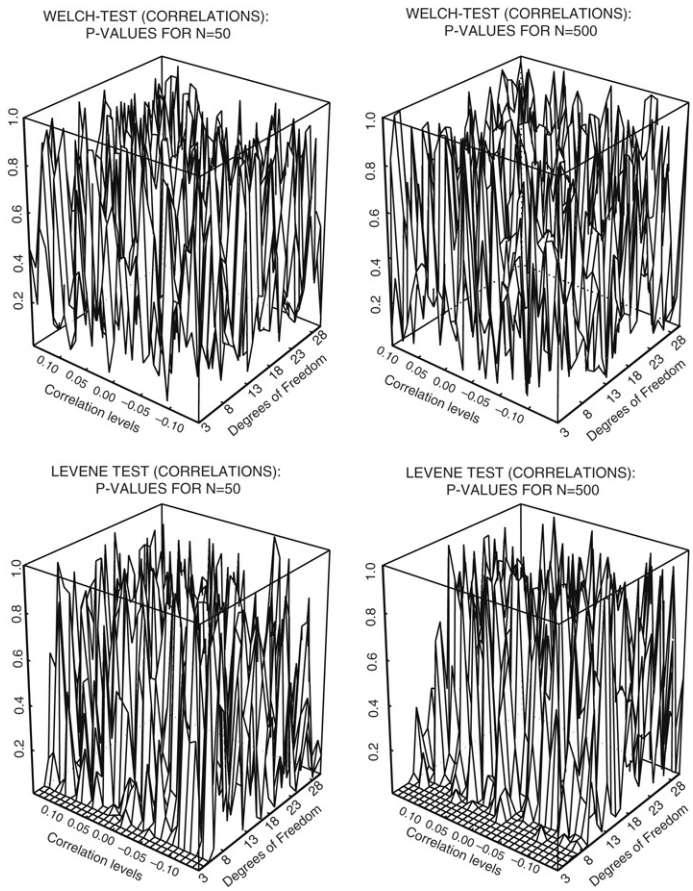


Fig. 6. The four plots report the p-values for the Welch tests (first row) and Levene tests (second row). The two samples compared in the tests are the *correlation parameters* of the T-copula with correlation matrix reported in Table 1, estimated with the ML and KME–CML methods.

$$\bar{x}^* = \sum_{g=1}^{G} h_g \bar{x}_g .$$

The numerator of the adjusted statistic is the weighted between-group mean squares and the denominator is the weighted within-group mean squares. Under the null hypothesis of equal means but possibly unequal variances, F^* has an approximate F-distribution with $(G-1, DF^*)$ degrees of freedom, where

$$DF^* = \frac{(G^2-1)}{3\sum_{g=1}^{G} \frac{(1-h_g)^2}{n_g-1}}.$$

Note that for $G = 2$ groups, this test reduces to the Satterthwaite (1946) test.

The standard F-tests for the equality of means and variance of two groups are not reported in the paper, but are available from the author upon request.

4.1. Simulation results: Bivariate T-copula

For the sake of interest and space, we report in the main manuscript only a limited set of results, while most of them are reported in the Technical Report by Fantazzini (2009). Particularly, Fig. 1 reports the coverage rates for the 95% confidence intervals based on a normal approximation of the bivariate T-copula degrees of freedom, for $n = 50$, while Figs. 2 and 3 report the p-values for the Welch tests and Levene tests, under the null hypothesis that the sample means and variances of the KME–CML and ML methods are equal, respectively.

The remaining set of results are reported in the Technical report by Fantazzini (2009): particularly, Figures 12–13 there report the Mean bias (in %), Median bias (in %) and Relative RMSE of the correlation and degrees of freedom parameters,

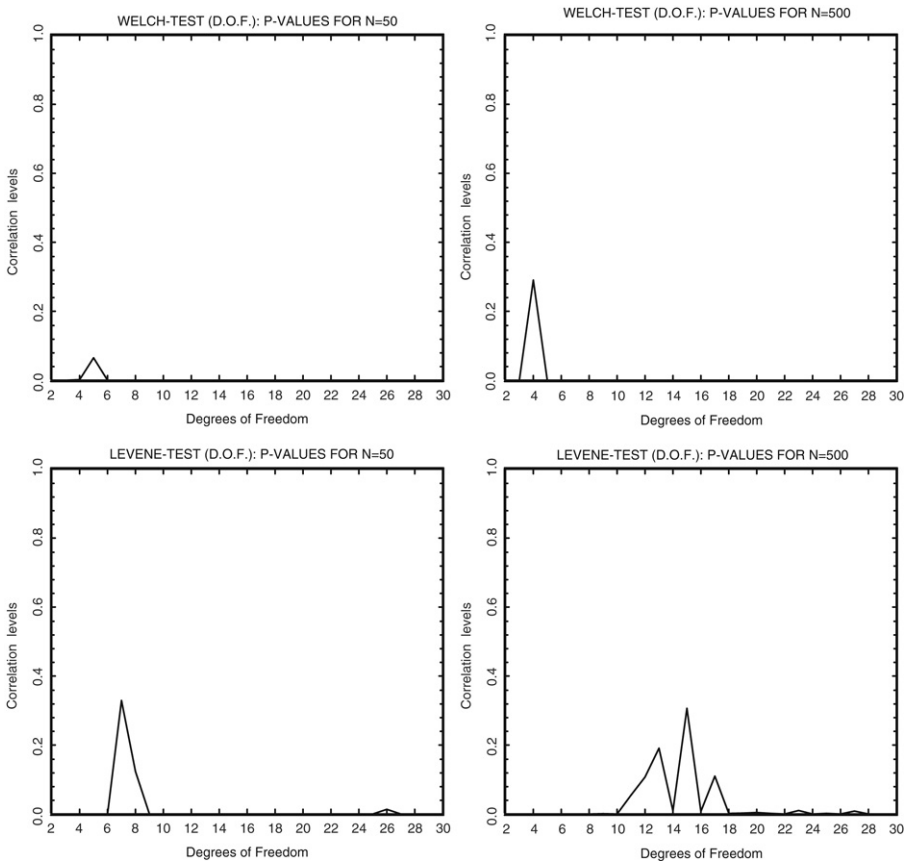


Fig. 7. The four plots report the p-values for the Welch tests (first row) and Levene tests (second row). The two samples compared in the tests are the *degrees of freedom parameters* of the T-copula with correlation matrix reported in Table 1, estimated with the ML and KME–CML methods.

respectively, for the KME–CML method, across different correlation levels and degrees of freedom parameters, as well as different data samples. Figure 14 in Fantazzini (2009) reports the coverage rate (in %) for a large sample 95% confidence interval based on a normal approximation for ν and ρ, together with the % of convergence failures when maximizing the log-likelihood. Figures 15–17 in Fantazzini (2009) report the same simulation statistics but for the ML method. Finally, Figures 26 (first column)–27 (first column) report the t-tests under the null hypothesis that the sample means across estimated correlations and degrees of freedom are equal to the true values, respectively.

The simulation studies show some interesting results:

With regard to the ***degrees of freedom*** "ν", when considering a small sample with $n = 50$ observations, both the ML and the KME–CML methods show poor results, with high mean and median biases, as well as high percentage of numerical convergence failures. However, if a low value for ν is used, the KME–CML method shows better results than the ML estimator, while the reverse is true if ν is high, that is if the T-copula becomes no more distinguishable from a Normal copula (the T-copula tends to the Normal copula when $\nu \to \infty$). The same results are confirmed by the t-tests in Fig. 27 in Fantazzini (2009) (first column).

When $n = 50$ and ν is high, the % of time when the numerical maximization of the log-likelihood failed to converge is much higher for the ML method than for the KME–CML method (over 50% vs 30%, respectively). However, while the coverage rates at the 95% level for the ML estimates do not show any particular bias or trend, the KME–CML estimates for ν show very low coverage rates when ν becomes close to 30 and the correlation is not too strong (between -0.5 and 0.5). Therefore, the lower number of convergence failures for the KME–CML method comes at a cost: the estimates show much stronger negative mean (and median) biases than the ML method, around the 70% of the true value vs the 50%, respectively. As a consequence, the confidence intervals based on a normal approximation are very poor. When a larger sample with $n = 500$ observations is considered, the two methods perform rather well and the ML estimator shows better properties, as expected. Moreover, the problem with the coverage rates for the KME–CML method disappears.

As for the ***correlation*** "ρ", the two methods produce consistent estimates already with very small samples and the analysis reveals no major difference between the two. However, if the variables are close to be uncorrelated, both methods present small biases and RRMSE that decrease when n increases (see Figures 12 and 15 in the Technical Appendix in Fantazzini (2009)).

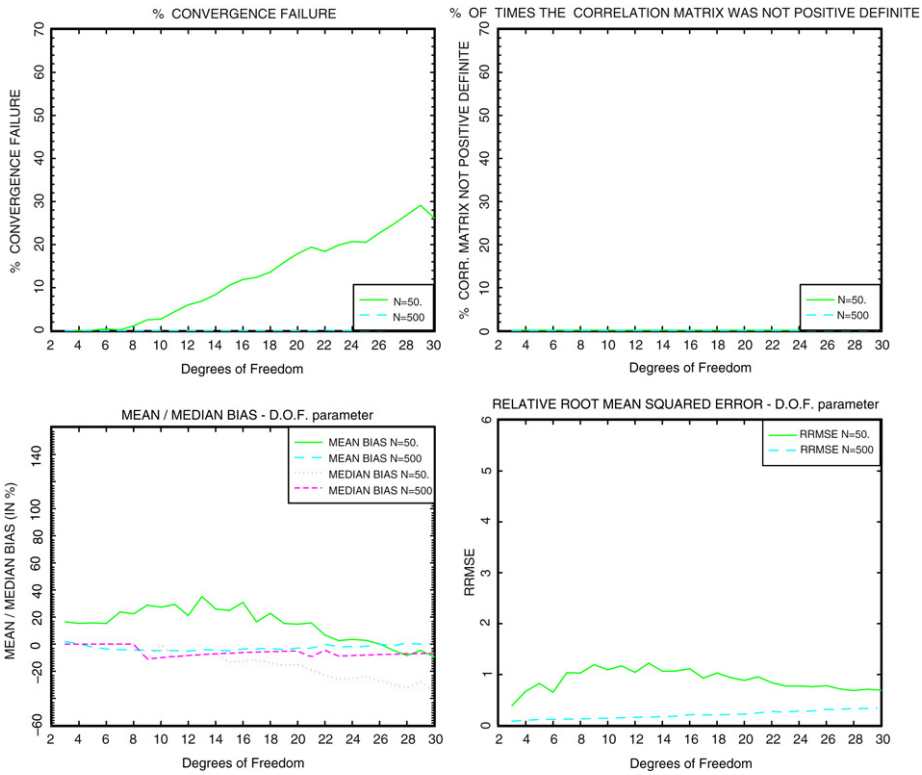


Fig. 8. The first plot reports the % of convergence failures when maximizing the log-likelihood. The second reports the % of times when the correlation matrix was not positive definite and the eigenvalue method was used. The third and the fourth plots report the Mean bias (in %), the Median bias (in %) and the Relative RMSE of the d.o.f. parameter, for the KME–CML method for the 10-variate T-copula with correlation matrix reported in Table 2.

With regard to ***ANOVA tests***, when the correlation parameters are of concern, the Welch tests in Fig. 2 show that the sample estimates delivered by the two methods (KME–CML and ML) have means which are not statistically different. A similar evidence is delivered by the Levene tests for the variances which, however, are statistically different when the true degrees of freedom ν are low. As for the degrees of freedom parameter, the Welch and Levene tests in Fig. 3 point out that for $n = 50$ the two methods deliver statistically different estimates, particularly with high values for ν and low correlations, thus confirming the previous evidence with coverage rates.

As regards the ***computational aspects***, as anticipated in the previous points, the ML method shows much higher convergence failures than the the KME–CML method when $n = 50$ and ν is higher than 10, while when $n = 500$ the numerical performances of the two methods are quite close. Furthermore, the analysis shows that when ν is high and the KME–CML is employed, the % of convergence failures is higher when T-copulas with *weaker* correlations are considered, reaching the maximum when the variables are uncorrelated (see the lower plots in Figure 14 in Fantazzini (2009)). This pattern is not present for the ML method, instead. This result together with the stronger negative biases of $\hat{\nu}_{KME-CML}$, helps us to explain the U-shape drop in coverage rates observed in Fig. 1 for the KME–CML method.

4.2. Simulation results: 10-variate T-copula — Ill-specified correlation matrix

For the sake of interest and space, we report in Figs. 4 and 5 the % of convergence failures when maximizing the log-likelihood, the % of times when the correlation matrix was not positive definite and the eigenvalue method was used, as well as the Mean bias (in %), the Median bias (in %) and the Relative RMSE of the d.o.f. parameter, for both the KME–CML and ML methods. Figs. 6 and 7 report the p-values for the Welch tests and Levene tests.

The remaining set of results are reported in the Technical Appendix in Fantazzini (2009): particularly, Figure 18 reports the Mean bias (in %), Median bias (in %) and Relative RMSE of the correlation parameters, for the KME–CML method, across different correlation levels and degrees of freedom as well as different data samples. The coverage rates for the 95% confidence intervals of the T-copula parameters based on a normal approximation are reported in Figure 20. Figures 19 and 21 in Fantazzini (2009) report the same simulation statistics but for the ML method. Finally, Figures 26 (second column)–27 (second column) in Fantazzini (2009) report the t-tests under the null hypothesis that the sample means across estimated correlations and degrees of freedom are equal to the true values, respectively.

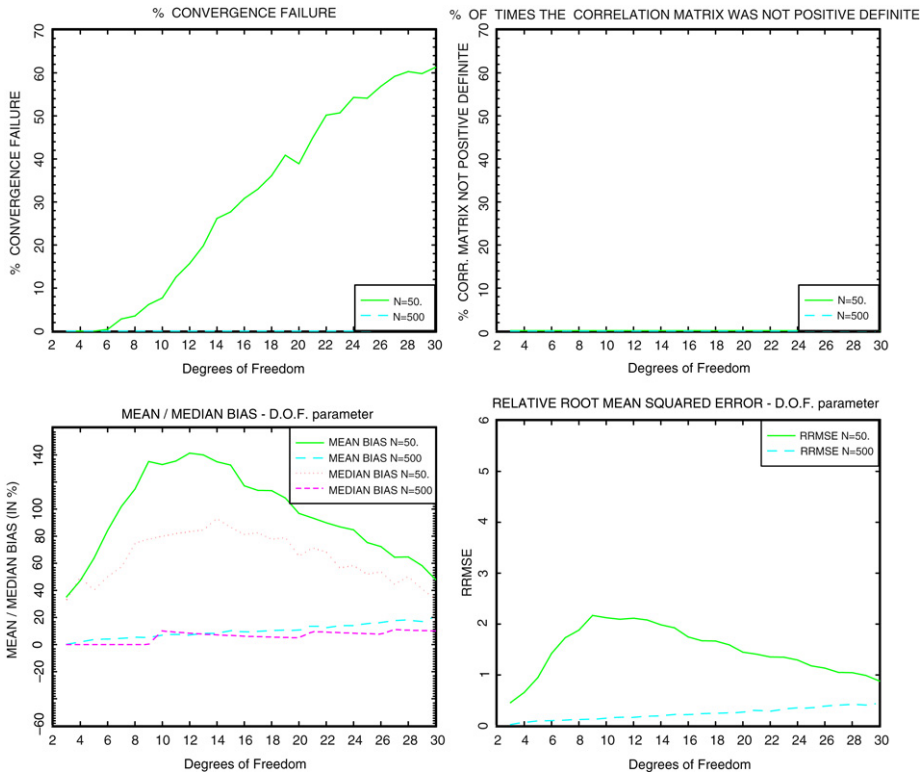


Fig. 9. The first plot reports the % of convergence failures when maximizing the log-likelihood. The second reports the % of times when the correlation matrix was not positive definite and the eigenvalue method was used. The third and the fourth plots report the Mean bias (in %), the Median bias (in %) and the Relative RMSE of the d.o.f. parameter, for the ML method for the 10-variate T-copula with correlation matrix reported in Table 2.

With regard to the ***degrees of freedom*** "ν", when $n = 50$, ν *is low* and the KME–CML method is employed, the analysis shows a strong positive mean bias and high RRMSE (see Fig. 4). However, the median bias is close to zero and the coverage rates are similar to the case with the ML method. This difference is due to the high % rates when the correlation matrix is not positive definite (between 10% and 20%) and the eigenvalue method by Rousseeuw and Molenberghs (1993) is used, see the second plot in Fig. 4. The analysis shows that this particular ad hoc fix has the effect to introduce a positive mean bias in $\hat{\nu}$, but the effect on the median is rather limited as well on the coverage rates. Besides, this bias quickly disappears when ν increases. Similar results can be retrieved by looking at the t-tests in Figure 27 in Fantazzini (2009) (second column).

Instead, when ν *is high* and the KME–CML method is employed, the simulation results show a negative median bias that has the effect to decrease the coverage rates below the 95% level. However, we remark that the drop in the coverage rates is much lower than those observed with bivariate T-copulas.

The ML method shows low positive mean and median biases as well as low RRMSE when ν is low. Then, these biases increase when ν increases, reaching the maximum around $\nu = 15$, after which they finally decrease (see Fig. 5). The effect on the 95% coverage rate is just specular: it is high when ν is low, then it decreases and finally it converges to the true value when ν is higher than 20.

When a larger sample with $n = 500$ observations is considered, the two methods perform rather well and the properties of the two estimators are quite similar. Interestingly, when $n = 500$ the correlation matrix is always positive definite for all ν, and the eigenvalue method by Rousseeuw and Molenberghs (1993) is not needed.

As for the ***correlation matrix***, when $n = 50$, both the ML and the KME–CML methods show mean and median biases as well as high RRMSE when the correlations are close to zero, see Figures 18 and 19 in Fantazzini (2009). However, the effects on the coverage rates are rather limited (see upper plots in Figures 20 and 21 in Fantazzini (2009)). Besides, the KME–CML method shows slightly better results than the ML method.

When $n = 500$ the previous biases decrease but they still remain quite high for the ML method when ν is low and the variables are close to be uncorrelated.

With regard to ***ANOVA tests***, when the correlation parameters are of concern, the Welch tests in Fig. 6 show again that the sample estimates delivered by the two methods (KME–CML and ML) have means which are not statistically significant different. A similar evidence is delivered by the Levene tests for the variances which are statistically different only when the true degrees of freedom ν are low.

As for the degrees of freedom parameter, the Welch and Levene tests in Fig. 7 point out that the two methods deliver statistically different estimates.

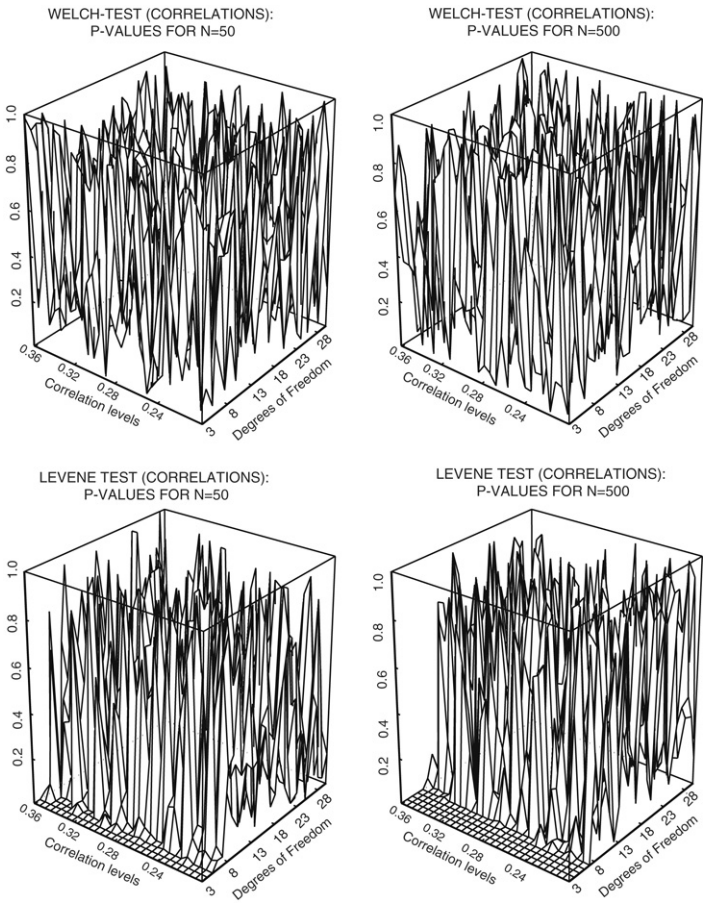


Fig. 10. The four plots report the p-values for the Welch tests (first row) and Levene tests (second row). The two samples compared in the tests are the *correlation parameters* of the T-copula with correlation matrix reported in Table 2, estimated with the ML and KME–CML methods.

Regarding the ***computational aspects***, similarly to the bivariate case, the ML method shows higher convergence failures than the KME–CML method when $n = 50$ and ν is higher than 10, while when $n = 500$ the numerical performances of the two methods are equal and there are no convergence failures. Furthermore, as previously discussed, when $n = 50$, ν is low and the KME–CML method is employed, the correlation matrix is not positive definite in 20% of cases and the eigenvalue method by Rousseeuw and Molenberghs (1993) has to be used. This fix induces a positive bias in the estimate of ν, but the effects on the coverage rates are rather limited. Besides, the number of times when this method has to be used quickly diminish when ν increases. Interestingly, the correlation matrix is always positive definite already with $n = 500$.

4.3. Simulation results: 10-variate T-copula — Dow Jones returns

For the sake of interest and space, we report in Figs. 8 and 9 the % of convergence failures when maximizing the log-likelihood, the % of times when the correlation matrix was not positive definite and the eigenvalue method was used, as well as the Mean bias (in %), the Median bias (in %) and the Relative RMSE of the d.o.f. parameter, for both the KME–CML and ML methods. Figs. 10 and 11 report the p-values for the Welch tests and Levene tests.

The remaining set of results are reported in the Technical Appendix in Fantazzini (2009): particularly, Figure 22 there reports the Mean bias (in %), Median bias (in %) and Relative RMSE of the correlation parameters, for the KME–CML method, across different correlation levels and degrees of freedom as well as different data samples. The coverage rate for the 95% confidence intervals of the T-copula parameters based on a normal approximation are reported in Figure 24 in Fantazzini (2009). Figures 23 and 25 report the same simulation statistics but for the ML method. Finally, Figures 26 (third column)–27 (third column) in Fantazzini (2009) report the t-tests under the null hypothesis that the sample means across estimated correlations and degrees of freedom are equal to the true values, respectively.

With regard to the ***degrees of freedom*** "ν", when $n = 50$, both the KME–CML and the ML methods show an inverted U-shape for the positive mean and median biases as well as for the RRMSE, as reported in Figs. 8 and 9. However, the biases for the KME–CML method are much lower than the ML method (20%–120% vs 0%–30%, respectively). Furthermore, while the biases for the ML method are very close to those observed for the ill-specified T-copula reported in Fig. 5, the mean

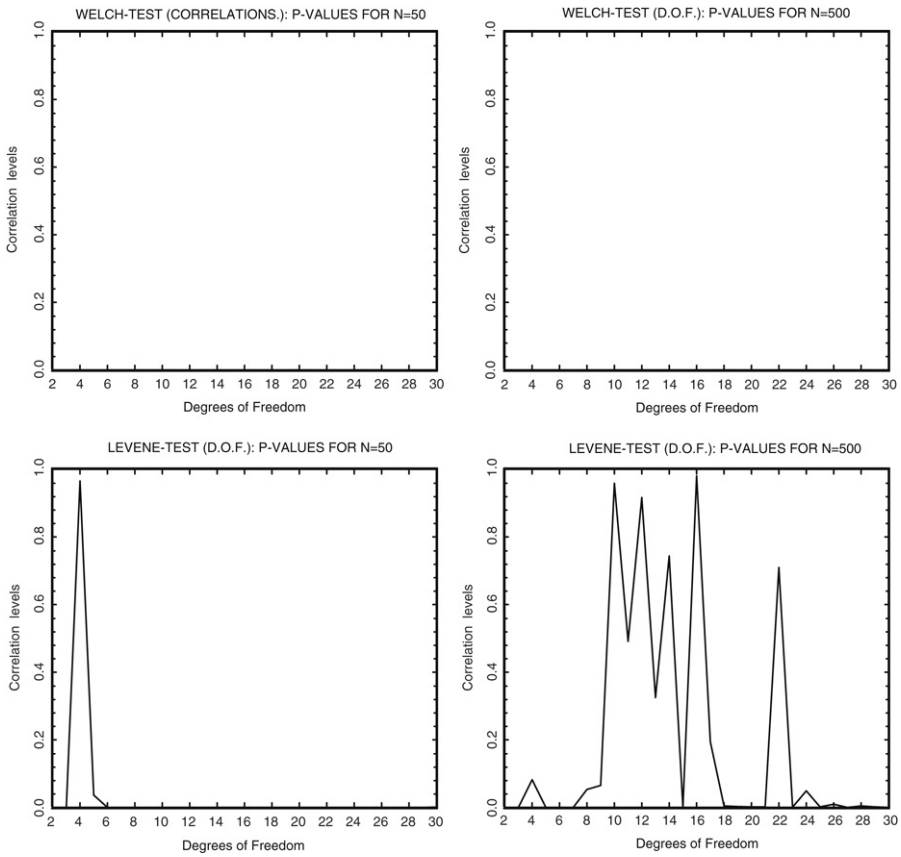


Fig. 11. The four plots report the p-values for the Welch tests (first row) and Levene tests (second row). The two samples compared in the tests are the *degrees of freedom parameters* of the T-copula with correlation matrix reported in Table 2, estimated with the ML and KME–CML methods.

biases observed for the KME–CML method are different from those observed for the same technique when dealing with the ill-specified T-copula, and reported in Fig. 4. This difference is due to the complete lack of negative definite correlation matrices (see Fig. 8), so that the eigenvalue method by Rousseeuw and Molenberghs (1993) is not needed.

Instead, both methods show median biases that are very close to those observed for the ill-specified T-copula, and similarly the 95% coverage rates reported in Figures 24 and 25 (lower plots) in Fantazzini (2009) are very close to those observed in Figures 20 and 21 (lower plots) in Fantazzini (2009), respectively: that is a U-shape that converges to the true rate when ν increases for the ML method, while a decreasing trend for the KME–CML method. However, we remark that the drop in the coverage rates for the KME–CML method is again much lower than that observed with bivariate T-copulas.

When a larger sample with $n = 500$ observations is considered, the two methods perform rather well and the properties of the two estimators are quite similar.

As for the ***correlation matrix***, the main results are close to those observed for the 10-variate T-copula with ill-specified correlation matrix: when $n = 50$, both the ML and the KME–CML methods show mean and median biases as well as high RRMSE when the correlations are close to zero, as shown in Figures 22 and 23 in Fantazzini (2009). However, the effects on the coverage rates are rather limited (see the upper plots in Figures 24 and 25 in Fantazzini (2009)). Besides, the KME–CML method shows slightly better results than the ML method. Differently from the ill-specified T-copula, the biases are much more lower, ranging between $-10/+10\%$ instead of $-100\%/+100\%$.

When $n = 500$ the previous biases decrease but they still remain quite high for the ML method when ν is low and the variables are close to be uncorrelated.

With regard to ***ANOVA tests***, when the correlation parameters are of concern, the Welch tests in Fig. 10 show again that the sample estimates delivered by the two methods (KME–CML and ML) have means which are not statistically significant different. A similar evidence is delivered by the Levene tests for the variances which are statistically different only when the true degrees of freedom ν are low.

As for the degrees of freedom parameter, the Welch and Levene tests in Fig. 11 point out that the two methods deliver statistically different estimates. However, when the dimension n increases, the Levene tests highlight that the variances of the sample estimates delivered by the two methods are no more statistically different.

Regarding the ***computational aspects***, similarly to the previous analysis, the ML method shows higher convergence failures than the the KME–CML method when $n = 50$ and ν is higher than 10, while when $n = 500$ the numerical

performances of the two methods are equal and there are no convergence failures. However, as previously discussed, the eigenvalue method by Rousseeuw and Molenberghs (1993) is not needed already with $n = 50$: this simulation evidence confirms previous empirical evidence in Demarta and McNeil (2005)and McNeil et al. (2005, chapter 5) who claim that the componentwise transformation of the empirical Kendall's tau matrix is positive definite in most cases.

Therefore, the previous results suggest to use the KME–CML method when dealing with small samples and low degrees of freedom, while the ML method is a better choice otherwise. A possible strategy would be to first use the KME–CML method: if the estimated degrees of freedom are higher than 20, then one should try to use the ML method if it converges. Otherwise, an alternative solution would be to use the simple normal copula, given that the T-copula tends to the Normal copula when $\nu \to \infty$, and the two copulas are already quite close when $\nu > 20$. Besides, the drop in the 95% coverage rates highlighted by our analysis for the KME–CML method when ν is high, is quite low if the number of variables is high: this is the usual case for financial professionals, whose managed portfolios are rarely bivariate, but include a large number of assets to diversify financial risk.

5. Conclusions

We developed the asymptotics of a recent semi-parametric estimation method used in the financial literature with the multivariate Student's T-copula, which involves empirical distribution functions, method-of-moments and maximum likelihood methods. We examined the finite-sample properties of this estimator via a Monte Carlo study designed to replicate different Data Generating Processes, and we found that this estimator was more efficient and less biased than the one-stage ML estimator when small samples and T-copulas with low degrees of freedom were of concern.

We then analyzed the pros and cons of this methodology in terms of numerical convergence and positive definiteness of the estimated T-copula correlation matrix. When small samples were of concern and ν was high, the number of times when the numerical maximization of the log-likelihood failed to converge was much higher for the ML method than for the KME–CML method. Yet, while the coverage rates at the 95% level for the ML estimates for ν did not show any particular bias or trend, the KME–CML estimates showed very low rates when ν became close to 30 and the correlations were not too strong. However, this drop in the coverage rates was large with bivariate T-copulas, only, while it was much lower when dealing with higher dimensional T-copulas, which is the usual case for real managed financial portfolios. Besides, both the ML and the KME–CML methods showed high mean and median biases for the estimated correlations when the true ones were close to zero. Nevertheless, the effects on the coverage rates for correlations were rather limited in this case.

Finally, we showed that the eigenvalue method by Rousseeuw and Molenberghs (1993) has to be used to obtain a positive definite correlation matrix only when dealing with very small samples ($n < 100$) and when the true underlying process has the lowest eigenvalue close to zero. This fix induces a positive bias in the estimate of ν, but the effects on the coverage rates are rather limited. Besides, the number of times when this method has to be used quickly decreases when ν increases.

Acknowledgements

For helpful comments and suggestions I would like to thank Paolo Giudici, Antonio Lijoi, Mario Maggi, Pietro Rigo, Eduardo Rossi. Special thanks go to the participants of the 63rd European Meeting of the Econometric Society, 27–31 August 2008, Milan, for helpful comments.

References

Barbe, P., Genest, C., Ghoudi, K., Remillard, B., 1996. On Kendalls process. Journal of Multivariate Analysis 58, 197–229.
Bouye', E., Durrleman, V., Nikeghbali, A., Riboulet, G., Roncalli, T., 2001. Copulas for finance a reading guide and some applications. Groupe de Recherche Operationnelle, Credit Lyonnais, Working Paper.
Cherubini, U., Luciano, E., Vecchiato, W., 2004. Copula Methods in Finance. Wiley.
Clayton, D.G., 1978. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65 (1), 141–151.
Clemen, M.N., Jouini, R.T., 1996. Copula models for aggregating expert opinions. Operations Research 44, 444–457.
Chen, X., Fan, Y., 2006. Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. Journal of Econometrics 135, 125–154.
Cochran, W.G., 1937. Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society 4 (Suppl. 1), 102–118.
Embrechts, P., McNeil, A., Straumann, D., 2002. Correlation and dependence in risk management: Properties and pitfalls. In: Dempster, M.A.H. (Ed.), Risk Management: Value at Risk and Beyond. Cambridge University Press, pp. 176–223.
Dalla Valle, L., Fantazzini, D., Giudici, P., 2008. Copulae and operational risks. International Journal of Risk Assessment and Management 9 (3), 238–257.
Demarta, S., McNeil, A.J., 2005. The t copula and related copulas. International Statistical Review 73, 111–129.
Dobric, J., Schmid, F., 2006. A goodness of fit test for copulas next term based on Rosenblatt's transformation. Computational Statistics and Data Analysis 51 (9), 4633–4642.
Fantazzini, D., 2009. Three-stage semi-parametric estimation of T-copulas: Asymptotics. Finite-Sample Properties and Computational Aspects. Moscow School of Economics Technical Report, Ferbuary 2009.
Fantazzini, D., 2009. The effects of misspecified marginals and copulas on computing the value at risk: A Monte Carlo study. Computational Statistics and Data Analysis, in press, doi:10.1016/j.csda.2008.02.002.
Genest, C., MacKay, R.J., 1986. Copules Archimédiennes et Familles de Lois Bidimensionnelles dont les Marges sont Donnes. Canadian Journal of Statistics 14, 145–159.
Genest, C., MacKay, R.J., 1986. The joy of copulas: Bivariate distributions with uniform marginals. The American Statistician 40, 280–285.
Genest, C., Rivest, L.-P., 1993. Statistical inference procedures for bivariate Archimedean copulas. Journal of the American Statistical Association 88, 1034–1043.

Genest, C., Ghoudi, K., Rivest, L.P., 1995. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82, 543–552.
Genest, C., Quessy, J.F., Remillard, B., 2006. Goodness-of-fit procedures for copula models based on the probability integral transformation. Scandinavian Journal of Statistics 33, 337–366.
Granger, C.W.J., Patton, A.J., Tërasvirta, T., 2006. Common factors in conditional distributions for bivariate time series. Journal of Econometrics 132, 43–57.
Greene, W., 2008. Econometric Analysis, 6th ed. Prentice Hall.
Höffding, D., 1940. Masstabinvariante Korrelationstheorie. Schriften des Mathematischen Seminars und des Instituts für Angewandte Mathematik der Universität 5, 181–233.
Huarda, D., Evin, G., Favre, A.C., 2006. Bayesian copula selection. Computational Statistics and Data Analysis 51 (2), 809–822.
Joe, H., 1997. Multivariate Models and Dependence Concepts. Chapman Hall.
Karlis, D., Nikoloulopoulos, A.K., 2008. Copula model evaluation based on parametric bootstrap. Computational Statistics and Data Analysis 52 (7), 3342–3353.
Kim, G., Silvapulle, M.J., Silvapulle, P., 2005. Semiparametric estimation of the joint distribution of two variables when each variable satisfies a regression model. Joint Statistical Meetings, August 2005, Minneapolis [Available as Working Paper Number 1/07, Department of Econometrics and Business Statistics, Monash University].
Kim, G., Silvapulle, M.J., Silvapulle, P., 2007. Comparison of semiparametric and parametric methods for estimating copulas. Computational Statistics and Data Analysis 51, 2836–2850.
Kim, G., Silvapulle, M.J., Silvapulle, P., 2008. Estimating the error distribution in multivariate heteroscedastic time series models. Journal of Statistical Planning and Inference 138 (5), 1442–1458.
Koul, H.L., Ling, S., 2006. Fitting an error distribution in some heteroscedastic time series model. Annals of Statistics 34, 994–1012.
Koul, H.L., 2002. Weighted Empirical Processes in Dynamic Nonlinear Models. In: Lecture Notes in Statistics, vol. 166. Springer Verlag.
Kowalczyka, T., Niewiadomska-Bugaj, M., 2002. An algorithm for maximizing Kendall's τ. Computational Statistics and Data Analysis 37 (2), 181–193.
Kruskal, W., 1958. Ordinal measures of association. Journal of the American Statistical Association 53, 814–861.
Levene, H., 1960. Robust tests for the equality of variances. In: Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G., Mann, H.B. (Eds.), Contribution to Probability and Statistics. Stanford University Press, Palo Alto, CA.
Lindskog, F., 2000. Modelling dependence with copulas. RiskLab Report, ETH Zurich.
Lindskog, F., McNeil, A.J., Schmock, U., 2002. Kendall's tau for elliptical distributions. In: Credit Risk Measurement Evaluation and Management. Physica-Verlag.
Marshall, R., Zeevi, A., 2002. Beyond correlation: Extreme co-movements between financial assets. Columbia University, working paper.
McNeil, A.J., Frey, R., Embrechts, P., 2005. Quantitative Risk Management. Princeton University Press.
Oakes, D., 1994. Multivariate survival distributions. Journal of Nonparametric Statistics 3 (3–4), 343–354.
Patton, A.J., 2004. On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. Journal of Financial Econometrics 2 (1), 130–168.
Patton, A.J., 2005. Estimation of copula models for time series of possibly different lengths. Journal of Applied Econometrics 21 (2), 147–173.
Nelsen, R.B., 1999. An Introduction to Copulas. In: Lecture Notes in Statistics, vol. 139. Springer.
Rosenberg, J.V., 1998. Pricing multivariate contingent claims using estimated risk-neutral density functions. Journal of International Money and Finance 17, 229–247.
Rosenberg, J.V., 2003. Nonparametric pricing of multivariate contingent claims. Journal of Derivatives 10 (3), 9–26.
Rousseeuw, P., Molenberghs, G., 1993. Transformation of non positive semidefinite correlation matrices. Communications in Statistics. Theory and Methods 22, 965–984.
Satterthwaite, F.E., 1946. An approximate distribution of estimates of variance components. Biometrics Bulletin 2 (6), 110–114.
Shih, J.H., Louis, T.A., 1995. Inferences on the association parameter in copula models for bivariate survival data. Biometrics 51, 1384–1399.
Sklar, A., 1959. Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de L'Université de Paris 8, 229–231.
Taniguchi, M., Kakizawa, Y., 2000. Asymptotic Theory of Statistical Inference for Time Series. Springer, NewYork.
Wang, W., Wells, M.A., 2000. Model selection and semiparametric inference for bivariate failure-time data. Journal of the American Statistical Association 95, 62–72.
Welch, B.L., 1951. On the comparison of several mean values: An alternative approach. Biometrika 38, 330–336.

Appendix 4 "Modelling and forecasting the global financial crisis: Initial findings using heterosckedastic log-periodic models."

Modelling and forecasting the global financial crisis: Initial findings using heterosckedastic log-periodic models. *Economics Bulletin*, 30(3), 1833-1841. http://www.accessecon.com/Pubs/EB/2010/Volume30/EB-10-V30-I3-P167.pdf

Permisson to post the Publisher's version/PDF: according to http://www.accessecon.com/Store/Economics%20Bulletin%20author%20guildlines-2012.pdf, "*authors are free to make any use of work they publish in the Economics Bulletin, be it commercial or non-commercia*"

Economics Bulletin

Volume 30, Issue 3

Modelling and forecasting the global financial crisis: Initial findings using heterosckedastic log-periodic models

Dean Fantazzini
Moscow School of Economics - Moscow State University

Abstract

The financial crisis of 2007-2009 has begun in July 2007 when a loss of confidence by investors in the value of securitized mortgages in the United States resulted in a liquidity crisis. World stock markets peaked in October 2007 and then entered a period of high volatility which culminated with the market crashes in September and October 2008. Since March 2009, the world stock markets have rebounded, but strong uncertainties still remain. In order to get more insights into the current world markets operation, we consider log-periodic models of price movements, which has been largely used in the past to forecast financial crashes and "anti-bubbles". Both the original and an extended model which accounts for heteroskedasticity and autocorrelation are fitted to the American S&P500 index. The empirical analysis reveal three interesting points: i) the log-periodic models outperform standard financial models when long-term out-of-sample forecasting is of concern. ii) the log-periodic-AR(1)- GARCH(1,1) model has residuals with better statistical properties than the original model and iii) the current market rebound should peak at the beginning of 2010.

The content in this paper is provided as general information only and should not be taken as investment advice. All paper content shall not be construed as a recommendation to buy or sell any security or financial instrument, or to participate in any particular trading or investment strategy. The ideas expressed in this paper are solely the opinions of the author and do not necessarily represent the opinions of sponsors or firms or public/private institutions affiliated with the author. Any action that you take as a result of information, or analysis, contained in this paper is ultimately your responsibility.

Citation: Dean Fantazzini, (2010) "Modelling and forecasting the global financial crisis: Initial findings using heterosckedastic log-periodic models", *Economics Bulletin*, Vol. 30 no.3 pp. 1833-1841.
Submitted: May 15 2009. **Published:** July 16, 2010.

1 Introduction

Sornette et al. (1996), Sornette and Johansen (1997), Johansen et al. (2000) and Sornette (2003) suggested that, prior to crashes, the mean function of a index price time series is characterized by a power law acceleration decorated with log-periodic oscillations, leading to a finite-time singularity that describes the onset of the market crash. Within this model, this behavior would hold for months and years in advance, allowing the anticipation of the crash from the log-periodic oscillations exhibited by the prices. The underlying hypothesis of this model is the existence of a growing cooperative action of the market traders due to an imitative behavior among them. In the pre-crash regime, clusters of correlated trades with arbitrary sizes would drive the financial system, and therefore the observed financial variables would exhibit scaling invariant properties. Particularly, Johansen et al. (2000) put forward the idea that stock market crashes are caused by the slow buildup of long-range correlations leading to a collapse of the stock market in one critical instant: in mathematical terms, complex dynamical systems can go through the so-called "critical" points, defined as the explosion to infinity of a normally well-behaved quantity. It is then natural to think of the stock market in terms of complex systems with analogies to dynamically driven out-of-equilibrium systems such as earthquakes, avalanches, crack propagation etc. However, there is a non-negligible and extremely important difference: the "microscopic" building blocks, i.e. the traders, are conscious of their action. This has been masterly captured by Keynes under the parable of the beauty contest. As a consequence, Johansen et al. (2000) develop an interesting model based on the interplay between economic theory and statistical physics. Since then, several authors have reported a large number of empirical results for a variety of unrelated crashes in worldwide stock markets indices. We refer to Sornette (2003) for a recent review of the theoretical framework of the log-periodic model and a compilation of empirical evidences. Johansen and Sornette (1999) and Zhou and Sornette (2005) examined the problem of whether the cooperative herding behavior of traders might also produce market evolutions that are symmetric to the accelerating speculative bubbles often ending in crashes. More specifically, they show that there seems to exist critical times t_c at which the market peaks with, either a power law increase with accelerating log-periodic oscillations ending with a crash (i.e. a bubble), or *a power law decrease after the critical time t_c with decelerating log-periodic oscillations*: in the latter case we have a so called "*anti-bubble*". Johansen and Sornette (1999) and Zhou and Sornette (2005) show that the traders' herding behavior can progressively occur and strengthen itself also in "bearish" decreasing market phases, thus forming anti-bubbles with decelerating market devaluations following market peaks.

What we do in this paper is to use log-periodic models to model and forecast the global financial crisis of 2007-2009 as a special case of "anti-bubble". By observing that world stock markets peaked in October 2007, which thus represents our critical time t_c, we present an econometric investigation of the log-periodic models, which looks at the residuals properties and tackles potential inferential problems arising from autocorrelation and heteroskedasticity. Particularly, we consider the AR(1)-GARCH(1,1) log-periodic model recently proposed by Gazola et al. (2008), which is intended to aggregate latent dynamical features and mechanisms of the normal phase of the market onto the critical long-range dynamics of price fluctuations encompassed by the original log-periodic model. We then compare the previous set of log-periodic models with standard times series models in terms of long-term out-of-sample forecasting performances. In this perspective, we perform forecasting exercises with the American SP500 stock index, considering 120-step ahead and 180-step ahead forecasts, showing that the AR(1)-GARCH(1,1) log-periodic model clearly outperforms standard financial models. Interestingly, we find out that the current rebound should peak at the beginning of 2010.

The rest of the paper is structured as follows. In Section 2 we review the log-periodic models for "anti-bubble" modelling, whereas in Section 3 we show an empirical application with the American S&P500 index. We perform extensive forecasting exercises in Section 4, while Section 5 briefly concludes.

2 Methodology

We consider the following log-periodic model for the price evolution trajectory of an anti-bubble:

$$p(t) = A + B(t - t_c)^{\beta} + C(t - t_c)^{\beta} \cos[w \ln(t - t_c) + \phi] + u_t \tag{1}$$

where $p(t)$ is the stock index price, $B < 0$ for a (bearish) anti-bubble to exist, $C \neq 0$ guarantees the significance of the log-periodic oscillations, β should be positive to ensure a finite price at the critical

initiation time t_c of the anti-bubble and quantifies the primary power law acceleration of prices, w is the angular log-frequency which measures the frequency of oscillations of the correction term in logarithmic time units, ϕ is a phase which can be absorbed in a re-definition of the unit of the time, while $t - t_c$ is the distance to the critical time t_c which is the starting point of the anti-bubble.
The previous deterministic component describes decreasing oscillations whose period increases as the time departs from the critical time t_c: as discussed in Sornette (2003 a,b) and Zhou and Sornette (2006), the power law acceleration $B(t - t_c)^\beta$ and the log-periodicity $\cos[w \ln(t - t_c) + \phi]$ are both intimately linked to the herding behavior of agents whose investments involve a competition between positive and negative feedbacks leading to a critical point. Furthermore, previous empirical studies as well as theoretical fundamentals discussed in Johansen et al. (2000) and Sornette (2003) suggest that the parameter governing the bubble growth should satisfy $0 < \beta < 1$, whereas the log-frequency parameter should satisfy $2 < w < 15$, meaning that there must be some oscillations embedded in the fit to give weight to the model. Particularly, for w close to zero, we would not observe any oscillations in the whole sample, so that the present modelling would not be very representative of the data under analysis. On the other hand, very high values for w would imply a lot of oscillations, thus resulting in an over-fitting of the noise. Besides, being ϕ a phase parameter, Gazola et al. (2008) proposed $0 < \phi < 2\pi$, which has also the advantage of ruling out some obvious non-identifiability problems. The time index t is converted in units of one year, so that 1 day = 1/365=0.0027397260 of the year.
Under normality assumption for the error term u_t, the maximum likelihood estimator for the parameter vector $\mathbf{\Pi} = [A, B, C, t_c, \beta, w, \phi]$ is obtained through the minimization of the squared residuals. Therefore, the estimates $\hat{\mathbf{\Pi}}$ for the log-periodic specification are estimated by nonlinear least squares (NLS).
In order to improve the optimization procedure, each parameter of the log-periodic model (1) denoted by θ and defined in a restricted interval denoted by $[a, b]$, can be re-parameterized according to the following monotonic transformation:

$$\theta = b\frac{\exp(\tilde{\theta})}{1 + \exp(\tilde{\theta})} + a\left(1 - \frac{\exp(\tilde{\theta})}{1 + \exp(\tilde{\theta})}\right) \tag{2}$$

This monotonic transformation turns the original estimation problem over a restricted space of solutions into an unrestricted problem, which eases estimation particularly when poor starting values are chosen. In this situation, the standard errors of the estimates can be computed by using the delta method. We remind that the delta method is used to compute an estimator for the variance of functions of estimators and the corresponding confidence bands. Let $\hat{V}[\tilde{\theta}]$ be an estimated variance-covariance matrix of $\tilde{\theta}$, then by the delta method, a variance-covariance matrix for a general nonlinear transformation $g(\tilde{\theta})$ is given by (see Greene (2002) or Hayashi (2000) for more details):

$$\hat{V}[g(\tilde{\theta})] = \frac{\partial g(\tilde{\theta})}{\partial \tilde{\theta}'}\hat{V}[\tilde{\theta}]\frac{\partial g(\tilde{\theta})}{\partial \tilde{\theta}}'$$

However, the delta method is only a first order approximation and may work very poorly in small samples like the ones used in our empirical analysis. A viable alternative is to maximize the log-likelihood with respect to $\tilde{\theta}$ in a first step, and then use the estimated transformed parameters $\hat{\theta}$ as starting values in a second maximization using only the original parameters θ.
While the previous log-periodic model (1) can model long-range dynamics of price fluctuations, nevertheless it is unable to consider the short-term market dynamics, thus showing residual terms $\hat{u}_t$ which are strongly autocorrelated and heteroskedastic. As a consequence, Gazola et al. (2008) proposed the following AR(1)-GARCH(1,1) log-periodic model:

$$\begin{aligned} p_t &= A + B(t - t_c)^\beta + C(t - t_c)^\beta \cos[w \ln(t - t_c) + \phi] + u_t \\ u_t &= \rho u_{t-1} + \eta_t \\ \eta_t &= \sigma_t \varepsilon_t, \quad \varepsilon_t \sim N(0, 1) \\ \sigma_t^2 &= \alpha_0 + \alpha_1 \eta_{t-1}^2 + \alpha_2 \sigma_{t-1}^2 \end{aligned} \tag{3}$$

where ε_t is a standard white noise term satisfying $\mathrm{E}[\varepsilon_t] = 0$ and $\mathrm{E}[\varepsilon_t^2] = 1$, whereas the conditional variance σ_t^2 follows a GARCH(1,1) process. We propose here a slightly different formulation for the

autoregressive structure in the conditional mean that we found much easier to estimate compared to (3):

$$\begin{aligned} p_t &= A + B(t-t_c)^\beta + C(t-t_c)^\beta \cos[w \ln(t-t_c) + \phi] + \rho p_{t-1} + \eta_t \\ \eta_t &= \sigma_t \varepsilon_t, \quad \varepsilon_t \sim N(0,1) \\ \sigma_t^2 &= \alpha_0 + \alpha_1 \eta_{t-1}^2 + \alpha_2 \sigma_{t-1}^2 \end{aligned} \tag{4}$$

The two formulations are very close and can be shown to differ only for the initial conditions. However, model (4) reached numerical convergence when applied to the times series of the SP500 in the years 2007-2009, while model (3) did not, thus confirming similar problems in Gazola et al. (2008). Moreover, the chosen lag orders in the mean and variance specifications in (4) minimized a set of information criteria: AIC, BIC and their modified versions proposed by Ng and Perron (2001). Besides, model (4) passed a large set of specification tests, as we will discuss in more details in Section 3.2 .
We remark that in case of processes with a deterministic linear trend and white noise shocks, the usual t and F-statistics have the same asymptotic distributions as those for the stationary processes, but with different convergence rates. Unfortunately, similar results for processes with deterministic non-linear trends (such as the ones exhibited by the log-periodic models) are not standard as in the linear case, and some care must be exercised when looking at the inferential results in the next section. However, an unreported preliminary small-scale simulation study seems to confirm that the t and F-statistics in case of nonlinear deterministic trends have the same asymptotic distributions as in the case of linear trends.

3 Empirical Analysis

We explore the possibility that log-periodic models provide a better empirical description than standard time series models of the long-run price dynamics of the American SP&500 market index, during the global financial crisis in the years 2007-2009. To answer this question, we estimate models (1) and (4) using data ranging between 10/10/2007 and 13/04/2009 for a total of 379 observations.

3.1 Estimation Results

When estimating the log-periodic models, we found out (not surprisingly) that the critical time t_c was very close to the market peak on the 09/10/2007, when the SP500 reached the value of 1565.15. Therefore, we fixed $t_c = 7.773$, i.e. October the 9th 2007, so to increase the degrees of freedom and improve the estimation efficiency. Besides, the numerical maximization of the log likelihood resulted in being much less cumbersome. The estimated parameters for the log-periodic model (1) and the AR(1)-GARCH(1,1) log-periodic model (4) are reported below in Tables I-II.

Table I: Estimation Results: Log-periodic model (1)

θ	$\hat{\theta}$	Std.err(*)	T-stat	P-value
A	1525.27	10.94	139.43	0.00
B	-388.71	16.50	-23.56	0.00
β	0.99	0.06	16.33	0.00
C	168.06	11.86	14.17	0.00
w	4.88	0.14	33.71	0.00
ϕ	2.11	0.08	25.34	0.00
Log lik.	-2022.90	Schwarz criterion		10.77

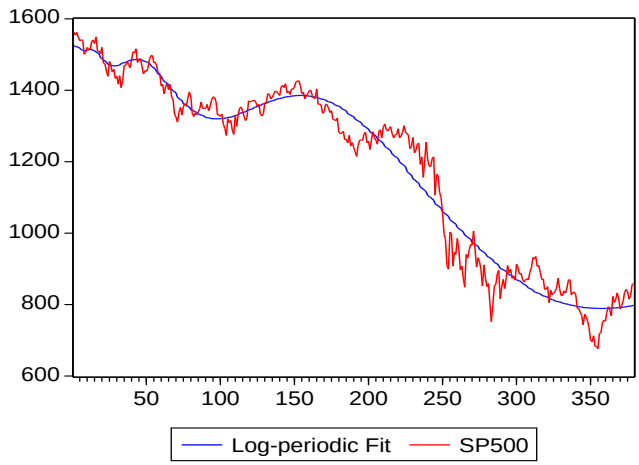
(*) Newey-West HAC Standard Errors

Table II: Estimation Results: AR(1)-GARCH(1,1) Log-periodic model (4)

θ	$\hat{\theta}$	Std.err	T-stat	P-value
A	184.78	35.14	5.26	0.00
B	-45.21	10.06	-4.49	0.00
β	0.96	0.15	6.60	0.00
C	21.20	4.18	5.08	0.00
w	4.52	0.30	15.29	0.00
ϕ	2.23	0.13	17.45	0.00
ρ	0.88	0.02	38.64	0.00
α_0	17.94	10.21	1.76	0.08
α_1	0.09	0.04	2.52	0.01
α_2	0.87	0.05	18.06	0.00
Log lik.	-1695.26	Schwarz criterion		9.13

Comparing Tables (I) and (II), it is possible to observe that all the log-periodic parameters are strongly significant and with the correct sign. Furthermore, the fundamental parameters β, w and ϕ are rather robust against the incorporation of the residual structure. On the other hand, the autocorrelation parameter ρ is strongly significant and smaller than 1 (we will see in the next section that the residuals are stationary), while the GARCH parameters take values well known in the financial literature, that is a small α_1 for the short-term shocks and a large α_2 for the long-term volatility persistence (see Tsay (2005) for a review of GARCH models). Figure 1 displays the plot of the SP500 stock index in the years 2007-2009 and the fitted curve according to the log-periodic specification given by model (1).

Figure 1: SP500 (10/10/2007 - 13/04/2009) and the log-periodic fit.



3.2 Specification Tests

The estimation of statistical models for non-stationary series like financial assets can be problematic, due to the possibility of a spurious regression. In this case, the residuals would be non-stationary and the parameter estimates would lack any statistical meaning. We test for unit roots in the residual series of the log-periodic specification (1) and (4) by using the Augmented Dickey-Fuller (ADF), the Dickey-Fuller test with GLS Detrending (DF-GLS) by Elliott et al. (1996) and the test by Kwiatkowski, Phillips, Schmidt and Shin (KPPS, 1992), which is based on the null of covariance stationarity rather than integratedness.

Table III: Tests statistics for ADF, DF-GLS and KPSS unit root tests applied to the residuals of the log-periodic models (1) and (4)

Residuals	**ADF**	**DF-GLS**	**KPSS**
Log-periodic model (1)	-4.823 (**)	-2.878 (**)	0.037
AR(1)-GARCH(1,1) Log-periodic model (4)	-21.236 (**)	-21.193 (**)	0.054

(**) Significant at the 1% level.

A careful analysis of the tests results reported in Table III shows that stationarity is the main feature of the residuals under scrutiny, similarly to the results found in Gazola et al. (2008). We then tested the goodness-of-fit of the log-periodic models employed for the conditional marginal distributions by using Ljung-Box tests on the residuals in levels and squares to test the null of no autocorrelation in the mean and in the variance, together with the specification tests discussed in Granger et al. (2006): we used standard empirical distribution tests for density specification (see D'Agostino and Stephens (1986)), together with the "Hit" test in order to test jointly for the adequacy of the dynamics and the density specification in the marginal distribution models, where the null hypothesis is that the density model is well specified. The latter test divides the support of the density into five regions and then applies interval forecast evaluation techniques to each region separately, and then to all regions jointly (see Granger et al. (2006) for more details). For the sake of space and interest, we report in Table IV only the p-values of each test.

The AR(1)-GARCH(1,1) log-periodic model (4) passes all the tests without any problem, thus highlighting that it is correctly specified, whereas the original log-periodic model (1) clearly shows some forms of

Table IV: P-values for the specification tests.

Model	**Ljung-Box (levels) (25)**	**Ljung-Box (squares) (25)**	**Jarque-Bera Test (H_0: Normal)**	**Lilliefors (D) (H_0: Normal)**
(1)	0.000	0.000	0.006	0.040
(4)	0.254	0.423	0.225	> 0.1
Model	**Cramer-von Mises (W2) (H_0: Normal)**	**Watson (U2) (H_0: Normal)**	**Anderson-Darling (A2) (H_0: Normal)**	**Joint Hit Test**
(1)	0.000	0.000	0.000	0.000
(4)	0.241	0.314	0.207	0.581

misspecifications. This result was expected, given that model (1) considers only the long-range dynamics of price movements, without taking the short-term dynamics into account.

4 Forecasting Exercises

In this section we perform some forecasting exercises with the American SP500 stock index, considering 120-step ahead and 180-step ahead forecasts in order to assess the accuracy of our approach with regard to a long-term forecasting. The competing models are the following ones:

1. The original log-periodic model (1);
2. The AR(1)-GARCH(1,1) log-periodic model (4);
3. An AR(p)-GARCH(1,1) model computed with log-returns;
4. An ARMA(1,1)-GARCH(1,1) model computed with log-returns.
5. A simple Random Walk.

The third and fourth models are standard models in the financial literature, while the random walk is the most classical benchmark in forecasting. We use the observations ranging between 10/10/2007 and 26/06/2008 as the first initialization sample, while those from 27/06/2008 till 13/04/2009 are used to perform 120-step ahead and 180-step ahead forecasts. A summary of the forecasting performances is reported in Table V.

Table V: Root Mean Squared Error and Mean Absolute Error for the 120 and 180-step ahead forecasts

	RMSE		**MAE**	
	120-step ahead	*180-step ahead*	*120-step ahead*	*180-step ahead*
LOG-PERIODIC	269.20	244.84	236.41	241.34
AR(1)-GARCH(1,1) -LOG-PERIODIC	250.48	170.17	223.50	158.22
AR(P)-GARCH(1,1)	266.16	220.95	244.81	214.03
ARMA(1,1)-GARCH(1,1)	266.95	227.05	244.78	221.52
RANDOM WALK	274.23	217.60	254.25	213.99

The previous table shows that the AR(1)-GARCH(1,1) log-periodic model yields better forecasting statistics than the competing models for both the 120-step ahead and the 180-step ahead forecasts, and the improvement increases with the forecasting distance. This result was expected since the main advantage of log-periodic models is to take the long range dynamics of price fluctuations into account, which is missing in standard financial models like the random walk. Instead, the original log-periodic model (1) shows mixed results, probably due to the misspecifications arising from not considering the short-term dynamics and which could have caused biased estimates and poor forecasts.
In order to compare the predictive accuracy of our models, we perform the Hansen and Lunde's (2005) and Hansen's (2005) Superior Predictive Ability (SPA) test, which compares the performances of two or more forecasting models. The forecasts are evaluated using a loss function like the MAE and the RMSE. The best forecasting model is the model that produces the smallest expected loss. The SPA test compares for the best standardized forecasting performance relative to a benchmark model, and the null hypothesis is that none of the competing models is better than the benchmark, see Hansen (2005) for

Table VI: Hansen's SPA test. P-values smaller than 0.05 are reported in bold fonts.

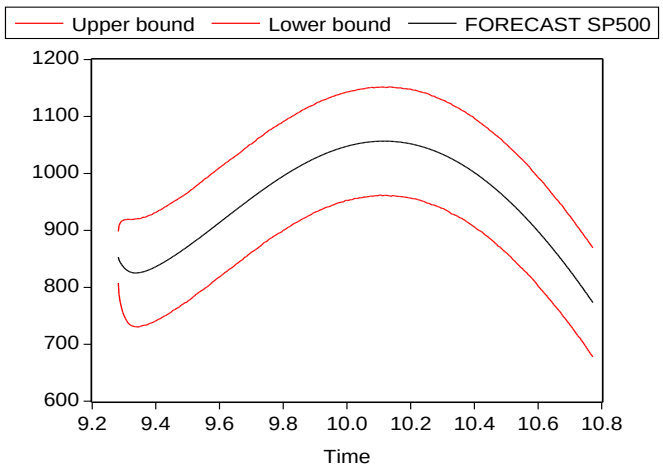
	RMSE		MAE	
Benchmark	*120-step ahead*	*180-step ahead*	*120-step ahead*	*180-step ahead*
LOG-PERIODIC	**0.01**	**0.00**	0.11	**0.00**
AR(1)-GARCH(1,1) -LOG-PERIODIC	0.52	0.51	0.61	0.51
AR(P)-GARCH(1,1)	0.31	**0.00**	0.34	**0.00**
ARMA(1,1)-GARCH(1,1)	0.27	**0.00**	0.32	**0.00**
RANDOM WALK	**0.00**	**0.00**	**0.00**	**0.00**

more details. The p-values produced by the SPA test with the RMSE and the MAE as loss functions are presented in Table VI.

The reported Hansen's SPA-consistent p-values show whether there is evidence against the hypothesis that the benchmark model is the best forecasting one. A low p-value (less that 0.05) means that the benchmark model is inferior to one or more of the competing models: the empirical results highlight that the sample being analyzed does yield strong evidence that all the competing models are outperformed by the AR(1)-GARCH(1,1) log-periodic model in case of the 180 step-ahead forecasts. As for the 120 step-ahead forecasts, the random walk and the original log-periodic model (1) can be outperformed by some other competing model, but this is not true for the AR(p)-GARCH(1,1), the ARMA(1,1)-GARCH(1,1) and the AR(1)-GARCH(1,1)-log-periodic models. This confirms our previous discussion with forecasting performances reported in Table V.

Finally, we report in Figure 2 the long-term out-of-sample forecast produced by the AR(1)-GARCH(1,1) log-periodic model from the 14/04/2009 till the 09/10/2010, together with 95% bootstrap confidence bands.

Figure 2: SP500 forecast: 14/04/2009 - 09/10/2010. Time t converted in units of one year.



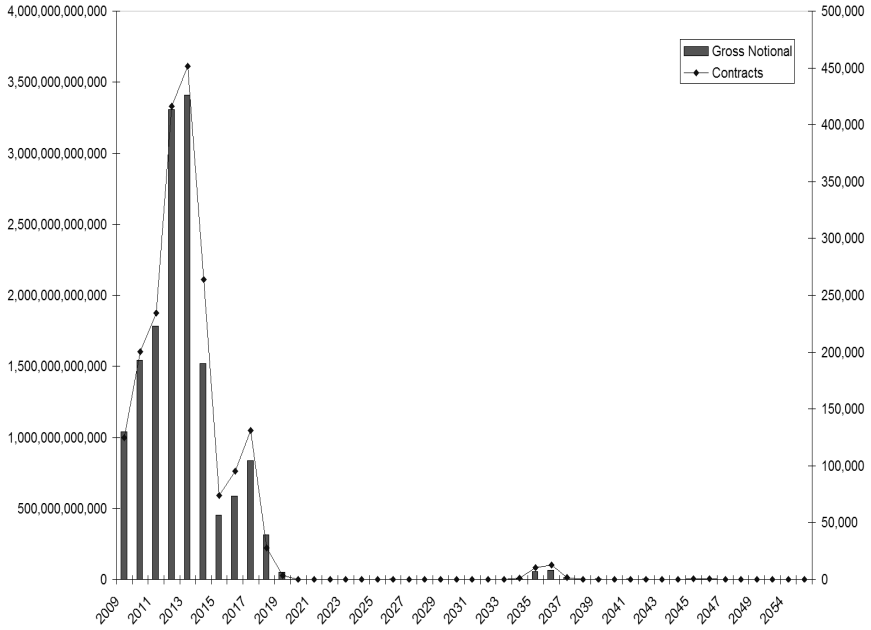
We considered here a forecast reaching a maximum distance of 3 years from the critical point t_c, i.e. the market peak on the 09/10/2007. We do so because the log-periodic models considered here are first order approximations valid only in the neighborhood of the critical point. Sornette and Johansen (1997) proposed a more general formula with additional degrees of freedom to better capture behavior away from the critical point up to 10 years. Moreover, Zhou and Sornette (2005) highlighted that, as time flows, the cumulative effect of exogenous news may detune progressively the anti-bubble pattern. This phenomenon may be accelerated in the presence of strong exogenous shocks such as the Federal Reserve interest rate and monetary policies.

The previous forecast of a renewed future market downturn in 2010 seems to be indirectly confirmed by the data on credit default swaps (CDS) reported online by the Depository Trust & Clearing Corporation's (DTCC's) Trade Information Warehouse[1]. These figures include the aggregate CDS gross notional

[1]DTCC is the world's largest post-trade financial services company and in 2007 it settled the vast majority of securities transactions in the United States, more than $ 1.86 quadrillion in value.

positions for Warehouse records as well as the aggregate number of contracts (see Figure 3 below).

Figure 3: CDS Gross notional positions and number of contracts (data retrieved on the 27/03/2009)



These data seem to suggest that the peak of the current crisis will be reached in the years 2012-2013, after which the situation should start normalizing. It will be very interesting to see whether this rather simple but extremely powerful forecasting tool will be confirmed by reality.

5 Conclusions

This paper proposed the use of log-periodic models to model and forecast the global financial crisis of 2007-2009 as a special case of "anti-bubble". By observing that world stock markets peaked in October 2007, we presented an econometric investigation of the log-periodic models, which tackled potential inferential problems arising from autocorrelation and heteroskedasticity. We considered the AR(1)-GARCH(1,1) log-periodic model recently proposed by Gazola et al. (2008), which aggregates latent dynamical features and mechanisms of the normal phase of the market, with the critical long-range dynamics of price fluctuations encompassed by the original log-periodic model. We then compared the previous set of log-periodic models with standard times series models in terms of long-term out-of-sample forecasting performances. We performed forecasting exercises with the American SP500 stock index, considering 120-step ahead and 180-step ahead forecasts. We showed that the AR(1)-GARCH(1,1) log-periodic model yielded better forecasting statistics than the competing models for both the 120-step ahead and the 180-step ahead forecasts, and the improvements increased with the forecasting distance. Instead, the original log-periodic model (1) showed mixed results, probably due to the misspecifications resulting from not considering the short-term dynamics: such misspecifications may have determined biased parameter estimates and poor forecasts. Interestingly, we found out that the current market rebound should peak at the beginning of 2010.

The main implication that we draw here is that the traders' herding behavior during the "Second Great Contraction" (using the definition by Reinhart and Rogoff (2009)) has progressively strengthened itself in bearish market phases, thus forming an anti-bubble. It remains to be seen whether the latter will last only 3 years, similarly to what happened with the SP500 in the years 2000-2003 and discussed by Zhou and Sornette (2005), or much longer like the Japanese anti-bubble in the 90's which was discussed in Johansen and Sornette (1999).

The future work will be directed to studying the small sample properties and the computational aspects of the ML estimators in case of models with nonlinear trends, similarly to the Monte Carlo studies recently performed in Fantazzini (2009, 2010).

References

[1] D'Agostino, R.B., and M. A., Stephens (eds.) (1986) *Goodness-Of-Fit Techniques*, Marcel Dekker, Inc.: New York.

[2] Elliott, G., Rothenberg, T., and J., Stock (1996) "Efficient tests for an autoregressive unit root", *Econometrica*, **64**, 813-836.

[3] Fantazzini, D. (2009) "The Effects of Misspecified Marginals and Copulas on Computing the Value at Risk: A Monte Carlo Study", *Computational Statistics and Data Analysis*, **53(6)**, 2168-2188.

[4] Fantazzini, D. (2010) "Three-Stage Semi-parametric Estimation of T-Copulas: Asymptotics, Finite-Sample Properties and Computational Aspects", *Computational Statistics and Data Analysis*, forthcoming.

[5] Gazola,L., Fernandes, C., Pizzinga, A., and R. Riera (2008) "The log-periodic-AR(1)-GARCH(1,1) model for financial crashes", *The European Physical Journal B*, **61**, 355-362.

[6] Granger, C.W.J., Patton, A.J. and T., Tërasvirta (2006) "Common factors in conditional distributions for bivariate time series", *Journal of Econometrics*, **132**, 43-57.

[7] Greene, W. (2002) *Econometric Analysis*, Princeton University Press, 5th edition.

[8] Hansen, P. (2005) "A Test for Superior Predictive Ability", *Journal of Business and Economic Statistics*, **23(4)**, 365-380.

[9] Hansen, P., and A., Lunde (2005) "A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?", *Journal of Applied Econometrics*, **20**, 873-889.

[20] Hayashi, F. (2000) *Econometrics*. Princeton University Press.

[11] Kwiatkowski, D., Phillips, P., Schmidt, P., and Y., Shin (1992). "Testing the null hypothesis of stationary against the alternative of a unit root", *Journal of Econometrics*, **54**, 159-178.

[12] Johansen, A., and D., Sornette (1999) "Financial "antibubbles": Log-periodicity in Gold and Nikkei collapses", *International Journal of Modern Physics C*, **10(4)**, 563-575.

[13] Johansen, A., Ledoit, O. and D., Sornette (2000) "Crashes as critical points", *International Journal of Theoretical and Applied Finance*, **3(2)**, 219-255.

[15] Ng, S. and P. Perron (2001) "Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power", *Econometrica*, **69(6)**, 1519-1554.

[15] Reinhart, C. and K. Rogoff (2009) *This Time is Different: Eight Centuries of Financial Folly*, Princeton University Press: Princeton, NJ.

[16] Sornette, D., Johansen, A. and J.P., Bouchaud (1996) "Stock market crashes, Precursors and Replicas", *Journal de Physique I*, **6**, 167-175.

[17] Sornette, D. and A., Johansen (1997) "Large financial crashes", Physica A, **245**, 411-422.

[18] Sornette, D. (2003a) *Why Stock Markets Crash (Critical Events in Complex Financial Systems)*, Princeton University Press: Princeton, NJ.

[19] Sornette, D. (2003b) "Critical market crashes", *Physics Reports*, **378(1)**, 1-98.

[20] Tsay, R. (2005) *Analysis of Financial Time Series*. Wiley.

[21] Zhou, W.X., and D., Sornette (2005) "Testing the Stability of the 2000-2003 US Stock Market "Antibubble"", *Physica A*, **348**, 428-452.

[22] Zhou, W.X., and D., Sornette (2006) "Predictability of Large Future Changes in Major Financial Indices", *International Journal of Forecasting*, **22**, 153-168.

Appendix 5 "Forecasting the global financial crisis in the years 2009-2010: Ex-post analysis."

Forecasting the global financial crisis in the years 2009-2010: Ex-post analysis. *Economics Bulletin*, 31(4), 3259-3267. http://www.accessecon.com/Pubs/EB/2011/Volume31/EB-11-V31-I4-P295.pdf

Permisson to post the Publisher's version/PDF: according to http://www.accessecon.com/Store/Economics%20Bulletin%20author%20guildlines-2012.pdf, "*authors are free to make any use of work they publish in the Economics Bulletin, be it commercial or non-commercia*"

Economics Bulletin

Volume 31, Issue 4

Forecasting the Global Financial Crisis in the Years 2009-2010: Ex-post Analysis

Dean Fantazzini
Moscow School of Economics - Moscow State University

Abstract

The ex-ante forecast of the SP500 index discussed in Fantazzini (2010a), covering the time sample 14/04/2009 - 09/10/2010, and originally submitted to the Economics Bulletin on the 15/05/2009 is analyzed. It is found that the realized values of the SP500 index trailed the forecasted values quite well, moving inside the forecast confidence bands for over a year. Interestingly, it is also found that the confidence bands worked very good as resistance levels, while the forecasted values as support levels. Moreover, an important turning point in April 2010 was also correctly forecasted. However, in July-August 2010, the SP500 started to diverge upwards, and after the speech by the FED Chairman at Jackson Hole (Wyo., USA) on the 27/08/2010, the stock market index never returned inside the forecast confidence bands. Additional evidence is then provided to show that the anti-bubble started in October 2007 ended almost three years later in August 2010, similarly to the first anti-bubble on the SP500, which started in August 2000 and ended in August 2003.

The author wants to thank Mario Maggi for stimulating comments on a previous version.
Citation: Dean Fantazzini, (2011) "Forecasting the Global Financial Crisis in the Years 2009-2010: Ex-post Analysis", *Economics Bulletin*, Vol. 31 No. 4 pp. 3259-3267.
Contact: Dean Fantazzini - fantazzini@mse-msu.ru.
Submitted: June 06, 2011. **Published:** November 28, 2011.

1 Introduction

Fantazzini (2010a) used log-periodic power law models to model and forecast the global financial crisis of 2007-2009 as a special case of "anti-bubble". He then compared a set of log-periodic models with standard times series models in terms of long-term out-of-sample forecasting performances using the American SP500 stock index, and found out that the AR(1)-GARCH(1,1) log-periodic model clearly outperformed standard financial models. Moreover, he reported in Figure 2 an *ex-ante* out-of-sample forecast produced by the AR(1)-GARCH(1,1) log-periodic model from the 14/04/2009 till the 09/10/2010, together with 95% bootstrap confidence bands. We remark that the paper was submitted to the Economics Bulletin on the 15/05/2009 (submission number: EB-09-00287).

What we do in this paper is to review that forecast by comparing it with the realized values of the SP500 index. We found that the SP500 trailed the forecasted values quite well, moving inside the forecast confidence bands for over a year. Interestingly, we also found that the confidence bands worked very good as resistance levels, while the forecasted values as support levels. Moreover, an important turning point on April the 23rd 2010, which worked as temporary market peak, was also correctly forecasted: in this regard, Fantazzini (2010a) expected that ... "the current market rebound should peak at the beginning of 2010". Besides, we also show how the market behavior before this turning point could have been modelled as a short-term bubble nested in a longer-term anti-bubble, and how the author diagnosed the end of this small bubble ex-ante by uploading on `www.scribd.com` a preliminary analysis about it on the 13th of March 2010. Furthermore, the importance of this temporary market peak and the following market fall is also highlighted by the Zivot-Andrews (1992) and Lee-Strazicich (2003) unit root tests allowing for possible structural break(s): using the SP500 daily data from 2009 till 2011, these tests reject the null hypothesis of a unit root in favor of a stationary model with a structural break at the beginning of May 2010, that is in the middle of the market fall following the (temporary) market peak on the 23/04/2010.

However, in July-August 2010, the SP500 started to diverge upwards, and after the speech by the FED Chairman at Jackson Hole (Wyo., USA) on the 27/08/2010, the stock market index never returned inside the forecast confidence bands, thus highlighting the end of the anti-bubble pattern. This conclusion is further confirmed by the recursive estimated parameters of the AR(1)-GARCH(1,1) log-periodic model, as well as by the test proposed by Zhou and Sornette (2005) to detect the possible end of this log-periodic structure. Therefore, we found that the anti-bubble started in October 2007 ended almost three years later in August 2010, similarly to the first anti-bubble on the SP500, which started in August 2000 and ended in August 2003. It may well be the case that the severity of the market downturn in the spring and summer 2010, together with the potential worsening of the US economy, may have prompted the US Federal Reserve to act decisively to stave off a potential renewed economic recession: this decision was then strongly signalled to the financial markets by the FED Chairman's speech at Jackson Hole (Wyo., USA) on the 27/08/2010, thus paving the way to an additional round of Quantitative Easing. See the full transcript of the Chairman's speech at http://www.federalreserve.gov/newsevents/speech/bernanke20100827a.htm for more details.

The rest of the paper is structured as follows. In Section 2 we provide a literature review regarding bubbles and anti-bubbles, whereas in section 3 we compare the ex-ante forecast discussed in Fantazzini (2010a) with the realized values of the SP500 index. In Section 4 we provide an additional ex-post analysis about the temporary market peak in April 2010 and the short-term bubble that preceded it, while in Section 5 we test the end of the anti-bubble with the approach proposed by Zhou and Sornette (2005). Section 6 briefly concludes.

2 Literature Review

2.1 Bubble Modelling: The Log Periodic Power Law approach

Johansen, Ledoit and Sornette (2000) proposed a model that assumes the presence of two types of agents in the market: a group of traders with rational expectations and a second group of so called "noise" traders, that is irrational agents with herding behavior. These traders are organized into networks and can have only two states, that is buy or sell. Moreover, their trading actions depend on the decisions of other traders and on external influences, so that agents can form groups with self-similar behavior which can lead the market to a bubble situation, which can be considered a situation of "order", compared to

the "disorder" of normal market conditions. Furthermore, this model allows for positive feedbacks which are generated by the increasing risk and the agents interactions, so that a bubble can be a self-sustained process. Interestingly, this approach allowed to diagnose bubbles ex-ante, as shown in a series of real-life tests, see Sornette and Zhou (2006), Sornette, Woodard and Zhou (2008) and Zhou and Sornette (2003, 2006, 2008, 2009). See Jiang et al. (2010), Geraskin and Fantazzini (2011) and Fantazzini (2010c) for a recent review of these type of models and for some extensions with applications.
The fundamental equation that describes the temporal growth of (log) prices before a crash or a simple bubble deflation is given by the following Log Periodic Power Law (LPPL),

$$\ln[p(t)] \approx A + B(t_c - t)^{\beta} + C(t_c - t)^{\beta} \cos[\omega \ln(t_c - t) + \phi] \tag{1}$$

where $t < t_c$ is any time before the end of the bubble, $A > 0$ is the value of $[\ln p(t_c)]$ at the critical time, $B < 0$ the increase in $[\ln p(t)]$ over the time unit before the crash if C were to be close to zero, $C \neq 0$ is the proportional magnitude of the oscillations around the exponential growth, $0 < \beta < 1$ should be positive to ensure a finite price at the critical time t_c of the bubble and quantifies the power law acceleration of prices, ω is the frequency of the oscillations during the bubble, while $0 < \phi < 2\pi$ is a phase parameter.
Even though eq. (1) can be estimated by Nonlinear Least Squares, Fantazzini (2010b) resorted to the Taboo Search (TS) algorithm by Cvijovic and Klinowski (1995), as well as the Pure Random Search (PRS), with 1 million draws. These algorithms do not need the computation of gradients and hessians, and are usually preferred for highly nonlinear functions. See Geraskin and Fantazzini (2011) for a recent discussion of different estimation methods for the LPPL approach.

2.2 Anti-Bubbles

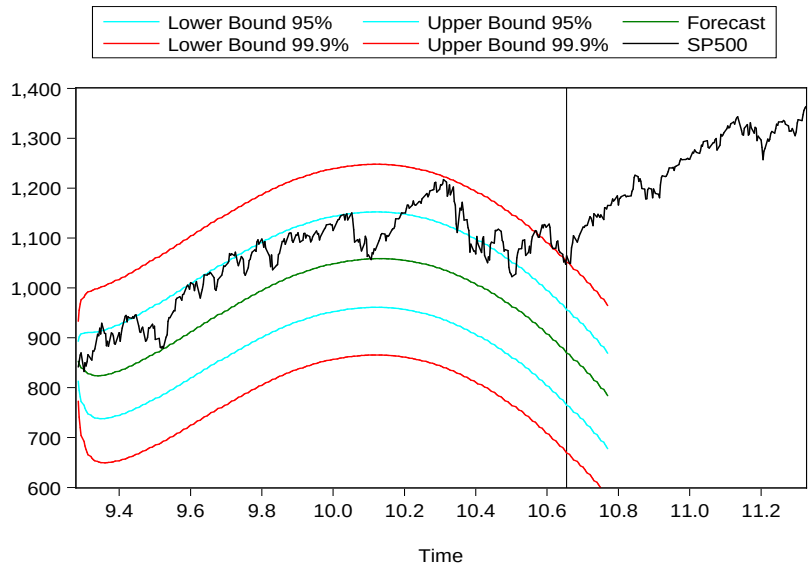
An "anti-bubble" is symmetric to a bubble, and represents a situation when the market peaks at a critical time t_c, after which it follows a power law decrease with decelerating log-periodic oscillations. Johansen and Sornette (1999) and Zhou and Sornette (2005) explained this phenomenon by showing how traders' herding behavior can progressively occur and strengthen itself in bearish decreasing market phases, thus forming anti-bubbles with decelerating market devaluations following market peaks.
As discussed in Zhou and Sornette (2005) and Fantazzini (2010a), as time flows, the cumulative effect of exogenous news may detune progressively the anti-bubble pattern, and this phenomenon may be accelerated in the presence of strong exogenous shocks such as the Federal Reserve interest rate and monetary policies.

3 Empirical Analysis: Forecast vs Real Values in 2009-2010

We report in Figure 1 the original ex-ante forecast of the SP500 reported as Figure 2 in Fantazzini (2010a, p.6) and covering the time sample between 14/04/2009 and 09/10/2010, the 95% and 99.9% forecast confidence bands over the same period, together with the realized values of the SP500 between 14/04/2009 and 29/04/2011 (i.e. the last data available at the time of writing this paper), as well as a vertical line highlighting the day of the FED Chairman's speech at Jackson Hole (Wyo., USA):
The previous figure shows that the SP500 trailed the ex-ante forecast quite well, moving inside the 95% forecast confidence bands till the end of March 2011, and inside the 99.9% forecast confidence bands till the end of August 2011. Interestingly, we can observe that the confidence bands worked very good as resistance levels, while the forecasted values as support levels. Moreover, Figure 1 also highlights that the FED Chairman's speech at Jackson Hole (Wyo., USA) had a significant impact on the financial markets: without the assurance of a second round of Quantitative Easing which was given that day, the SP500 would have most likely kept going down below the critical 1000 level and beyond. Needless to say, such an outcome would have paved the way to a renewed economic recession.
In this regard, it is important to remark that Bernanke himself openly confirmed on the 13/01/2011 that the FED's actions led to the stock market rally that began at the end of August 2010[1].

[1]During a forum sponsored by the Federal Deposit Insurance Corp., the government agency that backs bank deposits, the FED Chairman said that "*...the stock market rally that began last summer was fueled by the FED's efforts which improved U.S. economic activity...*", and "*I do think that our policies have contributed to a stronger stock market, just as they did in March of 2009*". See the full article by Chris Isidore at CNN-Money at http://money.cnn.com/2011/01/13/news/economy/bernanke_qe2_stock_market/index.htm for more details.

Figure 1: *SP500 ex-ante forecast, SP500 realized values, 95% and 99.9% confidence bands over the time sample* ***14/04/2009 - 29/04/2011****. The vertical black line on the 27/08/2010 signals the day of the Chairman's speech at Jackson Hole (Wyo., USA). Time* **t** *converted in units of one year (0 is set at Jan. 1st 2000).*



Clearly, discussing the advantages and disadvantages of Quantitative Easing is out of the scope of this paper, and we leave it as avenue for further research.

4 Empirical Analysis: The Turning Point in April 2010

The temporary end in April 2010 of the market upward trend which started in March 2009, may be one of the reasons behind the FED decision to proceed with QE2. In this regard, the AR(1)-GARCH(1,1) log-periodic model used in Fantazzini (2010a) was clearly expecting such an event, as shown by the previous Figure 1.

However, additional evidence can be provided about the criticality of this turning point: for example, the market behavior from February 2010 till April 2010 could have been modelled as a short-term bubble nested in a longer-term anti-bubble (see section 4.1 below). Moreover, the importance of this temporary market peak and the following market fall is also highlighted by unit root tests allowing for possible structural breaks, which show that the market suffered a significant structural break at the beginning of May 2010 (see section 4.2 below).

4.1 A Short-Term Bubble Nested in a Longer-Term Anti-Bubble?

On the 13/03/2010, Fantazzini (2010b) uploaded a preliminary analysis on `www.scribd.com`, about a possible small bubble developing in the future E-mini SP500 using hourly data from February 2010 till March 2010. More specifically, he employed the Log Periodic Power Law (LPPL) approach by Johansen, Ledoit and Sornette (2000) reviewed in Section 2.1 to verify the presence of a possible bubble.

Fantazzini (2010b) fitted the LPPL formula (1) to the hourly data of the E-mini SP500, using a rolling estimation window of 250 data to be robust against parameter variation, considering both the TS and the PRS algorithms. Given the stochastic nature of the initial parameter selection with the TS and PRS algorithms and the noisy nature of the underlying generating processes, after saving the parameters that minimized the cost function, he re-estimated the model a large number of times in order to build an empirical distribution for the estimated parameters. The estimation results were then filtered by the

following LPPL conditions, which were also used in Jiang et al. (2010) and Geraskin and Fantazzini (2011): $t_c > t_N$ (where t_N is the last observation in the estimation sample), $B < 0$ and $0 < \beta < 1$.
The main empirical quantiles for the critical time t_c when the bubble should have ended (at least temporarily), either by a market crash or by a simple bubble deflation, and for $\exp(A) \approx p(t_c)$, that is the price at time t_c, are given below:

Table 1: Quantiles of the empirical distribution of the estimated parameters t_c and $\exp(A)$

Quantile	t_c (converted in units of 1 year)	t_c (date)	$\exp(A) \approx p(t_c)$
2.5%	10.167	02.03.2010	1113
5.0%	10.169	03.03.2010	1115
25.0%	10.174	05.03.2010	1127
50.0%	10.179	11.03.2010	1145
75.0%	10.209	17.03.2010	1198
95.0%	10.266	08.04.2010	1632
97.5%	10.291	19.04.2010	1880

The results are quite encouraging, also considering that the SP500 reached a first peak on the 15/04/2010 at 1211.67, quite close to the peak reached on the 23/04/2010 at 1217.28. The strongly skewed distributions of the parameters were expected, see Jiang et al. (2010) and Geraskin and Fantazzini (2011) for similar results.

4.2 The End of the Short-Term Bubble: Was There a Structural Break in the SP500?

The previous subsection highlighted that an end to the short-term bubble in the SP500 was expected around March-April 2010. Therefore, the temporary market peak observed in April 2010 could have been a matter of concern if it represented a structural break in the upward market trend started in March 2009. To verify such a possibility, we considered a battery of unit root tests allowing for structural breaks in the level and the slope of the time series:

- *Zivot and Andrews (ZA, 1992) unit root test with endogenous structural break*: this is a sequential test which utilizes the full sample and uses a different dummy variable for each possible break date. Three types of structural breaks are allowed: *i*) a break in the level (or intercept) of the series; *ii*) a break in the slope; *iii*) a break both in the level and in the slope of the series. Zivot and Andrews (1992) proposed a test which endogenously determines the most likely occurrence of a structural break: more specifically, the break date is selected where the t-statistic from the (modified) ADF test regressions is the most negative, thus reaching a minimum. Due to space limits, we refer to Zivot and Andrews (1992) for more details and the test critical values, Pfaff (2008) for an implementation with `R` software, the routines by prof. J. Lee available at `http://www.cba.ua.edu/ jlee/gauss` for an implementation with `Gauss`, and the ZA add-in for Eviews software at `www.eviews.com`.

- *Lee and Strazicich (LS, 2003) minimum Lagrange Multiplier (LM) unit root test*: Lee and Strazicich (2003) highlighted that the previous test (and others similar to it) is derived assuming no break(s) under the null hypothesis, which may lead to a potential spurious rejection of the null hypothesis when there is a unit root with breaks: in such a situation, one may erroneously conclude that the time series is trend stationary when it is non-stationary with break(s). Moreover, Lumsdaine and Papell (1997) and Maddala and Kim (2003) argued that only one endogenous break may not be enough and lead to a loss of information when actually more than one break exists. In order to deal with these issues, Lee and Strazicich (2003) developed a two-break minimum LM unit root test, which allows for endogenous break(s) both under the null and the alternative hypothesis: therefore, this test avoid both the spurious rejection and the trend misspecification problems, and the rejection of the null hypothesis unambiguously implies trend-stationarity. Due to space limits we refer to Lee and Strazicich (2003) for the full description of the test implementation.

Since we want to check whether there was a structural break in the upward market trend, which could have been of great concern for policy makers, we consider Model *iii*) for both tests (i.e. possible breaks in the levels and the slope). The results using the time sample 14/04/2009 - 29/04/2011 are reported in Table 2.
The two tests find indeed a structural break in May 2010, with significant negative dummies after the trend break and the date for the break ranging between May 4 till May 19: this period of time was

Table 2: Zivot-Andrews Test and Lee and Strazicich Minimum LM Two Breaks Unit Root Tests

	Zivot-Andrews Test			Lee and Strazicich - LM Two Breaks					
	Test	Break	LM Test	First Break			Second Break		
	Statistic	Date	Statistic	Date	B_1	D_1	Date	B_2	D_2
SP500 (logs)	-5.18**	04/05/2010	-5.20*	01/09/2009	-0.01	0.00	19/05/2010	-0.395	-0.01
					(-0.84)	(1.67)		(-3.62)	(-3.68)
SP500 (levels)	-5.60**	04/05/2010	-5.00*	11/05/2010	24.46	-13.01	20/09/2010	-5.10	8.95
					(2.01)	(-4.82)		(-0.43)	(-4.24)

[The symbols * and ** denote statistical significance at the 10% and 5% levels. See Zivot and Andrews(1992) and Lee and Strazicich (2003) for the critical values of the test statistics. T-statistics for the dummies B_1,D_1,B_2 and D_2 are reported in parentheses. B_1 and B_2 are the dummies for the levels, while D_1 and D_2 for the slope.]

characterized by a strong market fall following the (temporary) market peak on the 23/04/2010. Once the structural breaks are included in the analysis, the null of a unit root is rejected by both tests. Moreover, if we consider the LM unit root test for the SP500 in levels, we can observe a second structural break in September 2010, that is after the FED Chairman's speech at Jackson Hole, with a positive dummy after the trend break. However, this structural break is not present if we consider the SP500 in logs and the LM test statistic is lower when considering the SP500 in levels, so that the evidence for this second break is weaker.

5 Testing the End of the Anti-Bubble: The Approach by Zhou and Sornette (2005)

The evidence reported in section 3 seems to point out that the anti-bubble pattern started in 2007 may have ended. However, in order to find more precise and conclusive evidence, it is necessary to use a statistical approach which is able to consider the specific nonlinear structure of the log-periodic anti-bubble pattern. Therefore, we decided to follow Zhou and Sornette (2005), which suggested to check the stability of the recursive estimates of the log-periodic model and proposed a test which belongs to a large class of pattern recognition methods.

5.1 Stability of the Anti-Bubble Equation Parameters

The stability of the parameters is an important condition to verify whether the anti-bubble pattern is a correct approach to model market prices. Therefore, similarly to Zhou and Sornette (2005), we fitted the SP500 with the AR(1)-GARCH(1,1) log-periodic model used in Fantazzini (2010a, eq. 4),

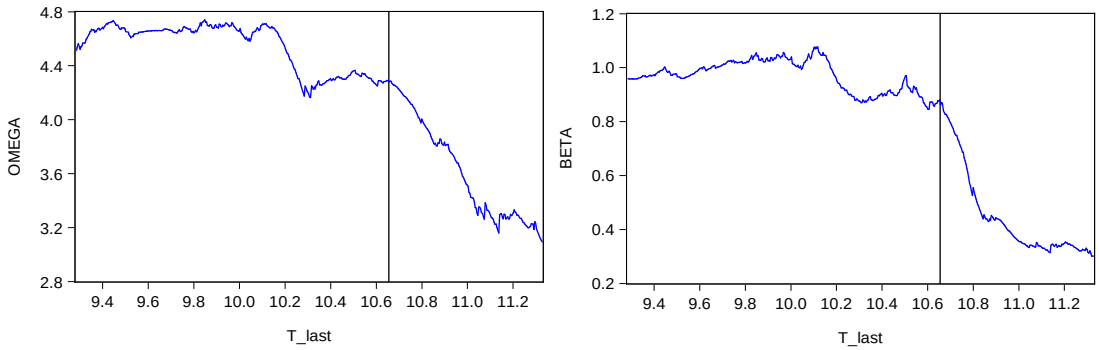
$$\begin{aligned} p_t &= A + B(t-t_c)^{\beta} + C(t-t_c)^{\beta}\cos[\omega \ln(t-t_c)+\phi] + \rho p_{t-1} + \eta_t \\ \eta_t &= \sigma_t \varepsilon_t, \quad \varepsilon_t \sim N(0,1) \\ \sigma_t^2 &= \alpha_0 + \alpha_1 \eta_{t-1}^2 + \alpha_2 \sigma_{t-1}^2 \end{aligned} \tag{2}$$

over a running window from t_c = 09/10/2007 (the date of the SP500 market peak) to t_{last}, where t_{last} ranges from 2009/04/14 (the first date of the ex-ante forecast reported in Fantazzini (2010a)) till 2011/04/29, that is the last data available. Due to space limits and for sake of interest, we report in Figure 2 only the recursive estimates of ω and β, which are the key parameters of the log-periodic model in (2), since A, B, C and ϕ are just units distributions of betas and omegas, see Sornette and Johansen (2001) and Geraskin and Fantazzini (2011) for more details. However, the results for the remaining parameters are qualitatively the same, and are available from the author upon request.
Figure 2 clearly shows that the market underwent a change of regime after the Chairman's speech at Jackson Hole (Wyo., USA) and the FED policies quickly detuned the anti-bubble pattern.

5.2 The Test by Zhou and Sornette (2005)

Zhou and Sornette (2005) proposed a test to verify the possibility that an antibubble may have ended, with a potential regime switching. Following Zhou and Sornette (2005), we considered the SP500 from the beginning of the antibubble t_c to the time t_{last} = 27/08/2010 and we then simulated $N = 1000$ price trajectories, from t_{last} up to six months into the future (i.e. up to 28/02/2011), for the following two classes:

Figure 2: *Recursive estimates of ω and β with the log-periodic equation (2), where t_{last} ranges from 2009/04/14 till 2011/04/29. The vertical black line on the 27/08/2010 signals the day of the FED Chairman's speech at Jackson Hole (Wyo., USA). Time* `t` *converted in units of one year.*



- Class I: Continuation of the log-periodic formula with noise added to it, where the noise follows a GARCH model with parameters given by the LPPL fitted values using data from t_c till t_{last};
- Class II: Random walk with daily volatility equal to the historical volatility over the same period (we also tried GARCH noise, but the results remain qualitatively the same).

Class I represents the continuation of the anti-bubble pattern, whereas Class II corresponds to a regime switch at t_{last} from the antibubble to a random walk price trajectory. We then fitted each trajectory by the anti-bubble formula (2) and obtained two sets of N parameters vectors for each class.
The idea of the test by Zhou an Sornette (2005) is to quantify the differences in the distributions of the parameters vectors in the two classes, to decide whether the realized values of the SP500 belongs to Class I or Class II: if the anti-bubble pattern proceeds up to t_{last} + 6 months with a price trajectory close to the extrapolation of the log-periodic formula made at t_{last}, then we can expect that the fitted parameters of (2) using data from t_c up to t_{last} + 6 months should be rather close to the values found for Class I and far from those found for Class II. Instead, if the market prices switch to a random walk after t_{last}, we should find the estimated parameters of (2) to be closer to the values found for Class II and far from those found for Class I. This specific test belongs to a a large class of pattern recognition methods and we refer to Zhou an Sornette (2005) and references therein for more details. Particularly, a key element in pattern recognition methods is to define two types of errors that can occur in a classification scheme using a parameter x: a type I error occurs when the hypothesis, which is true, is rejected (i.e. a so-called "false negative") and it is measured as the proportion of the objects in class I with a deviation $|X - x_0|$ greater than a certain $|x - x_0|$:

$$P_1(x) = \lim_{N\to\infty} \frac{\sharp\{X : |X - x_0| > |x - x_0| \& X \in I\}}{N} \tag{3}$$

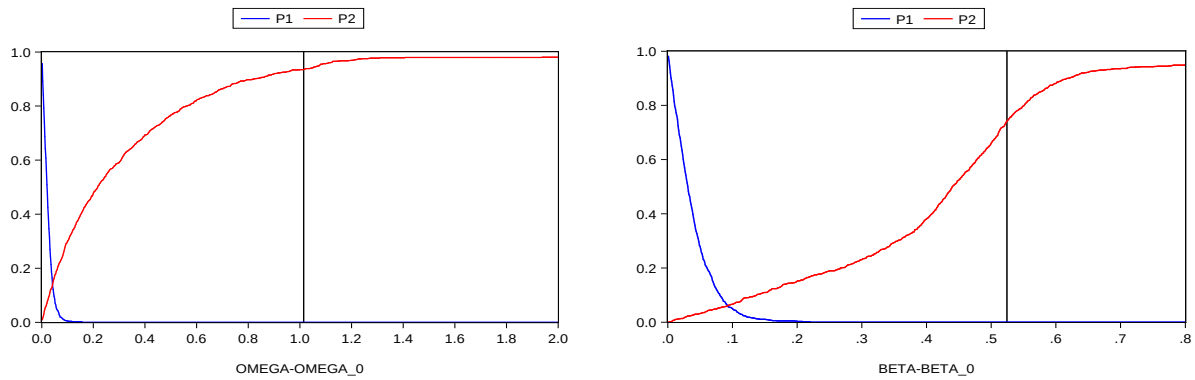
where $\sharp$ is the operator counting the number of elements in a given set. Instead, an error of type II takes place when an hypothesis, which is false, is accepted (a so called "false positive") and it is measured as the proportion of the objects in class II with a deviation $|X - x_0|$ smaller than $|x - x_0|$:

$$P_2(x) = \lim_{N\to\infty} \frac{\sharp\{X : |X - x_0| < |x - x_0| \& X \in II\}}{N} \tag{4}$$

and where $\lim_{x\to x_0} P_1(x) = 1$, $\lim_{x\to x_0} P_2(x) = 0$, $\lim_{|x-x_0|\to\infty} P_1(x) = 0$, and $\lim_{|x-x_0|\to\infty} P_2(x) = 1$. In order to have a confirmation for the continuation of the anti-bubble, the ideal case would be to have both $P_1(x)$ large and $P_2(x)$ small: the first condition informs us that the realized deviation of the estimated parameter x is within the normal fluctuations of objects in Class I, whereas the second condition highlights that it is improbable to obtain such a value when the SP500 does not follow an anti-bubble. Moreover, Zhou and Sornette (2005) highlight that P_1 can also be interpreted as the probability of existence of the LPPL antibubble. Summarizing, a small P_1 and a large P_2 are good signals of the end of the LPPL anti-bubble. Instead, a large P_1 and a small P_2 are good signals of the continuation of the LPPL anti-bubble.

We now apply this test to the realized values of the SP500, and we report in Figure 3 the probabilities P_1 and P_2 corresponding to the reference anti-bubble from t_c = 09/10/2007 to t_{last} = 27/08/2010 as functions of ω and β, i.e the key parameters of the log-periodic model (2). The vertical lines indicate the realized values of $|x - x_0|$, where x_0 is the reference value of the parameter estimated with data up to t_{last} = 27/08/2010, while x is the estimated parameter using data up to t_{last} + 6 months = 28/02/2011. The results with the remaining parameters are qualitatively the same and are not reported here for sake of interest and space.

Figure 3: *Probabilities P_1 (blue line), and P_2 (red line) corresponding to the reference anti-bubble from t_c = 09/10/2007 to t_{last} = 27/08/2010 as functions of ω and β. The vertical lines indicate the realized values of $|x - x_0|$, where x_0 is the reference value while x is the estimated parameter using data up to t_{last} + 6 months = 28/02/2011.*



It is immediate to see that the P_1's are very small while the P_2's are very large for both the parameters: this evidence therefore suggests that the anti-bubble in the SP500 has ended and seems to confirm the findings already emerged in the previous sections.

6 Conclusions

This paper reviewed the *ex-ante* forecast of the SP500 index discussed in Fantazzini (2010a) which covered the time sample 14/04/2009 - 09/10/2010. We showed that the realized values of the SP500 index trailed the forecasted values quite well, moving inside the forecast confidence bands for over a year. Moreover, an important turning point in April 2010 was also correctly forecasted. However, we also found that the speech by the FED Chairman at Jackson Hole (Wyo., USA) on the 27/08/2010 had a significant statistical impact on the stock market: had additional rounds of Quantitative Easing not been granted that day, the SP500 would have most likely kept going down below the critical 1000 level and beyond, with the all the potential negative effects on the real economy that such an event would have probably determined. Therefore, we found that the anti-bubble which started in October 2007 ended almost three years later in August 2010, similarly to the first anti-bubble on the SP500, which started in August 2000 and ended in August 2003, see Zhou and Sornette (2005). This similarity seems to suggest that it takes almost three years of aggressive monetary policies to gradually detune and end the decreasing market phases following the burst of serious bubbles like the "dot-com" bubble in 2000 and the real estate and oil bubbles in 2007-2008. Further research concerning the monetary policies necessary to deal with the aftermath of bubble bursts is therefore called for.

References

[1] Bernanke, B. (2010) Speech at the 2010 Annual Economic Policy Symposium, (hosted by) The Federal Reserve Bank of Kansas City, Jackson Hole, (Wyo., USA), 27-08-2010.

[2] Cvijović, D. and J. Klinowski (1995) "Taboo search An approach to the multiple minima problem", *Science*, **267(5)**, 664-666.

[3] Fantazzini, D. (2010a) "Modelling and forecasting the global financial crisis: Initial findings using heterosckedastic log-periodic models", *Economics Bulletin*, **30(3)**, 1833-1841.

[4] Fantazzini, D. (2010b) "A Mini Bubble in the SP500 ready to burst?". Available at `http://www.scribd.com/doc/28310962/A-Mini-Bubble-in-the-SP500-ready-to-Burst`.

[5] Fantazzini, D. (2010c) "Modelling Bubbles And Anti-Bubbles In Bear Markets: A Medium-Term Trading Analysis", In *Handbook of Trading*, ed. G. Gregoriou, 365-388. New York: McGraw-Hill.

[6] Geraskin P. and D. Fantazzini (2011) "Everything You Always Wanted to Know about Log Periodic Power Laws for Bubble Modelling but Were Afraid to Ask", *European Journal of Finance*, forthcoming.

[7] Jiang, Z.Q., Zhou, W.H., Sornette, D., Woodard, R., Bastiaensen, K., and P. Cauwels (2010) "Bubble Diagnosis and Prediction of the 2005-2007 and 2008-2009 Chinese stock market bubbles", *Journal of Economic Behavior and Organization*, **74(3)**, 149-162.

[8] Johansen, A., Ledoit, O. and D., Sornette (2000) "Crashes as critical points", *International Journal of Theoretical and Applied Finance*, **3(2)**, 219-255.

[9] Lee, J. and M. Strazicich (2003), "Minimum lm unit root test with two structural breaks", *Review of Economic and Statistics*, **85**, 1082-1089.

[10] Lumsdaine, R. and Papell, D. (1997) "Multiple trend breaks and the unit-root hypothesis", *Review of Economics and Statistics*, **79**, 212-218.

[11] Maddala, G.S. and Kim, I.M. (2003), *Unit Root, Cointegration and Structural Change*, Cambridge University Press, Fifth Edition.

[12] Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R*. New York: Wiley.

[13] Sornette, D. (2003a) *Why Stock Markets Crash (Critical Events in Complex Financial Systems)*, Princeton University Press: Princeton, NJ.

[14] Sornette, D. (2003b) "Critical market crashes", *Physics Reports*, **378(1)**, 1-98.

[15] Sornette, D. and W.X. Zhou (2006) "Predictability of Large Future Changes in major financial indices", *International Journal of Forecasting*, **22**, 153-168.

[16] Sornette, D., Woodard, R. and W.X. Zhou (2008) "The 2006-2008 Oil Bubble and Beyond. ETH Zurich preprint", http://arXiv.org/abs/0806.1170

[17] Sornette, D., Johansen, A. and J.P., Bouchaud (1996) "Stock market crashes, Precursors and Replicas", *Journal de Physique I*, **6**, 167-175.

[18] Zhou, W.X., and D., Sornette (2003) "2000-2003 Real Estate Bubble in the UK but not in the USA", *Physica A*, **329**: 249-263.

[19] Zhou, W.X., and D., Sornette (2005) "Testing the Stability of the 2000-2003 US Stock Market "Antibubble"", *Physica A*, **348**, 428-452.

[20] Zhou, W.X., and D., Sornette (2006) "Predictability of Large Future Changes in Major Financial Indices", *International Journal of Forecasting*, **22**, 153-168.

[21] Zhou, W.X., and D., Sornette (2006), "Is There a Real-Estate Bubble in the US?", *Physica A*, **361**, 297-308.

[22] Zhou, W.X., and D., Sornette (2008) "Analysis of the real estate market in Las Vegas: Bubble, seasonal patterns, and prediction of the CSW indexes", *Physica A*, **387**, 243-260.

[23] Zhou, W.X., and D., Sornette (2009), "A case study of speculative financial bubbles in the South African stock market 2003-2006", *Physica A*, **388**, 869-880.

[24] Zivot, E. and D. Andrews, (1992) "Further evidence on the great crash, the oilprice shock and the unit root hypothesis", *Journal of Business and Economic Statistics*, **10**, 251-270.

Appendix 6 "Analysis of multidimensional probability distributions with copula functions - Part I."

Fantazzini, D. (2011b). Analysis of multidimensional probability distributions with copula functions - Part I. *Applied Econometrics*, 22(2), 98-134. http://pe.cemi.rssi.ru/pe_2011_2_98-134.pdf

Permisson to post the Publisher's version/PDF: according to http://appliedeconometrics.cemi.rssi.ru/AppEc.files/contents_en.html, "*full texts are available for downloading with approximately one year lag*". In some cases, this 1-year lag can be shorter.

Д. Фантаццини

Моделирование многомерных распределений с использованием копула-функций. I

Проблематика копула-функций, их свойств, способов подбора под конкретные исходные данные, оценивания, прикладных возможностей крайне скупо представлена в мировой специальной литературе и почти никак — в отечественной. При этом уже существуют впечатляющие примеры их прикладного использования в ситуациях, когда построение, статистическое оценивание и анализ многомерных распределений вероятностей оказывается необходимым инструментом исследования, а использование для таких задач модели многомерного нормального (гауссовского) закона не отражает специфики имеющихся исходных статистических данных. Есть основания утверждать, что модели, основанные на копула-функциях, окажутся особенно востребованными в прикладных эконометрических исследованиях, посвященных задачам оценки, анализа и управления финансовыми и страховыми рисками, доходностями разных финансовых инструментов. Предлагаемый в данном номере журнала материал является, по существу, фрагментом готовящегося к изданию учебника С. А. Айвазяна и Д. Фантаццини «Методы эконометрики. Продвинутый уровень».

Перевод на русский язык осуществлен А. В. Кудровым под научной редакцией С. А. Айвазяна.

Ключевые слова: многомерное распределение, эллиптические копула-функции, архимедовы копула-функции, иерархические копула-функции.

JEL classification: C69, C49.

Содержание

Введение

1. Копула-функции
- 1.1. Определения и основные свойства
- 1.2. Плотности копула-функций
- 1.3. Эмпирические приложения в статистическом пакете *R*: двумерные границы Фреше–Хёффдинга и копула-функция, отвечающая случаю независимости

2. Эллиптические копула-функции
- 2.1. Обзор
- 2.2. Нормальная копула-функция
- 2.3. *T* копула-функция (копула Стьюдента)
- 2.4. Численные примеры: оценивание эллиптических копула-функций
- 2.5. Эмпирические приложения в статистическом пакете *R*: эллиптические копула-функции

3. Архимедовы копула-функции
- 3.1. Основные понятия и определения
- 3.2. Некоторые однопараметрические двумерные архимедовы копула-функции
- 3.3. Многомерные перестановочные архимедовы копула-функции
- 3.4. Вложенные архимедовы копула-функции
- 3.5. Моделирование наблюдений из архимедовых копула-функций
- 3.6. Эмпирические приложения в статистическом пакете *R*: архимедовы копула-функции

Список литературы

Введение

Доказательства того, что не все совместные распределения экономических характеристик могут быть описаны при помощи многомерного нормального закона, широко представлены в различных работах, начиная с (Mills, 1927). Наиболее распространенными статистиками, позволяющими охарактеризовать степень отклонения от нормального распределения, являются коэффициенты эксцесса и асимметрии. Последние исследования показывают, что для финансовых данных также наблюдается отклонение их совместного распределения от многомерного нормального закона, а именно, отклонение в форме зависимостей компонент анализируемого многомерного признака. Один из таких примеров — асимметрия зависимостей доходностей рыночных активов, которая больше в периоды спада на рынке, чем в периоды рыночного роста, см. (Erb et al., 1994; Longin, Solnik, 2001; Ang, Chen, 2002; Patton, 2004, 2006а, б) и др.

Основной вывод этих статей заключается в том, что многомерное нормальное распределение не является хорошей моделью для описания совместного распределения многих экономических и финансовых переменных. Это приводит к проблеме поиска более адекватных многомерных моделей. Теория копула-функций — один из возможных способов ее решения.

Начало теории копула-функций было положено работами (Hoeffding, 1940) и (Sklar, 1959), но статистическое моделирование с использованием копула-функций в приложениях появилось относительно недавно и относится к концу 1990-х гг. Количество статистических исследований на тему копула-функций увеличивается достаточно быстро, см. (Wang, 1998; Frees, Valdez, 1998; Embrechts et al., 1999). Сегодня копула-функции успешно применяются в финансах и страховании, биостатистике (например, (Lambert, Vandenhende, 2002)), гидрологии (например, (Zhang, Singh, 2006)) и климатологии (например, (Salvadori, De Michele, 2007)). Среди учебников по теории копула-функций следует отметить (Joe, 1997) и (Nelsen, 1999), где представлено введение в теорию, тогда как в (Cherubini et al., 2004) и (Malevergne, Sornette, 2006) излагается использование техники копула-функций в финансовых приложениях.

Копула-функция является функцией, агрегирующей всю информацию относительно структуры зависимости между компонентами случайного вектора. Когда в качестве компонент копула-функции берутся частные функции распределения, которые необязательно принадлежат одному и тому же семейству распределений, получаем многомерную функцию распределения. Как следствие, эта теория позволяет достаточно гибко моделировать структуру зависимости между различными переменными, которые могут иметь разные частные распределения.

Несмотря на то что теория копула-моделей исследована относительно полно, проблема оценивания и статистические выводы для копула-моделей, в определенном контексте, все еще требуют дальнейших исследований, см. (Genest, Favre, 2007). За последние годы предложены различные методы оценивания параметров копула-функций, начиная с параметрических (Jondeau, Rockinger, 2003; Patton, 2004, 2006а, б; Fantazzini, 2009б), полупараметрических (Genest et al., 1995; Breymann et al., 2003; Fantazzini, 2010), и заканчивая непараметрическими методами (Fermanian, Scaillet, 2003). Более того, недавние исследования направлены на исследование «смеси» указанных выше методов, которые позволили бы сэкономить время на вычислениях (Bouyé et al., 2000; Marshal, Zeevi, 2002; Cherubini et al., 2004).

В данной статье будет представлен некоторый обзор теории копула-функций, основные методы оценивания, а также сделана попытка ответить на, вероятно, самый важный вопрос теории копула-функций: как следует подбирать корректную копула-функцию?

1. Копула-функции

1.1. Определения и основные свойства

С помощью копула-функций описываются законы многомерного распределения вероятностей. Они определяются частными одномерными распределениями анализируемого многомерного закона и характером зависимостей, существующих между компонентами рассматриваемой многомерной случайной величины.

Определение 1. Функция $C(u_1,u_2,...,u_n)$ от n переменных, определенная на единичном гиперкубе $I^n=[0,1]^n$ (т.е. $u_i\in[0,1],\ i=1,2,\ldots,n$), называется *копула-функцией*, если она обладает следующими свойствами:

1) область значений функции — единичный интервал [0,1];
2) если $u_i=0$ по крайней мере для одного $i\in\{1,2,...,n\}$, то $C(u_1,u_2,...,u_n)=0$;
3) $C(1,...,1,u_i,1,...,1)=u_i$ для любых $u_i\in[0,1]$;
4) $C(u_1,u_2,...,u_n)$ является *n*-возрастающей функцией в том смысле, что для всех $(a_1,\ldots,a_n)$, $(b_1,\ldots,b_n)\in[0,1]^n$ с $a_i\le b_i$ справедливо неравенство:

$$\sum_{i_1=1}^{2}\ldots\sum_{i_n=1}^{2}(-1)^{i_1+\ldots+i_n}C(u_{1i_1},\ldots,u_{ni_n})\ge 0,$$

где $u_{j1}=a_j$ и $u_{j2}=b_j$ для всех $j\in\{1,\ldots,n\}$.

Например, если $n=2$, четвертое свойство принимает следующий вид:
для всех $(a_1,a_2),(b_1,b_2)\in[0,1]^2$ с $a_i\le b_i$ имеем,

$$\begin{aligned}
&\sum_{i_1=1}^{2}\sum_{i_2=1}^{2}(-1)^{i_1+i_2}C(u_{1i_1},u_{2i_2})\ge 0\\
&\Leftrightarrow \sum_{i_1=1}^{2}[(-1)^{i_1+1}C(u_{1i_1},u_{21})+(-1)^{i_1+2}C(u_{1i_1},u_{22})\ge 0\\
&\Leftrightarrow (-1)^2C(u_{11},u_{21})+(-1)^3C(u_{11},u_{22})+(-1)^3C(u_{12},u_{21})+(-1)^4C(u_{12},u_{22})\ge 0\\
&\Leftrightarrow C(u_{11},u_{21})-C(u_{11},u_{22})-C(u_{12},u_{21})+C(u_{12},u_{22})\ge 0,
\end{aligned}$$

где $u_{j1}=a_j$ и $u_{j2}=b_j$ при $j\in\{1,2\}$. Таким образом, свойство 4 равносильно тому, что

$$C(a_1,a_2)-C(a_1,b_2)-C(b_1,a_2)+C(b_1,b_2)\ge 0\,. \qquad (1)$$

Следовательно, (1) можно интерпретировать как:

$$P(0 \le u_1 \le a_1,\ 0 \le u_2 \le a_2) - P(0 \le u_1 \le a_1,\ 0 \le u_2 \le b_2) -$$
$$-P(0 \le u_1 \le b_1,\ 0 \le u_2 \le a_2) + P(0 \le u_1 \le b_1,\ 0 \le u_2 \le b_2) \ge 0.$$

Теперь сформулируем теорему Склара, которая показывает роль копула-функций в описании многомерных распределений вероятностей.

Теорема 1 (теорема Склара). *Пусть $H(\cdot)$ — n-мерная функция распределения с частными распределениями $F_1,...,F_n$. Тогда существует n-мерная копула-функция $C(\cdot)$ такая, что для всех действительных $x_1,...,x_n$:*

$$H(x_1,\ldots,x_n) = C(F_1(x_1),\ldots,F_n(x_n)). \tag{2}$$

Если все частные функции распределения непрерывны, то копула-функция определена единственным образом; в противном случае $C(\cdot)$ определена единственным образом лишь на области определения $\mathrm{Ran}\,F_1 \times \mathrm{Ran}\,F_2 \times \ldots \times \mathrm{Ran}\,F_n$, где Ran — область значений частных функций распределения. Обратно, если $C(\cdot)$ — копула-функция, а $F_1,\ldots,F_n$ — функции распределения, то функция $H(\cdot)$, определяемая выражением (2), *является совместной функцией распределения с частными распределениями $F_1,\ldots,F_n$.*

Доказательство. См. (Sklar, 1959, 1996; Joe, 1997; Nelsen, 1999).

Последнее утверждение теоремы представляет большой интерес в задаче моделирования многомерных функций распределения, поскольку из него следует, что можно связывать вместе любые n $(n \ge 2)$ одномерных функций распределения разного типа (не обязательно из одного семейства), используя любую копула-функцию, для того чтобы получить двумерные или многомерные функции распределения.

Следствие. *Пусть $F_1^{(-1)}(\cdot),\ldots,F_n^{(-1)}(\cdot)$ — обратные (в обобщенном смысле) функции частных распределений. Тогда для каждого $(u_1,...,u_n)$ из единичного n-мерного куба существует единственная копула-функция $C:[0,1]\times\ldots\times[0,1]\to[0,1]$ такая, что*

$$C(u_1,...,u_n) = H(F_1^{(-1)}(u_1),\ldots,F_n^{(-1)}(u_n)). \tag{3}$$

Доказательство. См. теорему 2.10.9 в (Nelsen, 1999).

Таким образом, копула-функция — это такая функция, которая, с использованием знания об одномерных частных распределениях, позволяет получить многомерную функцию распределения, поскольку функция распределения случайного вектора полностью описывает его вероятностную структуру, куда, в частности, входит структура зависимости его компонент. Копула-функции дают возможность разделить описание распределения случайного вектора на две части: частные распределения компонент и структура их зависимостей.

Перечислим кратко варианты использования результатов теоремы Склара и рассмотренного следствия:

- извлечение копула-функций из хорошо известных многомерных распределений, например, из многомерного нормального распределения, из многомерного распределения Стьюдента и т. д.;
- конструирование новых многомерных распределений, связывая произвольные одномерные распределения при помощи некоторой копула-функции;
- в то время как соотношение (2) обычно служит отправной точкой в задаче моделирования, соотношение (3) является теоретическим инструментом для получения копула-функции из многомерной функции распределения.

Следует отметить, что для дискретных распределений копула-функции имеют не столь простую форму, как в случае непрерывных распределений. Покажем это на следующем примере.

Пример 1. Рассмотрим две одинаковые кости, на которых выпадают два числа X_1 и X_2 из множества $\{1,2,\ldots,6\}$. Предположим, что мы знаем значение X_1 и можем делать ставки на значения X_2, которые выпадают на второй кости. Важным вопросом является следующий: существует ли взаимосвязь или зависимость между этими двумя случайными величинами, т. е. возможно ли получить некоторую информацию относительно значений X_2, зная X_1. Мы знаем, что каждая случайная величина полностью описывается своей функцией распределения $F_i(x) = P(X_i \le x)$. В нашем случае мы имеем $F_1(\cdot) = F_2(\cdot) = F(\cdot)$.

Однако в общем случае одномерная функция распределения не дает нам никакой информации относительно совместного распределения двумерной случайной величины (X_1, X_2). Правда, в нашем случае предполагается независимость (X_1, X_2), поэтому совместное распределение (X_1, X_2) имеет вид:

$$P(X_1 \le x_1, X_2 \le x_2) = F(x_1) \cdot F(x_2).$$

Для того чтобы получить полное описание (X_1, X_2), мы использовали два элемента: частные распределения и тип взаимосвязи (в нашем случае — независимость). Возможность разделения между частными распределениями и зависимостью — это именно то, что утверждает теорема Склара. Поскольку мы рассматриваем дискретные частные распределения, то копула-функция будет определена единственным образом только на $\operatorname{Ran} F_1 \times \operatorname{Ran} F_2$. В случае независимости имеем:

$$C\left(\frac{i}{6}, \frac{j}{6}\right) = \frac{i}{6} \cdot \frac{j}{6}, \quad i, j = 1,\ldots,6.$$

Любая копула-функция, удовлетворяющая этому ограничению, является подходящей: например, $C(u,v) = u \cdot v$ будет естественным выбором.

Важное свойство, которым обладают все копула-функции — это инвариантность относительно строго возрастающих преобразований случайных величин (см. приведенную ниже теорему 2).

Теорема 2. *Рассмотрим n случайных величин* $X_1,\ldots,X_n$, *зависимость между которыми определяется копула-функцией* $C(\cdot)$. *Если преобразования* $T_i:\mathrm{R}\to\mathrm{R},\ i=1,\ldots,n$ *определяются строго возрастающими функциями (т. е.* $\partial_x T_i>0$*), то структура зависимости случайных величин* $T_1(X_1),\ldots,T_n(X_n)$ *определяется той же самой копула-функцией* $C(\cdot)$.

Доказательство. См. (Schweizer, Wolff, 1976, 1981), для двумерного случая — (Malevergne, Sornette, 2006).

Последнее утверждение означает, что *строго возрастающие преобразования не меняют структуру зависимости*. Это значит, что при заданных частных распределениях сохраняется структура зависимости, которая определяется копула-функцией.

Более того, можно показать, что любая копула-функция ограничена так называемыми *границами Фреше–Хёффдинга*:

$$W(u_1,...,u_n)\ \le\ C(u_1,...,u_n)\ \le M(u_1,...,u_n),$$

где

$$W(u_1,\ldots,u_n)=\max(u_1+\ldots+u_n-n+1,0),\qquad M(u_1,\ldots,u_n)=\min(u_1,\ldots,u_n).$$

В качестве примера рассмотрим две равномерно распределенные на [0,1] случайные величины U_1 и U_2. Если $U_1=U_2$, то эти две случайные величины имеют копула-функции вида

$$C(u_1,u_2)=P(U_1\le u_1,U_1\le u_2)=\min(u_1,u_2)$$

и называются абсолютно зависимыми.

Ту же самую копула-функцию можно получить, если взять $X_2=T(X_1)$, где $T(\cdot)$ — монотонно возрастающее преобразование. Такие случайные величины X_1 и X_2 называются *комонотонными*. Прямо противоположным понятию комонотонности является *противомонотонность* случайных величин. В качестве примера возьмем равномерные случайные величины (U_1,U_2), для которых $U_2=1-U_1$. Копула-функция случайного вектора (U_1,U_2) равна

$$C(u_1,u_2)=P(U_1\le u_1,1-U_1\le u_2)=P(U_1\le u_1,1-u_2\le U_1)=u_1+u_2-1\ \text{ при }\ u_1+u_2>1.$$

В противном случае она равна нулю (рис. 1).

В двумерном случае границы Фреше–Хёффдинга являются копула-функциями, но для случаев большей размерности нижняя граница Фреше–Хёффдинга W уже не является n-возрастающей (см. свойство 4 в Определении 1). Однако, неравенство в левой части нельзя улучшить, поскольку для любого фиксированного u из единичного n-куба существует $C(\cdot)$ такая, что $W(u)=C(u)$, см. доказательство теоремы 2.10.12 в (Nelsen, 1999).

Копула-функции равномерно непрерывны и (почти всюду) дифференцируемы по каждому из аргументов — см. теоремы 2.2.4 и 2.2.7 в (Nelsen, 1999). Более детальная информация относительно этих и других свойств приведена в (Joe, 1997; Nelsen, 1999, 2006).

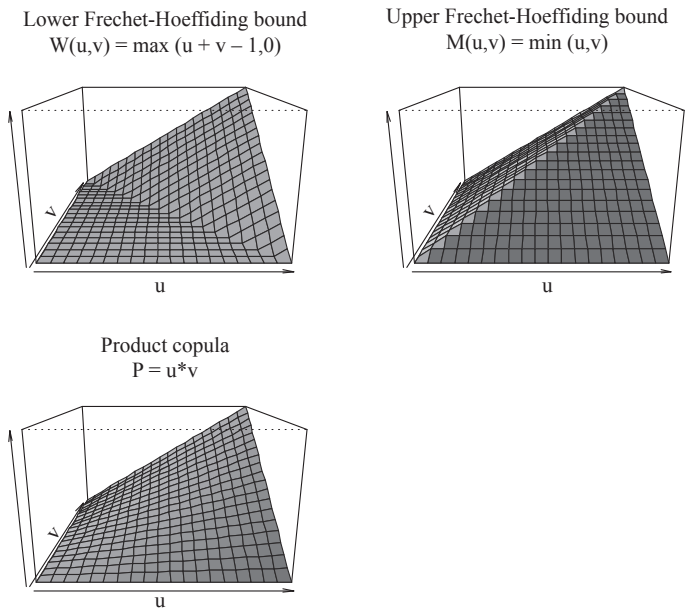


Рис. 1. Графики копула-функций *W* и *M*, а также копула-функция *P*, соответствующая независимым случайным величинам

1.2. Плотности копула-функций

Для копула-функций, аналогично функциям распределения, можно определить понятие плотности. В частности, плотность $c(u_1,u_2,\ldots,u_n)$, ассоциированная с копула-функцией $C(u_1,u_2,\ldots,u_n)$, определяется соотношением

$$c(u_1,u_2,\ldots,u_n)=\frac{\partial^n C(u_1,u_2,\ldots,u_n)}{\partial u_1\ldots\partial u_n}.$$

Плотность может быть использована для того, чтобы определить так называемые *абсолютно непрерывную* A_C и *сингулярную* S_C компоненты копула-функции $C(\cdot)$:

$$A_C(u_1,u_2,\ldots,u_n)=\int_0^{u_1}\int_0^{u_2}\ldots\int_0^{u_n}\frac{\partial^n C(s_1,s_2,\ldots,s_n)}{\partial s_1\ldots\partial s_n}ds_1\ldots ds_n,$$
$$S_C(u_1,u_2,\ldots,u_n)=C(u_1,u_2,\ldots,u_n)-A_C(u_1,u_2,\ldots,u_n).$$

Копула-функция, для которой $C(\cdot)=A_C(\cdot)$ на I^n, называется *абсолютно непрерывной*. В противном случае копула-функция называется *сингулярной.*

Существуют копула-функции, которые имеют как абсолютно непрерывную, так и сингулярную компоненты. Например, копула-функция $P=u_1u_2u_3$ является абсолютно непрерывной, поскольку

$$\frac{\partial C(s_1,s_2,s_3)}{\partial s_1 \partial s_2 \partial s_3} = 1$$

и для любого $(u_1,u_2,u_3) \in I^3$ имеем:

$$A_C(u_1,u_2,u_3) = \int_0^{u_1}\int_0^{u_2}\int_0^{u_3} \frac{\partial C(s_1,s_2,s_3)}{\partial s_1 \partial s_2 \partial s_3} ds_1 ds_2 ds_3 = u_1 u_2 u_3 = P.$$

Более того, применяя следствие теоремы Склара и рассматривая непрерывные случайные величины, можно видеть, что плотность копула-функции $c(F_1(x_1),\ldots,F_n(x_n))$ ассоциирована с плотностью совместной функции распределения $H(\cdot)$, обозначенной как $f_H(\cdot)$, следующим образом (*каноническое представление*):

$$f_H(x_1,\ldots,x_n) = \frac{\partial^n \left[C(F_1(x_1),\ldots,F_n(x_n))\right]}{\partial F_1(x_1)\ldots\partial F_n(x_n)} \cdot \prod_{i=1}^n f_i(x_i) = c(F_1(x_1),\ldots,F_n(x_n)) \cdot \prod_{i=1}^n f_i(x_i),$$

где

$$c(F_1(x_1),\ldots,F_n(x_n)) = \frac{f(x_1,\ldots,x_n)}{\prod_{i=1}^n f_i(x_i)}. \quad (4)$$

1.3. Эмпирические приложения в статистическом пакете *R*: двумерные границы Фреше–Хёффдинга и копула-функция, отвечающая случаю независимости

Если необходимо изобразить линии уровня двумерных границ Фреше–Хёффдинга и копула-функции, соответствующей случаю независимости, то для этого можно использовать следующий *R* код (результаты представлены на рис. 2):

```
library (fCopulae)
# Generate Grid:
N = 50; uv = grid2d (x = (0:N)/N); u = uvx; v = uvy
# Compute Frechet and Product Copulae:
W = matrix (apply (cbind (u+v-1, 0), 1, max), ncol = N+1)
Pi = matrix (u*v, ncol = N+1)
M = matrix (apply (cbind (u, v), 1, min), ncol = N+1)
# Create Perspective Plots:
persp (z = W, theta = -40, phi = 30, main = "Lower Frechet",
cex = 0.5, ticktype = "detailed", col = "steelblue")
persp (z = Pi, theta = -40, phi = 30, main = "Pi Copula",
cex = 0.5, ticktype = "detailed", col = "steelblue")
persp (z = M, theta = -40, phi = 30, main = "Upper Frechet",
cex = 0.5, ticktype = "detailed", col = "steelblue")
```

```
# Create Contour Plots:
contour (W, xlab = "u", ylab = "v", main = "Lower Frechet")
contour (Pi, xlab = "u", ylab = "v", main = "Pi Copula")
contour (M, xlab = "u", ylab = "v", main = "Upper Frechet")
```

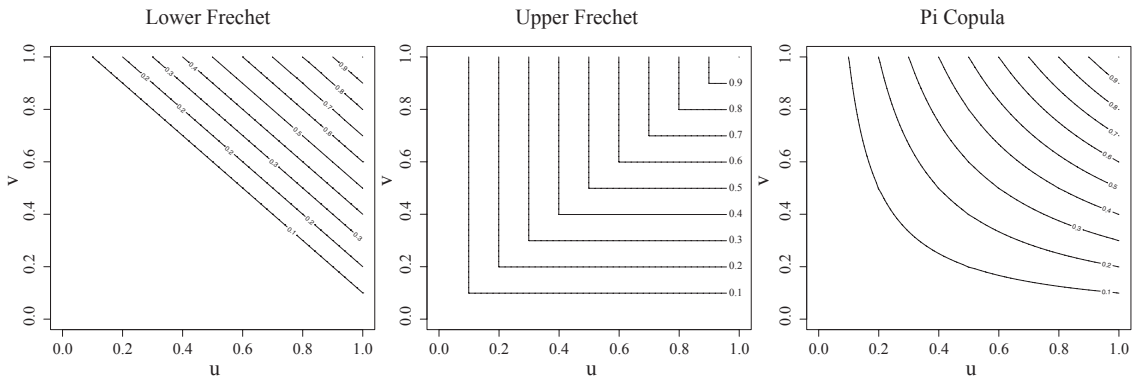


Рис. 2. Графики линий уровня двумерных границ Фреше–Хёффдинга и копула-функции, отвечающей случаю независимости

2. Эллиптические копула-функции

2.1. Обзор

Класс эллиптических распределений включает, главным образом, класс симметричных распределений, который весьма популярен в актуарной математике, страховании и финансах. Этот класс включает интересные примеры многомерных распределений, и многие из них имеют некоторые общие свойства с многомерным нормальным распределением. Они позволяют моделировать многомерные экстремальные события, формируя зависимость, не совпадающую с зависимостью многомерного нормального распределения, и использовать распределение с большим эксцессом, чем эксцесс нормального распределения. Кроме того, с их помощью можно моделировать феномен «тяжелых хвостов», который часто наблюдается для финансовых данных, см. (Embrechts et al., 1999; Schmidt, 2002).

Эллиптические копула-функции — это попросту копула-функции многомерных распределений эллиптического типа. Если следовать работе (Fang et al., 1987), то эллиптические распределения определяются следующим образом.

Определение 2 (эллиптические распределения). Пусть $\mathbf{X}$ есть n-мерный случайный вектор и $\Sigma \in \mathbb{R}^{n\times n}$ — симметричная и неотрицательно определенная матрица. Если существует $\mu \in \mathbb{R}^n$ и функция $\psi : \mathbb{R} \to \mathbb{R}^+$ такие, что характеристическая функция $\mathbf{X} - \mu$ имеет вид

$$\phi_{\mathbf{X}-\mu}(t) = \exp(it^{\mathrm{T}}\mu)\psi\left(-\frac{1}{2}t^{\mathrm{T}}\Sigma t\right),$$

для любого $t \in \mathbb{R}^n$, то $\mathbf{X}$ называют случайным вектором, имеющим *распределение эллиптического типа* с параметрами μ, Σ.

Копула-функция $C(\cdot)$ называется *эллиптической*, если она отвечает распределению эллиптического типа. Функция плотности распределения эллиптического типа, если она существует, имеет вид

$$f(\mathbf{x}) = \frac{\gamma}{\sqrt{|\Sigma|}} g\left((\mathbf{x}-\mu)^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}-\mu)\right), \quad \mathbf{x} \in \mathbb{R}^n$$

для некоторой функции $g : \mathbb{R} \to \mathbb{R}^+$, которая удовлетворяет условию: $\int_0^\infty g(x)dx < \infty$. Функцию $g(\cdot)$ называют оператором плотности распределения эллиптического типа или *генерирующей функцией*, см. (Fang et al., 1987). Нормирующая константа γ может быть определена явным образом с использованием перехода к полярным координатам, см. (Landsman, Valdez, 2003). В результате получим:

$$\gamma = \left(2\pi \int_0^\infty g(z)dz\right)^{-1}.$$

В качестве примеров распределений эллиптического типа следует привести нормальное распределение, распределение Коши и распределение Стьюдента, логистическое распределение, распределение Лапласа, распределение Котца и экспоненциальное распределение, см. табл. 1.

Таблица 1. Список генерирующих функций $g(t)$ и нормирующих констант γ для некоторых копула-функций эллиптического типа

Распределение	Генератор	Константа
Нормальное	$e^{-t/2}$	$(2\pi)^{-1}$
Коши	$(1+2t)^{-3/2}$	$(2\pi)^{-1}$
Стьюдента	$\left(1+\frac{2t}{\nu}\right)^{\frac{-\nu-2}{2}}$	$(2\pi)^{-1}$
Логистическое	$e^t(1+e^{-t})^{-2}$	π^{-1}
Лапласа	$e^{-\sqrt{t}}$	$(2\pi)^{-1}$
Котца	$e^{-rt/2}$	$r(2\pi)^{-1}$
Экспоненциальное	$e^{-r(t/2)^s}$	$sr^{1/s}\left(2\pi\Gamma(1/s)\right)^{-1}$

Отметим, что в некоторых работах используются иные определения генерирующих функций $g(\cdot)$, где $t/2$ заменено на t, что влияет на значение нормирующих констант γ.

Пример 2. Рассмотрим двумерный случай. Предположим, что $\mathbf{X} = (X, Y)^{\mathrm{T}}$ и

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad |\Sigma| = 1 - \rho^2, \quad -1 < \rho < 1.$$

Тогда $\mathbf{X}$ имеет функцию плотности распределения

$$f_\rho(x, y) = \frac{\gamma}{\sqrt{1-\rho^2}} g\left(\frac{x^2 - 2\rho xy + y^2}{1-\rho^2}\right), \quad (x, y) \in \mathbb{R}^2.$$

Следовательно, в соответствии с теоремой Склара, двумерная копула-функция эллиптического типа имеет вид:

$$C(u_1, u_2; \rho) = \frac{\gamma}{\sqrt{1-\rho^2}} \int_{-\infty}^{F_g^{-1}(u_1)} \int_{-\infty}^{F_g^{-1}(u_2)} g\left(\frac{x^2 - 2\rho xy + y^2}{1-\rho^2}\right) dx\, dy. \tag{5}$$

Далее более детально рассмотрим две наиболее распространенные копула-функции эллиптического типа, а именно, нормального распределения и распределения Стьюдента.

2.2. Нормальная копула-функция

Копула-функция двумерного нормального распределения с коэффициентом корреляции ρ может быть вычислена из уравнения (5) при $g = e^{-t/2}$ и $\gamma = 1/(2\pi)$:

$$C^{Normal}(u_1, u_2; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\phi^{-1}(u_1)} \int_{-\infty}^{\phi^{-1}(u_2)} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right) dx\, dy,$$

что может быть представлено в виде:

$$C(u_1, u_2; \rho) = \Phi^2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho),$$

где $\Phi^2(\cdot)$ обозначает функцию двумерного нормального распределения с единичными дисперсиями компонент, корреляцией ρ и нулевым средним значением, а $\Phi^{-1}(\cdot)$ — обратная функция для одномерного стандартного нормального распределения.

В общем случае многомерная нормальная копула-функция определяется следующим образом:

$$C(u_1, u_2, \ldots, u_n; \Sigma) = \Phi^n(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_n); \Sigma),$$

где $\Phi^n(\cdot, \ldots, \cdot; \Sigma)$ — это n-мерная нормальная функция распределения с нулевыми средними значениями и ковариационной матрицей Σ.

Плотность нормальной копула-функции может быть получена из уравнения (4) с использованием канонического представления:

$$c(\Phi(x_1),\ldots,\Phi(x_n)) = \frac{f^{Gaussian}(x_1,...,x_n)}{\prod_{i=1}^{n} f_i^{Gaussian}(x_i)} = \frac{(2\pi)^{-n/2}\,|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}x^{\mathrm{T}}\Sigma^{-1}x\right)}{\prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(-\frac{1}{2}x_i^2\right)} =$$
$$= \frac{1}{|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}\zeta^{\mathrm{T}}(\Sigma^{-1}-\mathbf{I})\zeta\right),$$

где $\zeta = (\Phi^{-1}(u_1),...,\Phi^{-1}(u_n))^{\mathrm{T}}$ — вектор, компонентами которого являются значения обратной функции для стандартного одномерного гауссовского распределения в точках $u_i = \Phi(x_i)$, $i = 1,...,n$; $\mathbf{I}$ — единичная матрица.

Представим основные шаги алгоритма моделирования нормальной копула-функции (более детальное описание моделирования копула-функции см. в (Cherubini et al., 2004; McNeil et al., 2005)):

- найти разложение Холецкого $\mathbf{A}$ для корреляционной матрицы Σ;
- смоделировать из стандартного нормального распределения *n* независимых случайных величин $\mathbf{z} = (z_1,\ldots,z_n)^T$;
- взять $\mathbf{x} = \mathbf{A}\mathbf{z}$;
- вычислить компоненты $u_i = \phi(x_i)$, $i = 1,\ldots,n$, где $\phi(\cdot)$ — одномерное стандартное нормальное распределение.

В результате получим вектор $(u_1,\ldots,u_n)^{\mathrm{T}}$, являющийся реализацией случайного вектора из *n*-мерной гауссовской копула-функции.

2.3. *T* копула-функция (копула Стьюдента)

Копула-функция для двумерного распределения Стьюдента с корреляцией ρ может быть получена из формулы (5) при

$$g = \left(1+\frac{2t}{\nu}\right)^{\frac{-\nu-2}{2}}, \quad \gamma = \frac{1}{2\pi}.$$

Для *T* копула-функции справедливо соотношение:

$$C^{t-copula}(u_1,u_2;\rho,\nu) = T^2(t_\nu^{-1}(u_1),t_\nu^{-1}(u_2);\rho,\nu),$$

где $T^2(\cdot,\cdot;\rho,\nu)$ — функция распределения двумерного распределения Стьюдента с корреляцией ρ и ν степенями свободы; $t_\nu^{-1}(\cdot)$ — обратная функция одномерного распределения Стьюдента. При $\nu \to \infty$ копула-функция Стьюдента сходится к нормальной копула-функции, а при $\nu \to 1$ сходится к копула-функции распределения Коши.

В общем случае многомерная *T* копула-функция определяется следующим образом:

$$C(u_1,u_2,\ldots,u_n;\Sigma,\nu) = T^n(t^{-1}(u_1),t^{-1}(u_2),\ldots,t^{-1}(u_n);\Sigma,\nu),$$

а плотность *T* копула-функции может быть получена из уравнения (4) и канонического представления:

$$c(t_\nu(x_1),\ldots,t_\nu(x_n)) = |\Sigma|^{-1/2}\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left[\frac{\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)}\right]^n \frac{\left(1+\frac{\zeta^{\mathrm{T}}\Sigma^{-1}\zeta}{\nu}\right)^{-\frac{\nu+n}{2}}}{\prod_{i=1}^{n}\left(1+\frac{\zeta_i^2}{2}\right)^{-\frac{\nu+1}{2}}},$$

где $\zeta = (t_v^{-1}(u_1),...,t_v^{-1}(u_n))^{\mathrm{T}}$ — вектор, компонентами которого являются значения обратной функции распределения Стьюдента в точках $u_i = t_\nu(x_i)$.

Приведем алгоритм моделирования наблюдений, подчиняющихся распределению с Т копула-функцией:

- найти разложение Холецкого $\mathbf{A}$ матрицы Σ;
- смоделировать n независимых случайных величин из стандартного нормального распределения, $\mathbf{z} = (z_1,\ldots,z_n)^{\mathrm{T}}$;
- смоделировать случайную величину s из χ_ν-распределения, не зависящую от $\mathbf{z}$;
- определить вектор $\mathbf{y} = \mathbf{A}\mathbf{z}$;
- положить $\mathbf{x} = \frac{\sqrt{\nu}}{\sqrt{s}}\mathbf{y}$;
- определить компоненты $u_i = t_\nu(x_i),\ i = 1,\ldots,n$, где $t_\nu(\cdot)$ — одномерное стандартное распределение Стьюдента с ν степенями свободы.

Вектор $(u_1,\ldots,u_n)^{\mathrm{T}}$ смоделирован из n-мерной T копула-функции.

Нормальная и T копула-функции весьма популярны при анализе финансовых данных, поскольку они позволяют моделировать портфели высокой размерности, являясь при этом относительно легко оцениваемыми. Кроме того, они позволяют генерировать наблюдения при помощи относительно простых алгоритмов. На рисунке 3 представлены графики плотностей некоторых эллиптических копула-функций, построенных с использованием функции `fcopulae` из пакета *R*.

Обобщение: сгруппированные *T* копула-функции. Сгруппированные T копула-функции впервые были предложены в работах (Daul et al., 2003) и (Demarta, McNeil, 2005). Калибровка таких копула-функций не сложнее, чем калибровка T копула-функций, но для первых имеется возможность определять различные структуры зависимости для подгрупп компонент вектора из T копула-функции.

Пусть $\mathbf{Z} \in \mathbf{N}_n(0,\mathbf{R})$, где $\mathbf{R}$ — ковариационная матрица размерности $n \times n$ с единичными диагональными элементами. Далее рассмотрим случайную величину U, равномерно распределенную на отрезке [0,1] и не зависящую от $\mathbf{Z}$. Пусть через $G_\nu(\cdot)$ обозначена функция распределения случайной величины $\sqrt{\nu/\chi_\nu^2}$, где χ_ν^2 — случайная величина, имеющая хи-квадрат распределение с ν степенями свободы. Кроме того, задано разбиение множества $1,\ldots,n$ на m подмножеств с размерами $s_1,\ldots,s_m$ (т.е. $s_1,\ldots,s_m \geq 1,\ s_1+\ldots+s_m = n$) и для каждого $k = 1,2,\ldots,m$ задано число степеней свободы ν_k. Возьмем $W_k = G_{\nu_k}^{-1}(U)$, $k = 1,\ldots,m$, и $\mathbf{Y} = (W_1Z_1,\ldots,W_1Z_{s_1},W_2Z_{s_1+1},\ldots,W_2Z_{s_1+s_2},\ldots,W_mZ_n)$. Случайная величина $\mathbf{Y}$ имеет так называемое сгруппированное t-распределение. Наконец, определим

$$\mathbf{U} = (t_{\nu_1}(Y_1),\ldots,t_{\nu_1}(Y_{s_1}),t_{\nu_2}(Y_{s_1+1}),\ldots,t_{\nu_2}(Y_{s_1+s_2}),\ldots,t_{\nu_m}(Y_n)).$$

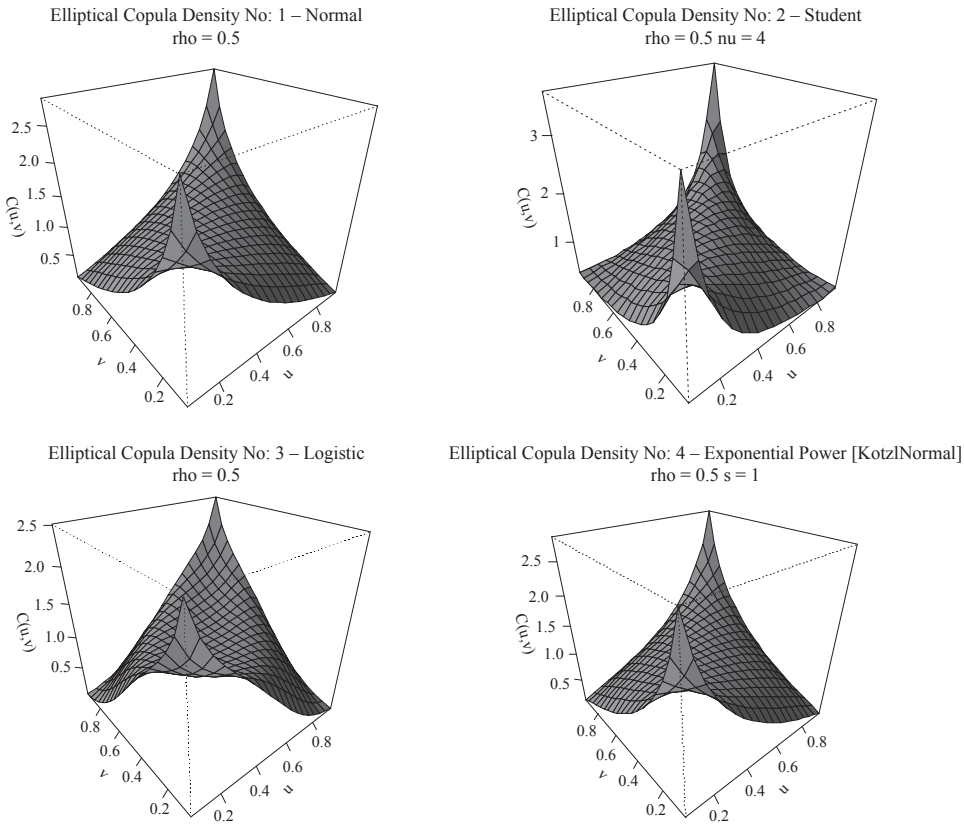


Рис. 3. Плотности эллиптических копула-функций: нормальное распределение, распределение Стьюдента, логистическое и экспоненциальное распределения. Параметры указаны над соответствующими им графиками

Случайный вектор $\mathbf{U}$ имеет распределение на $[0,1]^n$, а его компоненты имеют равномерное распределение на [0,1]. Будем называть распределение вектора $\mathbf{U}$ сгруппированной *T* копула-функцией.

Отметим, что $(Y_1,\ldots,Y_{s_1})$ имеет многомерное *T*-распределение с ν_1 степенями свободы, и для $k=1,\ldots,m-1$ вектор $(Y_{s_1+\ldots+s_k+1},\ldots,Y_{s_1+\ldots+s_{k+1}})$ имеет многомерное *T*-распределение с ν_{k+1} степенями свободы. Соответствующее распределение подвектора вектора $\mathbf{U}$ представляет собой *T* копула-функцию с ν_{k+1} степенями свободы, для $k=0,\ldots,m-1$. Для сгруппированной *T* копула-функции нет простого выражения для плотности. Однако существует весьма полезная аппроксимация для автокорреляции (в предположении ее постоянства), полученная в работе (Daul et al., 2003):

$$\rho_{ij}(z_i,z_j)\approx\sin(\pi\tau_{ij}(u_i,u_j)/2)\,,$$

где i и j принадлежат различным группам, а τ_{ij} — парный коэффициент ранговой корреляции Кендалла. Эта аппроксимация позволяет производить оценку параметров методом максимального правдоподобия независимо для каждой подгруппы. В работе (Fantazzini,

2009a) представлено обобщение модели из (Daul et al., 2003), в котором рассматривается динамическая структура корреляционной матрицы $\mathbf{R}_t$.

2.4. Численные примеры: оценивание эллиптических копула-функций

Нормальная копула-функция. Выше было показано, как, применяя теорему Склара и используя соотношение между функцией распределения и ее плотностью, можно получить плотность нормальной (гауссовской) копула-функции:

$$c(u_1,u_2,\ldots,u_n;\Sigma)=\frac{f^{Gaussian}(x_1,\ldots,x_n)}{\prod_{i=1}^{n} f_i^{Gaussian}(x_i)}=\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}\zeta_t^{\mathrm{T}}(\Sigma^{-1}-\mathbf{I})\zeta_t\right],$$

где $\zeta=(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_n))^{\mathrm{T}}$, а $u_i=\Phi(x_i)$; логарифмическая функция правдоподобия имеет вид:

$$l^{Gaussian}(\theta)=-\frac{T}{2}\ln|\Sigma|-\frac{1}{2}\sum_{t=1}^{T}\zeta_t^{\mathrm{T}}(\Sigma^{-1}-\mathbf{I})\zeta_t.$$

Если логарифмическая функция правдоподобия дифференцируема по параметрам θ, а решение уравнений $\partial_\theta l(\theta)=0$ дает точку *глобального* максимума функции $l(\theta)$, можно получить оценку максимального правдоподобия для ковариационной матрицы $\hat{\theta}_{ML}=\hat{\Sigma}$ гауссовской копула-функции:

$$\frac{\partial}{\partial\Sigma^{-1}}\,l^{Gaussian}(\theta)=\frac{T}{2}\Sigma-\frac{1}{2}\sum_{t=1}^{T}\zeta_t\zeta_t^{\mathrm{T}}=0,$$

а, следовательно,

$$\hat{\Sigma}=\frac{1}{T}\sum_{t=1}^{T}\zeta_t\zeta_t^{\mathrm{T}},$$

т. е. получаем классическую оценку для ковариационной матрицы. Следовательно, в качестве приближенной оценки корреляционной матрицы можно использовать матрицу:

$$\widehat{\mathbf{R}}=(\mathrm{diag}\ \hat{\Sigma})^{-1/2}\ \hat{\Sigma}\ (\mathrm{diag}\ \hat{\Sigma})^{-1/2}. \qquad (6)$$

***T* копула-функция.** Плотность *T* копула-функции равна:

$$c(u_1,u_2,\ldots,u_n;\Sigma)=\frac{f^{Student}(x_1,\ldots,x_n)}{\prod_{i=1}^{n} f_i^{Student}(x_i)}=\frac{1}{|\Sigma|^{\frac{1}{2}}}\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left[\frac{\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)}\right]^{n}\frac{\left(1+\frac{\zeta_t^{\mathrm{T}}\Sigma^{-1}\zeta_t}{\nu}\right)^{-\frac{\nu+n}{2}}}{\prod_{i=1}^{n}\left(1+\frac{\zeta_{it}^2}{\nu}\right)^{-\frac{\nu+1}{2}}},$$

а логарифмическая функция правдоподобия имеет вид

$$l^{Student}(\theta)=-T\ln\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}-nT\ln\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}-\frac{T}{2}\ln|\Sigma|-\frac{\nu+n}{2}\sum_{t=1}^{T}\ln\left(1+\frac{\zeta_t^{\mathrm{T}}\Sigma^{-1}\zeta_t}{\nu}\right)+\frac{\nu+1}{2}\sum_{t=1}^{T}\sum_{i=1}^{n}\ln\left(1+\frac{\zeta_{it}^2}{\nu}\right).$$

В случае T копула-функции невозможно получить аналитическое выражение для оценки максимального правдоподобия, поэтому необходимо использовать численные алгоритмы максимизации функции правдоподобия. Однако в случае большой размерности T копула-функции такие алгоритмы являются слишком затратными по времени. В связи с этим предлагаются альтернативные алгоритмы, многошаговые параметрические или полупараметрические подходы. Рассмотрим три наиболее часто применяемых метода: первый из них предложен в работе (Bouyé et al., 2000) и основан на процедуре рекурсивной оптимизации для корреляционной матрицы; второй представлен в работе (Marshal, Zeevi, 2002) и базируется на использовании оценки ранговой корреляции Кендалла; третий является смешанным параметрическим методом и основан на методах моментов и максимального правдоподобия (Chen et al., 2004).

Метод 1 (Bouyé et al., 2000).

1) Используя эмпирические функции одномерных частных распределений, преобразуем $(x_{1t},x_{2t},\ldots,x_{nt})$, $t=1,\ldots,T$, к приблизительно равномерно распределенным $(\hat{u}_{1t},\hat{u}_{2t},\ldots,\hat{u}_{nt})$.

2) *Для каждого значения степени свободы* ν из рассматриваемого интервала значений оцениваем корреляционную матрицу $\mathbf{R}_\nu$:

а) для каждого фиксированного t ($t=1,\ldots,T$) положим:

$$\zeta_t=(t_\nu^{-1}(\hat{u}_{1t}),\ldots,t_\nu^{-1}(\hat{u}_{nt}))^{\mathrm{T}};$$

б) пусть $\hat{\Sigma}$ — корреляционная матрица для гауссовской копула-функции, оцениваемая по формуле (6), положим $\hat{\Sigma}_{\nu,1}=\hat{\Sigma}$;

в) пусть $\hat{\Sigma}_{\nu,k+1}$ удовлетворяет следующей рекурсивной схеме:

$$\hat{\Sigma}_{\nu,k+1}=\frac{\nu+n}{T\nu}\sum_{t=1}^{T}\zeta_t\zeta_t^{\mathrm{T}}\left(1+\zeta_t\Sigma_{\nu,k}^{-1}\zeta_t^{\mathrm{T}}/\nu\right)^{-1};$$

г) перемасштабируем элементы матрицы следующим образом:

$$\left(\hat{\Sigma}_{\nu,k+1}\right)_{i,j}=\frac{\left(\hat{\Sigma}_{\nu,k+1}\right)_{i,j}}{\sqrt{\left(\hat{\Sigma}_{\nu,k+1}\right)_{i,i}\left(\hat{\Sigma}_{\nu,k+1}\right)_{j,j}}},$$

в результате получим, в частности, единичные диагональные элементы;

д) повторяем шаги в)–г) до тех пор, пока $\hat{\Sigma}_{\nu,k+1}=\hat{\Sigma}_{\nu,k}$, и положим $\hat{\Sigma}_\nu=\hat{\Sigma}_{\nu,k}$.

3) Оценим ν путем максимизации логарифмической функции правдоподобия плотности T копула-функции:

$$\hat{\nu} = \arg\max_{\nu} \sum_{t=1}^{T} \ln c^{Student}(\hat{u}_{1,t},\ldots,\hat{u}_{n,t};\hat{\Sigma}_{\nu},\nu).$$

Метод 2 (Marshall, Zeevi, 2002).

Предыдущая процедура является вычислительно трудоемкой в случае большой обучающей выборки. Более того, она может давать неустойчивые результаты в ситуациях, когда существует k $(k \geq 1)$, для которого матрица $\hat{\Sigma}_{\nu,k}$ близка к вырожденной. Для преодоления этой проблемы в работе (Marshall, Zeevi, 2002) предлагается следующий алгоритм:

1) Используя эмпирические функции одномерных частных распределений, преобразуем $(x_{1t}, x_{2t},\ldots,x_{nt})$, $t = 1,\ldots,T$ к приблизительно равномерно распределенным $(\hat{u}_{1t},\hat{u}_{2t},\ldots,\hat{u}_{nt})$.

2) Оценим $\hat{\Sigma}$ с использованием непараметрической оценки рангового коэффициента Кендалла (Lindskog et al., 2002):

$$\hat{\Sigma}_{ij} = \sin\left(\frac{\pi}{2}\hat{\tau}_{ij}\right), \quad i,j = 1,\ldots,n.$$

Как указано в работах (Lindskog, 2000; Genest et al., 1995), если имеется два временных ряда X_t и Y_t, $t = 1,\ldots,T$, то состоятельной оценкой коэффициента Кендалла будет:

$$\hat{\tau}_{i,j} = \frac{2}{T(T-1)} \sum_{i<j} \operatorname{sign}[(X_i - X_j)(Y - Y_j)], \quad i,j = 1,\ldots,n,$$

$$\text{где } \operatorname{sign}(x) = \begin{cases} 1, & \text{если } x \geq 0 \\ -1, & \text{если } x < 0. \end{cases}$$

3) Оценим ν путем максимизации логарифмической функции правдоподобия плотности T копула-функции:

$$\hat{\nu} = \operatorname{argmax} \sum_{t=1}^{T} \ln c^{Student}(\hat{u}_{1t},\ldots,\hat{u}_{nt};\hat{\Sigma};\nu).$$

Метод 3 (Chen et al., 2004).

1) Используя эмпирические или параметрически оцененные функции одномерных частных распределений, преобразуем $(x_{1t}, x_{2t},\ldots,x_{nt})$, $t = 1,\ldots,T$ к приблизительно равномерно распределенным $(\hat{u}_{1t},\hat{u}_{2t},\ldots,\hat{u}_{nt})$.

2) Пусть $\widehat{\mathbf{R}}$ — корреляционная матрица для гауссовской копула-функции, оцененной с использованием уравнения (6). Положим $\hat{\Sigma}_{Gaussian} = \widehat{\mathbf{R}}$.

3) Оценим ν путем максимизации логарифмической функции правдоподобия плотности T копула-функции:

$$\hat{\nu} = \operatorname{argmax} \sum_{t=1}^{T} \ln c^{Student}(\hat{u}_{1t},\ldots,\hat{u}_{nt};\hat{\Sigma}_{Gaussian};\nu).$$

4) Пусть $\zeta_{vt} = (t_v^{-1}(\hat{u}_{1t}),\ldots,t_v^{-1}(\hat{u}_{nt}))^{\mathrm{T}}$. Наконец, используя уравнение (6), вычислим:

$$\hat{\Sigma}_{Student} = \frac{1}{T}\sum_{t=1}^{T}\zeta_{vt}\zeta_{vt}^{\mathrm{T}}.$$

Здесь также могла быть применена итеративная процедура, но авторы работы (Chen et al., 2004) не делали этого. Однако после первого шага разница между фактической ковариационной матрицей и ее оценкой весьма мала.

2.5. Эмпирические приложения в статистическом пакете *R*: эллиптические копула-функции

В модуль «fCopulae» входит множество интересных функций, позволяющих исследовать эллиптические функции. Приведем некоторые команды:

fcopulae функция	Описание
ellipticalList	список доступных эллиптических копула-функций
ellipticalParam	установка параметров по умолчанию
ellipticalRange	диапазон допустимых значений коэффициента корреляции
ellipticalCheck	проверка принадлежности коэффициента корреляции допустимому диапазону

Для функции распределения и плотностей имеются следующие команды:

fcopulae функция	Описание
rellipticalCopula	моделирование наблюдений из эллиптической копула-функции
pellipticalCopula	вычисление значения эллиптической копула-функции
dellipticalCopula	вычисление значения плотности эллиптической копула-функции
rellipticalSlider	график случайных величин, смоделированных из эллиптической копула-функции
pellipticalSlider	график функции распределения
dellipticalSlider	график функции плотности

Если необходимо изобразить график функции плотности и линий уровня нормальной копула-функции, нужно использовать следующие команды (результат см. на рис. 4):

```
# Generate the 2D Grid:
N = 50; x = (0:N)/N; uv = grid2d (x); u = uv$x; v = uv$y
# Compute the Normal Copula Density:
c. uv = dellipticalCopula (u, v, rho = 3/4, type = "norm", output = "list")
# Create a Perspective Plot:
persp (c. uv, theta = -40, phi = 30, ticktype = "detailed", col = "steelblue",
main = "Normal Copula Density", cex = 0.5)
# Create a Contour Plot:
contour (c. uv, nlevels = 20, main = "Normal Copula Density", cex = 0.5)
```

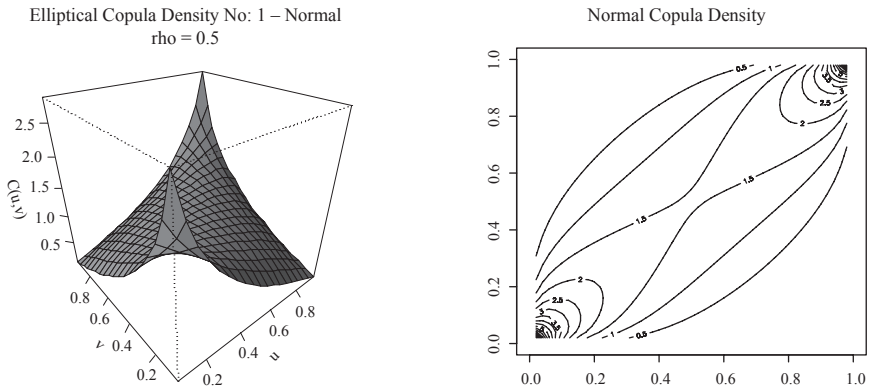


Рис. 4. График плотности нормальной копула-функции и ее линий уровня

Для представления о графике эллиптической копула-функции и ее плотности при различном выборе параметров рекомендуем воспользоваться «слайдером» функций:

```
# Start Slider for Perspective Distribution Plots:
pellipticalSlider ()
11cm!
# Start Slider for Perspective Density Plots:
dellipticalSlider ()
```

В случае, если необходимо оценить эллиптические копула-функции и сгенерировать соответствующие им наблюдения, можно использовать следующий код:

```
# Estimation of 100 Samples:
est = NULL
for (i in 1:100)
R = ellipticalCopulaSim (n = 100, rho = 0.6, param = 4, type = "t")
ans = ellipticalCopulaFit (R, type = "t")
ans=ans$par
est = rbind (est, ans)
# Print the Result:
for (i in 1:2) print (c ( mean = mean (est [, i]), sd = sd (est [, i])))
# Make a kernel density of the 100 estimates
dens1=density (est [,1]) plot (dens1,main="kernel density of estimated Rho")
dens2=density (est [,2]) plot (dens2,main="kernel density of estimated NU")
```

3. Архимедовы копула-функции

3.1. Основные понятия и определения

Архимедовы копула-функции обеспечивают аналитическую гибкость и широкий спектр различных мер зависимости. По следующим причинам эти копула-функции могут быть использованы в широком диапазоне приложений:

• *Архимедовы копула-функции могут быть представлены в явном аналитическом виде*, в отличие от семейства эллиптических копула-функций, которые определяются в неявной форме.

• *Архимедовы копула-функции допускают относительно простое построение, включая вычислительную реализацию* (см. табл. 2).

• *Многие параметрические семейства* копула-функций принадлежат этому классу.

• Архимедовы копула-функции *не ограничены обязательным наличием радиальной симметрии*, что свойственно случаю нормальной копула-функции или общему случаю эллиптических копула-функций. Это является преимуществом, т. к. во многих финансовых и страховых приложениях наблюдается более сильная зависимость между большими убытками, чем между большими доходами.

Двумерные архимедовы копула-функции могут быть определены следующим образом.

Определение 3. Рассмотрим непрерывную, строго убывающую и выпуклую функцию $\varphi(u)$ с неотрицательными значениями, определенную при $u \in [0,1]$ и удовлетворяющую условию $\varphi(1) = 0$. Определим псевдообратную функцию $\varphi^{[-1]}(t)$ соотношением:

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t) & \text{для } 0 \le t < \varphi(0); \\ 0 & \text{для } t \ge \varphi(0), \end{cases}$$

где $\varphi^{-1}(t)$ — обычная обратная функция к функции $\varphi(u)$.

Тогда функцию $C : [0,1]^2 \to [0,1]$, определенную как:

$$C(u_1, u_2) = \varphi^{[-1]}[\varphi(u_1) + \varphi(u_2)], \tag{7}$$

называют архимедовой копула-функцией с генератором $\varphi(\cdot)$. Более того, если $\varphi(0) = \infty$, то псевдообращение дает обычную обратную функцию (т. е. $\varphi^{[-1]}(\cdot) = \varphi^{-1}(\cdot)$), а $\varphi(\cdot)$ и $C(\cdot)$ в этом случае называют строгим генератором и строгой архимедовой копула-функцией, соответственно (в противном случае они называются нестрогими), см. рис. 5.

В работе (Nelsen, 1999) представлен список из 22 однопараметрических двумерных архимедовых копула-функций. Этот список вместе с генерирующими функциями $\varphi(\cdot)$ и соответствующими диапазонами параметра α воспроизведен в табл. 2.

Копула-функции с номерами 3 – 6, 9, 10, 12 – 14, 17, 19, 20, 22 являются строгими копула-функциями, тогда как копула-функции с номерами 1 и 16 — строгие только для $\alpha > 0$. Остальные копула-функции из табл. 2 не являются строгими. Некоторые из архимедовых копула-функций имеют специальные названия. Например, копула-функцию под номером 1 называют *копула-функцией Клейтона*, под номером 4 — *копула-функцией Гумбеля*,

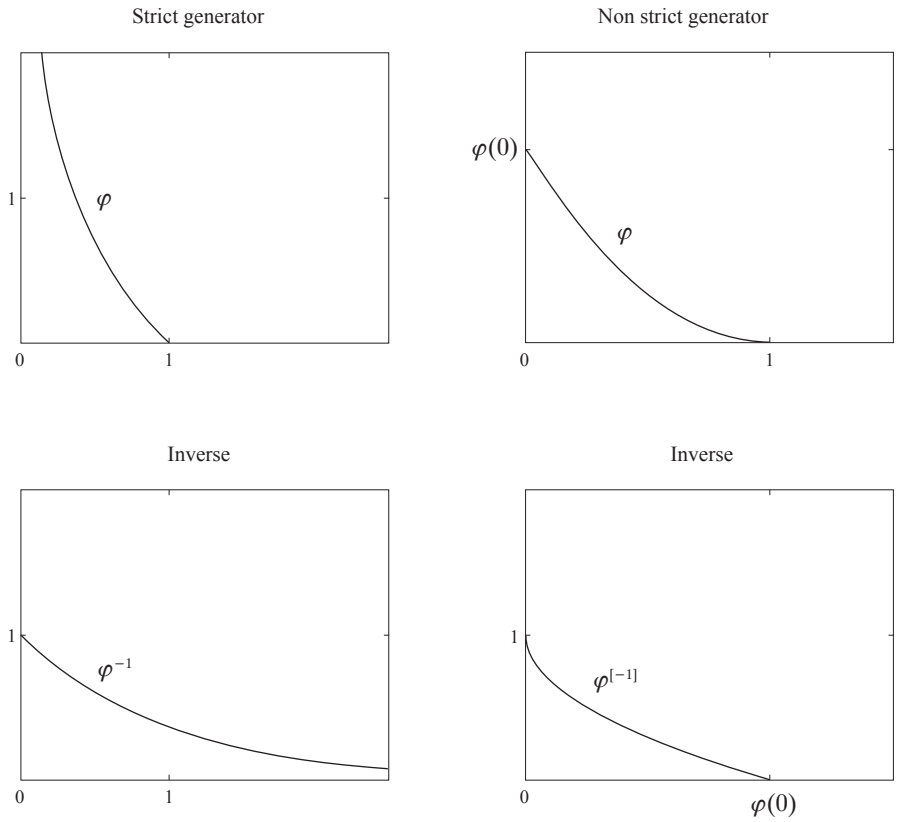


Рис. 5. Строгие и нестрогие генераторы и их обращения

а под номером 5 — *копула-функцией Франка*. Учитывая, что последние три копула-функции довольно популярны, они будут рассмотрены ниже более подробно.

Таблица 2. Список архимедовых копула-функций

Номер	$C_\alpha(u,v)$	Генератор $\varphi_\alpha(t)$	Диапазон параметра α
1	$\max([u^{-\alpha}+v^{-\alpha}-1]^{-1/\alpha},0)$	$(t^{-\alpha}-1)/\alpha$	$[-1,\infty)\setminus\{0\}$
2	$\max(1-[(1-u)^\alpha+(1-v)^\alpha]^{1/\alpha},0)$	$(1-t)^\alpha$	$[1,\infty)$
3	$\dfrac{uv}{1-\alpha(1-u)(1-v)}$	$\ln\dfrac{1-\alpha(1-t)}{t}$	$[-1,1)$
4	$\exp(-[(-\ln u)^\alpha+(-\ln v)^\alpha]^{1/\alpha})$	$(-\ln t)^\alpha$	$[1,\infty)$
5	$-\dfrac{1}{\alpha}\ln\left(1+\dfrac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{e^{-\alpha}-1}\right)$	$-\ln\dfrac{e^{-\alpha t}-1}{e^{-\alpha}-1}$	$(-\infty,\infty)\setminus\{0\}$
6	$1-[(1-u)^\alpha+(1-v)^\alpha-(1-u)^\alpha(1-v)^\alpha]^{1/\alpha}$	$-\ln[1-(1-t)^\alpha]$	$[1,\infty)$
7	$\max(\alpha uv+(1-\alpha)(u+v-1),0)$	$-\ln[\alpha t+(1-\alpha)]$	$(0,1]$
8	$\max\left(\dfrac{\alpha^2uv-(1-u)(1-v)}{\alpha^2-(\alpha-1)^2(1-u)(1-v)},0\right)$	$\dfrac{1-t}{1+(\alpha-1)t}$	$[1,\infty)$

Окончание табл. 2

Номер	$C_\alpha(u,v)$	Генератор $\varphi_\alpha(t)$	Диапазон параметра α
9	$uv\exp(-\alpha\ln u\ln v)$	$\ln(1-\alpha\ln t)$	$(0,1]$
10	$uv[1+(1-u^\alpha)(1-v^\alpha)]^{-1/\alpha}$	$\ln(2t^{-\alpha}-1)$	$(0,1]$
11	$\max([u^\alpha v^\alpha-2(1-u)^\alpha(1-v)^\alpha]^{1/\alpha},0)$	$\ln(2-t^\alpha)$	$(0,\frac{1}{2}]$
12	$(1+[(u^{-1}-1)^\alpha+(v^{-1}-1)^\alpha]^{1/\alpha})^{-1}$	$(t^{-1}-1)^\alpha$	$[1,\infty)$
13	$\exp(1-[(1-\ln u)^\alpha+(1-\ln v)^\alpha-1]^{1/\alpha})$	$(1-\ln t)^\alpha-1$	$(0,\infty)$
14	$(1+[(u^{-1/\alpha}-1)^\alpha+(v^{-1/\alpha}-1)^\alpha]^{1/\alpha})^{-\alpha}$	$(t^{-1/\alpha}-1)^\alpha$	$[1,\infty)$
15	$\max\left(\{1-[(1-u^{1/\alpha})^\alpha+(1-v^{1/\alpha})^\alpha]^{1/\alpha}\}^\alpha,0\right)$	$(1-t^{1/\alpha})^\alpha$	$[1,\infty)$
16	$\frac{1}{2}(S+\sqrt{S^2+4\alpha}),\ S=u+v-1-\alpha(u^{-1}+v^{-1}-1)$	$(\alpha t^{-1}+1)(1-t)$	$[0,\infty)$
17	$\left(1+\frac{[(1+u)^{-\alpha}-1][(1+v)^{-\alpha}-1]}{2^{-\alpha}-1}\right)^{-1/\alpha}-1$	$-\ln\frac{(1+t)^{-\alpha}-1}{2^{-\alpha}-1}$	$(-\infty,\infty)\setminus\{0\}$
18	$\max\left(1+\alpha/\ln[e^{\alpha/(u-1)}+e^{\alpha/(v-1)}],0\right)$	$e^{\alpha/(u-1)}$	$[2,\infty)$
19	$\alpha/\ln(e^{\alpha/u}+e^{\alpha/v}-e^\alpha)$	$e^{\alpha/t}-e^\alpha$	$[0,\infty)$
20	$[\ln(\exp(u^{-\alpha})+\exp(v^{-\alpha})-e)]^{-1/\alpha}$	$\exp(t^{-\alpha})-e$	$(0,\infty)$
21	$1-(1-\{\max([1-(1-u)^\alpha]^{1/\alpha}+$ $+[1-(1-v)^\alpha]^{1/\alpha}-1,0)\}^\alpha)^{1/\alpha}$	$1-[1-(1-t)^\alpha]^{1/\alpha}$	$[1,\infty)$
22	$\max\left(\left[1-(1-u^\alpha)\sqrt{1-(1-u^\alpha)^2}-\right.\right.$ $\left.\left.-(1-v^\alpha)\sqrt{1-(1-v^\alpha)^2}\right]^{1/\alpha},0\right)$	$\arcsin(1-t^\alpha)$	$(0,1]$

Плотность архимедовых копула-функций может быть вычислена по формуле:

$$c(u_1,u_2)=\frac{\partial^2 C(u_1,u_2)}{\partial u_1\partial u_2}$$

или, если генератор $\varphi(\cdot)$ дважды непрерывно дифференцируем,

$$c_\varphi(u_1,u_2)=\frac{\left(\varphi^{[-1]}[\varphi(u_1)+\varphi(u_2)]\right)''}{\left(\varphi^{[-1]}[\varphi(u_1)]\right)'\left(\varphi^{[-1]}[\varphi(u_2)]\right)'}.$$

Это означает, что, зная генератор $\varphi(\cdot)$, его обратную функцию $\varphi^{[-1]}(\cdot)$, ее первую и вторую производные, можно получить архимедову копула-функцию и ее плотность. Однако, за исключением некоторых простых случаев, плотности копула-функций оцениваются с использованием таких программ, как *Maple* или *Mathematica*.

3.2. Некоторые однопараметрические двумерные архимедовы копула-функции

Копула-функция Клейтона. Семейство копула-функций Клейтона было впервые предложено в работе (Clayton, 1978). Рассмотрим генератор $\varphi(t) = (t^{-\alpha} - 1)/\alpha$ с $\alpha \in [-1, \infty) \setminus \{0\}$. Тогда $\varphi^{-1}(t) = (1+t)^{-1/\alpha}$. Используя (7), получим:

$$C^{Clayton}(u_1, u_2; \alpha) = \max\left[\left(u_1^{-\alpha} + u_2^{-\alpha} - 1\right)^{-1/\alpha}, 0\right].$$

Если $\alpha > 0$, то $\varphi(0) = \infty$ и вышеприведенное выражение для копула-функции примет вид:

$$C(u_1, u_2) = \left(u_1^{-\alpha} + u_2^{-\alpha} - 1\right)^{-1/\alpha},$$

а плотность копула-функции Клейтона будет равна:

$$c(u_1, u_2) = (1+\alpha)(u_1 u_2)^{-\alpha-1}(u^{-\alpha} + v^{-\alpha} - 1)^{-\frac{1}{\alpha}-2}.$$

Отметим, что в отличие от случая $\alpha > 0$, при $-1 \le \alpha < 0$ копула-функция Клейтона не является строгой, поскольку $\varphi(0) = 1/\alpha$. При $\alpha \to 0$ получим копула-фунцию, отвечающую случаю независимости, а при $\alpha \to \infty$ — копула-функцию, соответствующую комонотонности. При $\alpha = -1$ получим нижнюю границу Фреше–Хёффдинга. Таким образом, копула-функция позволяет интерполировать промежуточные структуры зависимости между противомонотонностью, комонотонностью и независимостью.

Копула-функция Гумбеля. Семейство копула-функций Гумбеля было впервые представлено в работе (Gumbel, 1960), но поскольку подробный анализ этого семейства был приведен в работе (Hougaard, 1986), то его иногда называют семейством Гумбеля–Хоугарда. Генератор этого семейства имеет вид: $\varphi(t) = (-\ln t)^{\alpha}$, где $\alpha \ge 1$. Обратная функция генератора равна $\varphi^{-1}(t) = \exp(-t^{1/\alpha})$. Таким образом, копула-функции из этого семейства имеют вид:

$$C(u_1, u_2; \alpha) = \exp\left\{-\left[(-\ln u_1)^{\alpha} + (-\ln u_2)^{\alpha}\right]^{1/\alpha}\right\}$$

с функцией плотности:

$$c(u_1, u_2; \alpha) = C(u_1, u_2) \cdot u_1^{-1} u_2^{-1}\left[(-\ln u_1)^{\alpha} + (-\ln u_2)^{\alpha}\right]^{-2+1/\alpha} \times$$
$$\times [\ln u_1 \ln u_2]^{\alpha-1} \cdot \{[(-\ln u_1)^{\alpha} + (-\ln u_2)^{\alpha}]^{-1/\alpha} + \alpha - 1\}.$$

При $\alpha = 1$ получаем копула-функцию, соответствующую случаю независимости; а при $\alpha \to \infty$ копула-функция Гумбеля $C(u_1, u_2; \alpha)$ стремится к копула-функции, отвечающей комонотонности. Таким образом, копула-функция Гумбеля интерполирует структуру зависимости, промежуточную между независимостью и абсолютной положительной зависимостью.

Копула-функция Франка. Впервые копула-функция Франка была представлена в работе (Frank, 1979), а более подробный ее анализ проведен в работе (Genest, 1987). Эта копула-функция имеет генератор:

$$\varphi(t) = \ln\left(\frac{e^{-\alpha}-1}{e^{-\alpha t}-1}\right).$$

Обратная функция этого генератора равна:

$$\varphi^{-1}(t) = -\frac{1}{\alpha}\ln(1+e^{t}(e^{-\alpha}-1)),$$

а сама копула-функция и ее плотность имеют вид:

$$C(u_1,u_2;\alpha) = -\frac{1}{\alpha}\ln\left(1+\frac{(e^{-\alpha u_1}-1)\cdot(e^{-\alpha u_2}-1)}{e^{-\alpha}-1}\right),$$

$$c(u_1,u_2;\alpha) = \frac{-\alpha[e^{-\alpha}-1]e^{-\alpha(u_1+u_2)}}{\{[e^{-\alpha u_1}-1][e^{-\alpha u_2}-1]+e^{-\alpha}-1\}^2}$$

для $\alpha \neq 0$.

При $\alpha = 0$ копула-функция Франка имеет вид копула-функции, отвечающей случаю независимости; а при $\alpha \to -\infty$ и $\alpha \to +\infty$ копула-функция Франка принимает вид, соответственно, нижней и верхней границ Фреше–Хёффдинга.

Для того чтобы продемонстрировать различия между копула-функциями Клейтона, Гумбеля и Франка, на рис. 6 они представлены графически. Видно, что изображенные копула-функции имеют различный характер поведения в окрестностях точек (0,0) и (1,1). Например, копула-функция Гумбеля в окрестности (1,1) демонстрирует резкий рост, тогда как рост в окрестности точки (0,0) менее резок. Как будет показано далее, хорошим измерителем характера хвостов этих копула-функций является коэффициент «хвостовой зависимости». Копула-функция Гумбеля имеет «верхнюю хвостовую зависимость», копула-функция Клейтона — «нижнюю хвостовую зависимость», а копула-функция Франка не имеет хвостовой зависимости.

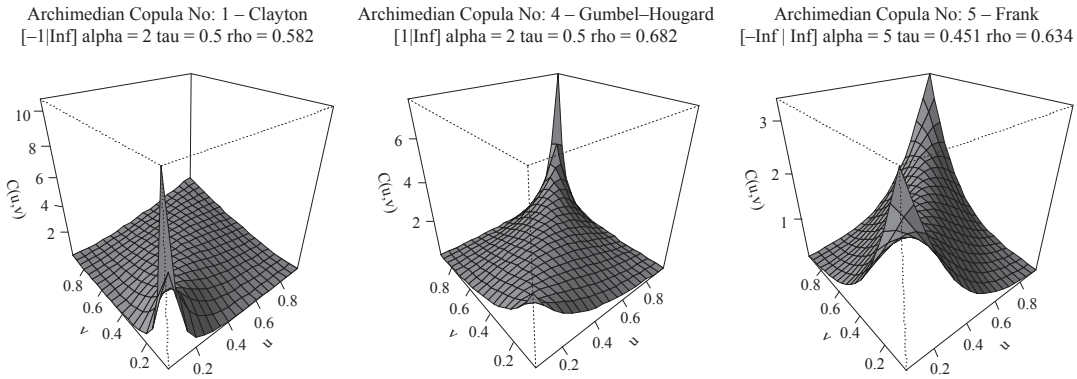


Рис. 6. Плотности копула-функций Клейтона, Гумбеля и Франка

3.3. Многомерные перестановочные архимедовы копула-функции

В работе (Nelsen, 2006) показано, что n-мерную копула-функцию для случая независимости можно записать в следующем виде:

$$\mathbf{P}(u_1,u_2,\ldots,u_n)=u_1\cdot u_2\cdot\ldots\cdot u_n=\exp(-[(-\ln u_1)+(-\ln u_2)+\ldots+(-\ln u_n)]).$$

Это наблюдение подсказывает возможность обобщения (7) на n-мерный случай:

$$C(u_1,u_2,\ldots,u_n)=\varphi^{[-1]}[\varphi(u_1)+\varphi(u_2)+\ldots+\varphi(u_n)]. \qquad (7a)$$

Функции $C(\cdot)$ в (7a) называют *серийно-итеративными* функциями (Schweizer, Sklar, 1983), основанными на двумерной архимедовой копула-функции с генератором $\varphi(\cdot)$. Если положить $C(u_1,u_2)=\varphi^{[-1]}[\varphi(u_1)+\varphi(u_2)]$, то для $n\geq 3$ получим следующее обобщение:

$$C(u_1,u_2,\ldots,u_n)=C(C(u_1,u_2,\ldots,u_{n-1}),u_n).$$

Двумерные архимедовы копула-функции обладают свойствами симметричности и ассоциативности (см. теорему 4.1.5 в (Nelsen, 2006)), а именно:

$$C(u,v)=C(v,u)\quad \forall u,v\in[0,1]\quad \text{(симметричность)},$$
$$C\big(C(u,v),w\big)=C\big(u,C(v,w)\big)\quad \forall u,v,w\in[0,1]\quad \text{(ассоциативность)}.$$

Копула-функции вида (7a) также известны как *перестановочные архимедовы копула-функции*: случайные величины X и Y называются *перестановочными*, если векторы (X,Y) и (Y,X) являются одинаково распределенными. В нашем случае перестановочность копула-функции эквивалентна ее симметричности (см. теорему 2.7.4 в (Nelsen, 2006)).

Для того чтобы $C(\cdot)$ была копула-функцией при $n\geq 3$, требуются некоторые дополнительные свойства функций $\varphi(\cdot)$ и $\varphi^{[-1]}(\cdot)$. Для этого понадобится следующее определение.

Определение 4. Неотрицательная функция $g(t)$ называется абсолютно монотонной на интервале J, если она непрерывна на этом интервале и имеет производные любого порядка, причем

$$(-1)^k\frac{d^k}{dt^k}g(t)\geq 0$$

для всех внутренних точек $t\in J$ и $k=1,2,\ldots$.

Например, функция $g(t)=e^{-t}$, $t\in R$ является абсолютно монотонной, т. к. $g'(t)=-e^{-t}$, $g''(t)=e^{-t},\ldots,g^{(k)}(t)=(-1)^k e^{-t},\ldots$.

Следует отметить, что если $g(t)$ абсолютно монотонна на $[0,\infty)$ и $g(c)=0$ для некоторого $c>0$, то $g(\cdot)\equiv 0$ на $[0,\infty)$. Более того, если псевдообратная функция $\varphi^{[-1]}(\cdot)$ генератора

архимедовой копула-функции $\varphi(\cdot)$ абсолютно монотонна, то на интервале $[0,\infty)$ функция $\varphi^{[-1]}(\cdot)$ положительна, функция $\varphi(\cdot)$ абсолютно монотонна и $\varphi^{[-1]}(\cdot)=\varphi^{-1}(\cdot)$.

Ниже приведем необходимое и достаточное условие, налагаемое на строгий генератор $\varphi(\cdot)$, для того чтобы функция (7а) была *n*-мерной копула-функцией при $n \geq 2$ (Kimberling, 1974).

Теорема 3. *Пусть* $\varphi:[0,1]\to[0,\infty]$ *— непрерывная и строго убывающая функция, такая, что* $\varphi(0)=\infty$ *и* $\varphi(1)=0$, *а* $\varphi^{-1}(\cdot)$ *— обратная функция для* $\varphi(\cdot)$, $C:[0,1]^n \to [0,1]$ *определяется соотношением* (*7а*). *Тогда необходимым и достаточным условием для того, чтобы* $C(\cdot)$ *была n-мерной копула-функцией для всех* $n \geq 2$, *является требование абсолютной монотонности функции* $\varphi^{-1}(\cdot)$ *на* $[0,\infty)$.

Доказательство. См. (Kimberling, 1974; Schweizer, Sklar, 1983; Alsina et al., 2005).

Однако требование абсолютной монотонности в вышеприведенном условии является весьма жестким, что приводит к ограничениям структуры зависимости. В работах (Genest, Rivest, 1993; Nelsen, 2005; Müller, Scarsini, 2005) было предложено заменить это условие более слабым, в котором для функции $\varphi(\cdot)$ требуется существование производных лишь до некоторого *конечного* порядка. Недавно в (McNeil, Nešlehová, 2009) представлено необходимое и достаточное условие для того, чтобы функция $\varphi(\cdot)$ обладала свойством *n*-монотонности (*d*-монотонности в терминах этой работы). Это условие показывает, что существуют *n*-мерные архимедовы копула-функции, не имеющие плотностей. Кроме того, из результатов работы следует *существование точных нижних границ* для множества всех *n*-мерных архимедовых копула-функций относительно так называемого конкордационного упорядочивания (для более детальной информации см. (McNeil, Nešlehová, 2009)).

Представленное выше многомерное обобщение архимедовых копула-функций является весьма ограничительным, поскольку в своей спецификации оно использует лишь один генератор, не зависящий от рассматриваемой размерности. Как следствие, все *k*-мерные частные распределения ($k<n$) имеют идентичный вид. Для того чтобы обеспечить большую гибкость, было предложено множество подходов, некоторые из которых мы рассмотрим ниже.

3.4. Вложенные архимедовы копула-функции

Построение полностью вложенных архимедовых копула-функций. Описание и исследование одного из обобщений для функций (7а) имеется в работах (Joe, 1997; Embrechts et al., 2003; Whelan, 2004; Savu, Trede, 2009; McNeil, 2008). Структура такого обобщения для четырехмерного случая показана на рис. 7.

Узлы u_1 и u_2 связываются при помощи копула-функции $C_1(\cdot)$, узел u_3 связывается с $C_1(u_1,u_2)$ при помощи копула-функции $C_2(\cdot)$ и, наконец, узел u_4 связывается с $C_2(u_3,C_1(u_1,u_2))$ при помощи копула-функции $C_3(\cdot)$. Таким образом, четырехмерный случай требует трех двумерных копула-функций $C_1(\cdot)$, $C_2(\cdot)$ и $C_3(\cdot)$, генераторы которых равны $\varphi_1(\cdot)$, $\varphi_2(\cdot)$ и $\varphi_3(\cdot)$ соответственно. Несмотря на то, что структура «связывания» довольно проста, ее выражение в виде аналитического уравнения является весьма непростой задачей:

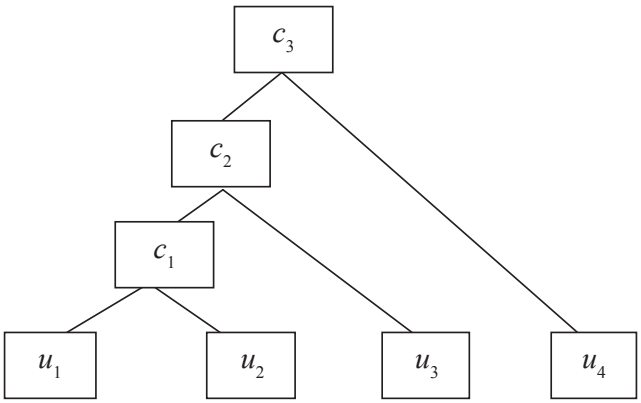


Рис. 7. Построение полностью вложенных архимедовых копула-функций

$$C(u_1,u_2,u_3,u_4) = C_3(u_4, C_2(u_3, C_1(u_1,u_2))) =$$
$$= \varphi_3^{-1}\{\varphi_3(u_4) + \varphi_3(\varphi_2^{-1}\{\varphi_2(u_3) + \varphi_2(\varphi_1^{-1}\{\varphi_1(u_1) + \varphi_1(u_2)\})\})\}.$$

Другими словами, пары (u_1,u_3) и (u_2,u_3) имеют одну и ту же копула-функцию $C_2(\cdot)$ с параметром зависимости α_2, тогда как пары (u_1,u_4), (u_2,u_4) и (u_3,u_4) имеют копула-функцию $C_3(\cdot)$ с параметром зависимости α_3. В общем *n*-мерном случае:

$$C(u_1,\ldots,u_n) = \varphi_{n-1}^{-1}\Big(\varphi_{n-1} \circ \varphi_{n-2}^{-1}[\ldots(\varphi_2 \circ \varphi_1^{-1}[\varphi_1(u_1) + \varphi_1(u_2)] + \varphi_2(u_3))$$
$$+ \ldots + \varphi_2(u_{n-1})] + \varphi_{n-1}(u_n)\Big). \qquad (8)$$

Эта копула-функция называется полностью вложенной архимедовой копула-функцией, поскольку получается путем пошагового увеличения размерности на единицу. Такая структура построения копула-функций является более общей, чем (7а). Кроме того, она обладает лишь частичной перестановочностью. Полностью вложенная архимедова копула-функция имеет $n(n-1)/2$ различных двумерных частных распределений, но только $n-1$ несовпадающих двумерных копула-функций со свободной спецификацией, остальные двумерные функции неявным образом определяются из структуры построения.

Выражение (8) будет *n*-мерной копула-функцией только тогда, когда, в дополнение к свойству абсолютной монотонности для обратных функций генераторов, «композитная» функция $\varphi_{i+1} \circ \varphi_i^{-1}$ будет иметь абсолютно монотонные первые производные для вложений всех уровней *i*.

Как оказывается, если все генераторы однотипны, степень зависимости, выражаемая параметром копула-функции, должна убывать с увеличением уровня вложения, т. е. $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_{n-1}$. Если же генераторы принадлежат разным семействам, то ограничения на параметры будут еще более жесткими, а список генераторов, которые могут быть связаны, не столь велик (более детальную информацию см., например, в (Whelan, 2004; Savu, Trede, 2009; McNeil, 2008)).

Построение частично вложенных архимедовых копула-функций. Построение частично вложенных архимедовых копула-функций — альтернативный способ обобщения многомерных копула-функций вида (7а). Эта структура связывания изначально была предложена в работе (Joe, 1997), а дальнейший подробный анализ представлен в (Whelan, 2004; McNeil et al., 2005; McNeil, 2008). Структура этого метода связывания представляет собой смесь структур, используемых при построении перестановочной и полностью вложенной копула-функций, и носит название *частично вложенной структуры*. Наименьшая размерность, для которой частично вложенная структура связывания дает копула-функции, равна четырем. Для нее:

$$C(u_1,u_2,u_3,u_4) = C_3(C_1(u_1,u_2),C_2(u_3,u_4)) = \\ = \varphi_3^{-1}\{\varphi_3(\varphi_1^{-1}\{\varphi_1(u_1)+\varphi_1(u_2)\})+\varphi_3(\varphi_2^{-1}\{\varphi_2(u_3)+\varphi_2(u_4)\})\}. \quad (9)$$

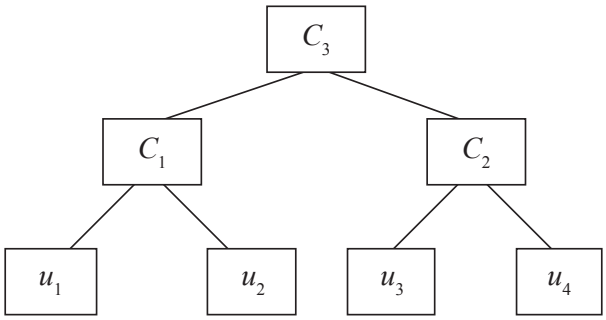


Рис. 8. Частично вложенные архимедовы копула-функции

Как и в предыдущих случаях, имеем громоздкие аналитические выражения, но весьма ясное описание: на первом шаге связываем пары (u_1,u_2) и (u_3,u_4) копула-функциями $C_1(\cdot)$ и $C_2(\cdot)$, генераторы которых равны φ_1 и φ_2, соответственно. Эти две копула-функции связываются при помощи третьей копула-функции $C_3(\cdot)$. Случайные величины u_1 и u_2 являются перестановочными, так же, как u_3 и u_4, но все другие пары перестановочными не являются. Однако пары (u_1,u_3), (u_1,u_4), (u_2,u_3) и (u_2,u_4) имеют копула-функцию $C_3(\cdot)$ (см. рис. 8). Для того чтобы получить *n*-мерную копула-функцию с использованием такого способа связывания, требуются ограничения, похожие на рассмотренные выше при построении полностью вложенных архимедовых копула-функций.

Общий случай: иерархически вложенные архимедовы копула-функции. Этот класс копула-функций впервые был предложен в работе (Joe, 1997), а в дальнейшем подробно изучался в (Whelan, 2004). В работе (Savu, Trede, 2009) представлена первая попытка разработки иерархически вложенных архимедовых копула-функций в общем виде. Такая структура связывания является расширением структуры частичного вложения в том смысле, что участвующие в процедуре связывания копула-функции не обязательно двумерные.

Основная идея структуры связывания, предложенная в (Savu, Trede, 2009), заключается в использовании иерархического вложения. При этом предполагается, что имеется L уровней, по n_l различных объектов на каждом уровне l (объектом может быть либо ко-

пула-функция, либо переменная). На уровне $l=1$ переменные $u_1,\ldots,u_n$ разбиваются на n_1 групп, в каждой из которых переменные связываются при помощи перестановочной архимедовой копула-функции. В свою очередь, на уровне $l=2$ эти копула-функции связываются при помощи n_2 копула-функций и т. д. Более формально, на уровне $l=1$ величины $u_1,\ldots,u_n$ связываются при помощи n_1 многомерных архимедовых копула-функций $C_{1,j}(\cdot)$, $j=1,\ldots,n_1$, имеющих вид:

$$C_{1,j}(\mathbf{u}_{1,j})=\varphi_{1,j}^{-1}\left(\sum_{\mathbf{u}_{1,j}}\varphi_{1,j}(\mathbf{u}_{1,j})\right),$$

где $\varphi_{1,j}(\cdot)$ — генератор копула-функции $C_{1,j}(\cdot)$; $\mathbf{u}_{1,j}$ $(j=1,\ldots,n_1)$ — подмножество элементов $u_1,\ldots,u_n$, связываемых при помощи $C_{1,j}(\cdot)$. Копула-функции $C_{1,1}(\cdot),\ldots,C_{1,n_1}(\cdot)$ могут принадлежать разным семействам архимедовых копула-функций, таким как семейство Франка или Гумбеля. Копула-функции уровня $l=1$, в свою очередь, агрегируются при помощи копула-функций уровня $l=2$. В результате получаем n_2 обобщенных архимедовых копула-функций $C_{2,j}(\cdot)$, $j=1,\ldots,n_2$, структура зависимости которых является частично перестановочной. Их компоненты — копула-функции предыдущего уровня. Более формально:

$$C_{2,j}(\mathbf{C}_{2,j})=\varphi_{2,j}^{-1}\left(\sum_{z\in\mathbf{C}_{2,j}}\varphi_{2,j}(\mathbf{C}_{z,j})\right),$$

где $j=1,\ldots,n_2$; $\varphi_{2,j}(\cdot)$ — генератор копула-функции $C_{2,j}(\cdot)$; $\mathbf{C}_{2,j}$ — множество всех копула-функций уровня $l=1$, агрегируемых копула-функцией $C_{2,j}(\cdot)$. Далее продолжаем шаг за шагом до тех пор, пока не достигнем уровня L, на котором получим единственную иерархическую архимедову копула-функцию $C_{L,1}(\cdot)$.

В работе (Savu, Trede, 2009) используются несколько отличные от используемых в данной статье, но эквивалентные, обозначения: поскольку *j*-ая копула-функция *l*-го ($l=1,\ldots,L$) уровня $C_{l,j}(\cdot)$ имеет в качестве аргументов $\mathbf{u}_{l,j}$, т. е. подмножество из $u_1,\ldots,u_n$, элементы которого в качестве аргументов входят явным или неявным образом в $C_{l,j}(\cdot)$, или (что эквивалентно) аргументы $\mathbf{C}_{l,j}(\cdot)$ (т. е. множество всех копула-функций уровня $l-1$, участвующих в $C_{l,j}(\cdot)$), то $C_{l,j}(\mathbf{C}_{l,j})$ и $C_{l,j}(\mathbf{u}_{l,j})$ суть одно и то же.

Для того чтобы иметь хорошо определенную иерархию, число используемых на каждом уровне копула-функций должно убывать с увеличением уровня, т. е. $n_l<n_{l-1}$ для всех $l=2,\ldots,L$, а на самом верхнем уровне должен использоваться лишь один агрегирующий объект $C_{L,1}(\cdot)$, т. е. $n_L=1$. Кроме того, размерность используемых копула-функций на каждом следующем уровне должна увеличиваться. На самом верхнем уровне размерность $C_{L,1}(\cdot)$ равна размерности исходных данных *n*, т. е. $n_{L,1}=n$.

Пример 3. Четырехмерные частично вложенные архимедовы копула-функции, описанные в (9) (см. рис. 8), имеют вид:

$$\begin{aligned}C_{2,1}(\mathbf{u})&=C_{2,1}(u_1,u_2,u_3,u_4)=C_{2,1}(C_{1,1}(u_1,u_2),C_{1,2}(u_3,u_4))=\\&=\varphi_{2,1}^{-1}\left(\varphi_{2,1}\circ\varphi_{1,1}^{-1}[\varphi_{1,1}(u_1)+\varphi_{1,1}(u_2)]+\varphi_{2,1}\left(\varphi_{1,2}^{-1}[\varphi_{1,2}(u_3)+\varphi_{1,2}(u_4)]\right)\right).\end{aligned}$$

Пример 4. Рассмотрим 9-мерную структуру, определяемую на рис. 9. Копула-функция для нее имеет вид:

$$C(u_1,\ldots,u_9) = C_{4,1}(C_{3,1}(C_{2,1}(C_{1,1}(u_1,u_2),u_3,u_4),u_5,u_6),C_{2,2}(u_7,C_{1,2}(u_8,u_9))).$$

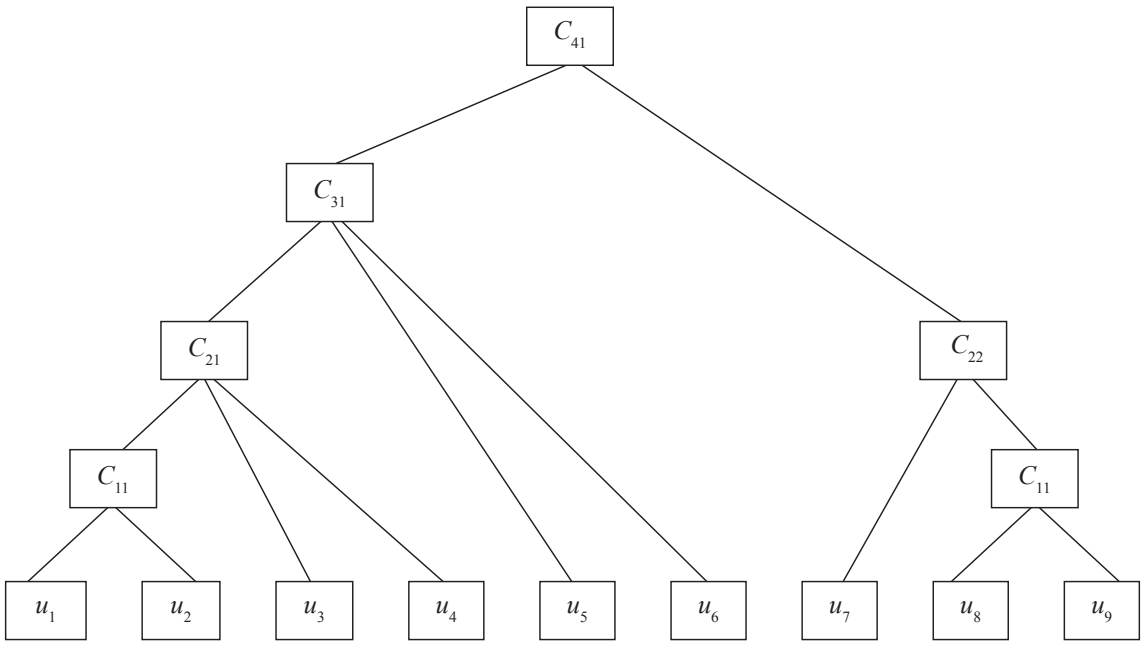


Рис. 9. Иерархически вложенная архимедова конструкция

В работе (Savu, Trede, 2009) указаны некоторые условия для того, чтобы иерархическое вложение давало многомерную функцию распределения. Во-первых, требуется, чтобы все обратные функции генераторов $\varphi_{l,j}^{-1}(\cdot)$ были абсолютно монотонными. Во-вторых, для $l=1,\ldots,L$ и $j=1,\ldots,n_l$, $i=1,\ldots,n_{l+1}$ сложные функции $\varphi_{l+1,i}\left(\varphi_{l,j}^{-1}(\cdot)\right)$ должны иметь *абсолютно монотонные производные* при всех $l=1,2,\ldots,L$ и $j=1,\ldots,n_l$. В работе (Embrechts et al., 2003) показано, что в случае полного вложения копула-функций либо из семейства Гумбеля, либо семейства Клейтона степень зависимости, выраженная параметром α, должна убывать с увеличением уровня иерархических вложений. В примере 4 это означает, что $\alpha_{1,1} \geq \alpha_{2,1} \geq \alpha_{3,1} \geq \alpha_{4,1}$ и $\alpha_{1,2} \geq \alpha_{2,2} \geq \alpha_{4,1}$.

Если же агрегируемые копула-функции принадлежат разным семействам архимедовых копула-функций, то, как показано в (McNeil, 2008), две архимедовы копула-функции из двух различных семейств a и b могут быть иерархически вложены только в том случае, если производная «композитной» функции $\varphi_a \circ \varphi_b^{-1}$ является абсолютно монотонной. Вопрос о том, какие семейства копула-функций могут быть агрегированы, в некоторой степени рассмотрен в работе (Joe, 1997).

В силу достаточно сложной структуры иерархических копула-функций, их плотности задаются весьма непростыми выражениями. В (Savu, Trede, 2009) используется рекурсивный подход для дифференцирования n-мерной копула-функции $C_{L,1}(\cdot)$ верхнего уровня по аргументам $\mathbf{u}_{L,1}$. А именно, в силу того, что

$$C_{L,1}(\mathbf{u}_{L,1}) = C_{L,1}(\mathbf{C}_{L,1}) = C_{L,1}(C_{L-1,1},\ldots,C_{L-1,n_{L-1}}) = C_{L,1}(C_{L-1,1}(\mathbf{C}_{L-1,1}),\ldots,C_{L-1,n_{L-1}}(\mathbf{C}_{L-1,n_{L-1}})) =$$
$$= C_{L,1}(C_{L-1,1}(\mathbf{u}_{L-1,1}),\ldots,C_{L-1,n_{L-1}}(\mathbf{u}_{L-1,n_{L-1}})),$$

плотность для $C_{L,1}(\mathbf{u}_{L,1})$ может быть вычислена следующим образом:

$$c_{L,1}(\mathbf{u}_{L,1}) = \frac{\partial^n C_{L,1}(\mathbf{u}_{L,1})}{\partial u_1 \ldots \partial u_n} = \sum \frac{\partial^{n-i} C_{L,1}(\mathbf{u}_{L,1})}{\partial C_{L-1,1}^{k_1} \ldots \partial C_{L-1,n_{L-1}}^{k_{n_{L-1}}}} \times$$
$$\times \prod_{r=1}^{n_{L-1}} \sum_{u=\{\sigma_1,\ldots,\sigma_r\}} \frac{\partial^{|\sigma_1|} C_{L-1,r}(\mathbf{u}_{L,1})}{\partial \sigma_1} \ldots \frac{\partial^{|\sigma_r|} C_{L-1,r}(\mathbf{u}_{L,1})}{\partial \sigma_r}, \tag{10}$$

где внешняя сумма берется по всем $k_1,\ldots,k_{n_{L-1}} \in N \cup \{0\}$ таким, что $\max_j k_j \le n_{L-1,j}$ и $\sum_{j=1}^{n_{L-1}} k_j = n - i$ для всех $i = 0,\ldots,n - n_{L-1}$.

Вторая часть формулы включает в себя производные копула-функций уровня $L-1$ с аргументами $\mathbf{u}_{L-1,j}$, $j = 1,\ldots,n_{L-1}$. Суммирование во второй части производится по $r \in \{0,1,\ldots,n_{L-1}\}$ различным подмножествам $\{\sigma_1,\ldots,\sigma_r\}$ множества $\mathbf{u}_{L-1,r}$, а произведение берется по всем порядковым номерам копула-функций уровня $L-1$.

Алгоритм нахождения плотности *n*-мерной копула-функции $C_{L,1}(\cdot)$ является рекурсивным: плотность для $C_{L,1}(\cdot)$ определяется через частные производные копула-функций $C_{L-1,1}(\cdot),\ldots,C_{L-1,n_{L-1}}(\cdot)$, которые в свою очередь могут быть вычислены с использованием (10) для копула-функций, используемых на более низких уровнях. Рекурсивная процедура завершается на самом низком уровне, когда требуется найти частные производные (различных порядков) лишь стандартных архимедовых копула-функций.

В силу рекурсивной природы алгоритма нахождения плотности для $C_{L,1}(\cdot)$ количество вычислительных операций растет вместе с усложнением копула-функции. На практике для вычисления плотности иерархической *n*-мерной копула-функции приходится прибегать к компьютерным алгебраическим системам, таким как *Mathematica* или *R* (функция *D*). Нет необходимости упоминать про то, что время, необходимое для вычисления оценок максимального правдоподобия, растет вместе с ростом размерности *n* (более детально см. (Aas et al., 2009)).

Пример 5. Используя формулу (10), получим, что плотность четырехмерной иерархической архимедовой копула-функции, представленной на рис. 9, определяется соотношением:

$$\frac{\partial^4 C_{2,1}}{\partial u_1 \ldots \partial u_4} = \frac{\partial^4 C_{2,1}}{\partial C_{1,1}^2 \partial C_{1,2}^2} \cdot \frac{\partial C_{1,1}}{\partial u_1} \cdot \frac{\partial C_{1,1}}{\partial u_2} \cdot \frac{\partial C_{1,2}}{\partial u_3} \cdot \frac{\partial C_{1,2}}{\partial u_4} + \frac{\partial^3 C_{2,1}}{\partial C_{1,1}^2 \partial C_{1,2}} \cdot \frac{\partial C_{1,1}}{\partial u_1} \cdot \frac{\partial C_{1,1}}{\partial u_2} \cdot \frac{\partial^2 C_{1,2}}{\partial u_3 \partial u_4} +$$
$$+ \frac{\partial^3 C_{2,1}}{\partial C_{1,1} \partial C_{1,2}^2} \cdot \frac{\partial^2 C_{1,1}}{\partial u_1 \partial u_2} \cdot \frac{\partial C_{1,2}}{\partial u_3} \cdot \frac{\partial C_{1,2}}{\partial u_4} + \frac{\partial^2 C_{2,1}}{\partial C_{1,1} \partial C_{1,2}} \cdot \frac{\partial^2 C_{1,1}}{\partial u_1 \partial u_2} \cdot \frac{\partial^2 C_{1,2}}{\partial u_3 \partial u_4}.$$

3.5. Моделирование наблюдений из архимедовых копула-функций

Прямое применение традиционных методов генерирования многомерных наблюдений, подчиняющихся заданному закону распределения, в принципе, возможно и к распределениям, описанным с помощью копула-функций, см. (Frees, Valdez, 1998; McNeil, 2008). К сожалению, эти методы оказываются чрезмерно трудоемкими в случае вложенных архимедовых копула-функций и многомерных перестановочных архимедовых копула-функций. Для последних моделирование наблюдений целесообразно проводить с использованием *метода условной инверсии*, который также можно использовать для любой копула-функции.

Если имеется копула-функция $C = C(u_1,\ldots,u_n)$ (не обязательно архимедова), и необходимо сгенерировать наблюдения $(u_1,\ldots,u_n)$ n-мерного распределения случайного вектора $(U_1,\ldots,U_n)$, копула-функция которого равна $C(\cdot)$, а частные распределения являются равномерными на отрезке $[0,1]$, то можно воспользоваться *методом условного распределения*. Пусть $C_k(u_1,u_2,\ldots,u_k) = C(u_1,u_2,\ldots,u_k,1,\ldots,1)$, $k = 1,\ldots,n$; при этом $C_1(u_1) = u_1$ и $C_n(u_1,u_2,\ldots,u_n) = C(u_1,\ldots,u_n)$. Условное распределение U_k, $k = 2,\ldots,n$ при заданных $U_1,\ldots,U_{k-1}$ равно:

$$C_k(u_k \mid u_1,u_2,\ldots,u_{k-1}) = P(U_k \le u_k \mid U_1 = u_1,\ldots,U_{k-1} = u_{k-1}) = \\ = C_k(u_1,u_2,\ldots,u_k) / C_{k-1}(u_1,u_2,\ldots,u_{k-1}).$$

Приведем алгоритм моделирования:

- смоделировать n независимых равномерно распределенных случайных величин $v_1,\ldots,v_n$;
- положить $u_1 = v_1$;
- аналитически или с использованием численных методов найти обратные функции условных распределений $C_k(u_1,u_2,\ldots,u_k)$ при $k = 2,\ldots,n$.

Результатом этого алгоритма является вектор наблюдений $(u_1,\ldots,u_n)$, смоделированный в соответствии с распределением $C(\cdot)$. Несмотря на то что моделирование с использованием условных распределений весьма элегантно, сама процедура может быть весьма трудоемкой с вычислительной точки зрения.

В случае архимедовых копула-функций, условное распределение U_k при заданных значениях $U_1,\ldots,U_{k-1}$ определяется теоремой 4.

Теорема 4. *Пусть* $C(u_1,u_2,\ldots,u_n) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2) + \ldots + \varphi(u_n))$ *есть n-мерная архимедова копула-функция с генератором* $\varphi(\cdot)$. *Тогда для* $k = 2,\ldots,n$

$$C_k(u_k \mid u_1,u_2,\ldots,u_{k-1}) = \frac{\varphi^{-1(k-1)}(\varphi(u_1) + \varphi(u_2) + \ldots + \varphi(u_k))}{\varphi^{-1(k-1)}(\varphi(u_1) + \varphi(u_2) + \ldots + \varphi(u_{k-1}))}.$$

Доказательство. См. (Cherubini et al., 2004, с. 183).

В работе (Marshall, Olkin, 1988) предложен метод построения архимедовых копула-функций с использованием преобразования Лапласа и обратной к нему функции. Этот метод

особенно полезен, учитывая то, что каждое абсолютно монотонное отображение $[0,\infty]$ в $[0,1]$ можно представить в терминах преобразования Лапласа и обратной к нему функции. Пусть $G(\cdot)$ — функция распределения на $\mathbb{R}^+$ такая, что $G(0)=0$, а ее преобразование Лапласа–Стилтьеса определяется соотношением

$$\widehat{G}(t)=\int_0^\infty e^{-tx}dG(x), \quad t\geq 0.$$

Можно показать, что $\widehat{G}:[0,\infty]\to[0,1]$ является непрерывной, абсолютно монотонной и строго убывающей функцией. Поэтому $\widehat{G}(\cdot)$ — хороший выбор для обратной функции генератора архимедовой копулы.

Как показано в работах (Frees, Valdez, 1998; Cherubini et al., 2004; McNeil et al., 2005), с использованием преобразования Лапласа–Стилтьеса можно предложить следующий алгоритм моделирования наблюдений из распределения, описанного многомерной архимедовой копула-функцией:

- смоделировать случайную величину V из распределения $G(\cdot)$ такого, что $\widehat{G}(\cdot)$ (преобразование Лапласа функции $G(\cdot)$) является обратной функцией генератора $\varphi(\cdot)$ копула-функции, из которой моделируются наблюдения. Например:
  - — для копула-функции Клейтона V имеет гамма-распределение $Ga(1/\alpha,1)$, $\alpha>0$, а $\widehat{G}(t)=(1+t)^{-1/\alpha}$; отметим, что $\widehat{G}^{-1}=(t^{-\alpha}-1)$ отличается от генератора под номером 1 из табл. 2 лишь на константу;
  - — для копула-функции Гумбеля V имеет устойчивое распределение $St(1/\alpha,1,\gamma,0)$ с

$$\gamma=cos^{\alpha}(\pi\alpha/2),\ \alpha>1 \text{ и } \widehat{G}(t)=\exp(-t^{1/\alpha});$$

  - — для копула-функции Франка V имеет дискретное распределение

$$P(V=k)=(1-e^{-\alpha})^k/(k\alpha) \text{ при } k=1,2,\ldots \text{ и } \alpha>0.$$

- сгенерировать независимые одинаково распределенные случайные величины $X_1,\ldots,X_n$.
- взять $u_i=\widehat{G}\left(-\frac{\ln X_i}{V}\right)$, $i=1,\ldots,n$.

Вектор $(u_1,\ldots,u_n)^{\mathrm{T}}$ смоделирован из n-мерной архимедовой копула-функции. Отметим, что для рассмотренных выше случаев алгоритм предполагает необходимость моделирования случайных величин из гамма-распределения, устойчивого распределения или дискретного распределения. Более подробную информацию см. в (Cherubini et al., 2004; Schoutens, 2003; McNeil et al., 2005; McNeil, 2008).

3.6. Эмпирические приложения в статистическом пакете *R*: архимедовы копула-функции

Для того чтобы вычислить значения функции распределения двумерной архимедовой копула-функции и ее плотности, необходимо воспользоваться некоторыми процедурами из модуля `fCopulae` в пакете *R*:

fcopulae функция	Описание
rarchmCopula	моделирование случайных величин из архимедовой копула-функции
parchmCopula	вычисление значения архимедовой копула-функции
darchmCopula	вычисление значения плотности архимедовой копула-функции
rarchmSlider	график наблюдений, смоделированных из архимедовой копула-функции
parchmSlider	график функции распределения
darchmSlider	график функции плотности
archmCopulaFit	оценка параметров архимедовой копула-функции

Например, для того чтобы смоделировать наблюдения из двумерной копула-функции Гумбеля, а затем по ним оценить параметры той же копула-функции, необходимо использовать код:

```
# Random Variates:
R = archmCopulaSim (n = 1000, alpha = 1, type = "4")
# Fit:
fit = archmCopulaFit (u = R [, 1], v = R [, 2], type = "4")
fit
```

Если же рассматривать многомерную перестановочную архимедову копула-функцию, то необходимо прибегнуть к использованию модуля `copula`. Ниже представлен пример кода для трехмерной перестановочной архимедовой копула-функции (для более детальной информации рекомендуем обратиться к руководству статистическим пакетом *R*), а результаты моделирования изображены на рис. 10.

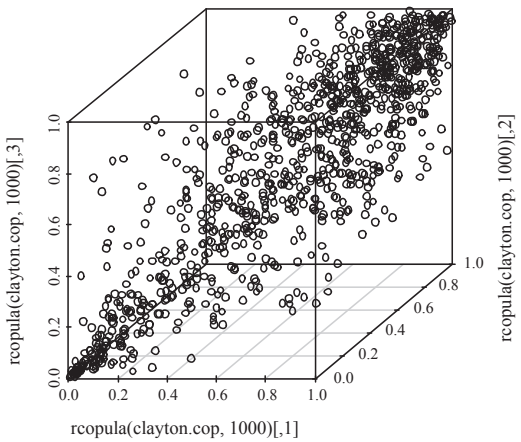


Рис. 10. График смоделированных наблюдений из трехмерной копула-функции Клейтона

```
library(copula)
#Some graphical examples
v <- rcopula(claytonCopula(2, dim = 3), 1000)
scatterplot3d(v)
frank.cop <- frankCopula(3)
persp(frank.cop, dcopula)
gumbel.cop <- archmCopula("gumbel", 5)
contour(gumbel.cop, dcopula)
# A 5-dim Frank copula
frank.cop <- frankCopula(3, dim = 5)
```

Для моделирования и оценки *четырехмерной полностью вложенной архимедовой копулы* можно было бы обратиться к модулю «copula GOF» (недавно разработанному Дэниелем Бергом). Однако к моменту написания данной работы этот модуль находился в состоянии «β-версии» и не был общедоступным, т. е. он еще не был загружен в среду пакета *R*.

Список литературы

Aas K., Czado C., Frigessi A., Bakken H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44 (2), 182–198.

Alsina M., Frank J., Schweizer B. (2005). Problems on associative functions. *Aequationes Mathematicae*, 66, 128–140.

Ang A., Chen J. (2002). Asymmetric correlations of equity portfolios. *Journal of Financial Economics*, 63 (3), 443–494.

Bouyé E., Durrleman V., Nikeghbali A., Riboulet G., Roncalli T. (2000). Copulas for finance a reading guide and some applications. Groupe de Recherche Operationnelle, Credit Lyonnais, *Working Paper*.

Breymann W., Dias A., Embrechts P. (2003). Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance*, 3, 1–14.

Chen X., Fan Y., Patton A. (2004). Simple tests for models of dependence between multiple financial time series, with applications to U. S. equity returns and exchange rates. London Economics Financial Markets Group, *Working Paper n. 483*.

Cherubini U., Vecchiato W., Luciano E. (2004). *Copula methods in finance.* Wiley.

Clayton D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65, 141–151.

Daul S., De Giorgi E., Lindskog F., McNeil A. (2003). The grouped t-copula with an application to credit risk. *Risk*, 1, 73–76.

Demarta S., McNeil A. (2005). The t copula and related copulas. *International Statistical Review*, 73, 111–129.

Embrechts P., McNeil A. J., Straumann D. (1999). Correlation and dependency in risk management: Properties and pitfalls. *Working paper*, Department of Mathematik, ETHZ, Zurich. (Now in M. A. H. Dempster (ed.) (2002), *Risk Management: Value at Risk and Beyond,* 176–223. Cambridge: Cambridge University Press).

Embrechts P., Lindskog F., McNeil A. (2003). Modelling dependence with copulas and applications to risk management. In: Rachev S. T. (ed.), *Handbook of Heavy Tailed Distributions in Finance*, Amsterdam: Elsevier/North-Holland.

Erb C., Harvey C., Viskanta T. (1994). Forecasting international equity correlations. *Financial Analysts Journal*, 50, 32–45.

Fang K., Kotz S., Ng K. (1987). *Symmetric multivariate and related distributions*. London: Chapman Hall.

Fantazzini D. (2009a). A dynamic grouped T copula approach for market risk management. In: G. Gregoriou (ed.), *A VaR Implementation Handbook*, 253–282, McGraw-Hill: New York.

Fantazzini D. (2009б). The effects of misspecified marginals and copulas on computing the value at risk: A Monte Carlo study. *Computational Statistics and Data Analysis*, 53 (6), 2168–2188.

Fantazzini D. (2010). Three-stage semi-parametric estimation of T-copulas: Asymptotics, finite-sample properties and computational aspects. *Computational Statistics and Data Analysis*, forthcoming.

Fermanian J., Scaillet O. (2003). Nonparametric estimation of copulas for time series. *Journal of Risk*, 5, 25–54.

Frank M. J. (1979). On the simultaneous associativity of $F(x,y)$ and $x+y-F(x,y)$. *Aequationes Mathematicae*, 19, 194–226.

Frees E. W., Valdez E. (1998). Understanding relationship using copulas. *North American Actuarial Journal*, 2, 1–25.

Genest C. (1987). Frank's family of bivariate distributions. *Biometrika*, 74 (3), 549–555.

Genest C., Favre A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12 (4), 347–368.

Genest C., Rivest L. (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association*, 88, 1034–1043.

Genest C., Ghoudi K., Rivest L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distribution. *Biometrika*, 82 (3), 543–552.

Gumbel E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistal Association*, 55, 698–707.

Hoeffding D. (1940). Masstabinvariante Korrelationstheorie. *Schriften des Mathematischen Seminars und des Instituts fur Angewandte Mathematik der Universität*, 5, 181–233.

Hougaard P. (1986). A class of multivariate failure time distributions. *Biometrika*, 73, 671–678.

Joe H. (1997). *Multivariate models and dependence concepts*. London: Chapman Hall.

Jondeau E., Rockinger M. (2003). Conditional volatility, skewness, and kurtosis: existence, persistence, and comovements. *Journal of Economic Dynamics and Control*, 27, 1699–1737.

Kimberling C. H. (1974). A probabilistic interpretation of complete monotonicity. *Aequationes Mathematicae*, 10, 152–164.

Lambert P., Vandenhende F. (2002). A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine*, 21, 3197–3217.

Lindskog F., McNeil A., Schmoch (2002). A note on Kendall's tau for elliptical distributions. *ETH preprint*.

Lindskog F. (2000). Modelling dependence with copulas and applications to risk management. *Master Thesis*, ETH Zurich.

Longin F., Solnik B. (2001). Extreme correlation of international equity markets. *Journal of Finance*, 56 (2), 649–676.

Malevergne Y., Sornette D. (2006). *Extreme financial risks (From dependence to risk management)*. Springer: Heidelberg.

Marshall A., Olkin I. (1988). Families of multivariate distributions. *Journal of the American Statistal Association,* 83, 834–841.

Marshall R., Zeevi A. (2002). Beyond correlation: Extreme co-movements between financial assets. Columbia University, *Working Paper*.

McNeil A. (2008). Sampling nested Archimedean copulas. *Journal of Statistical Computation and Simulation*, 78 (6), 567–581.

McNeil A., Nešlehová J. (2009). Multivariate Archimedean copulas, d-monotone functions and L1-norm symmetric distributions. *Annals of Statistics*, 37 (5), 3059–3097.

McNeil A., Frey R., Embrechts P. (2005). *Quantitative risk management: Concepts, techniques and tools*. New Jersey: Princeton Series in Finance.

Mills F. C. (1927). *The behaviour of prices*. New York: National Bureau of Economic Research.

Müller A., Scarsini M. (2005). Archimedean copulae and positive dependence. *Journal of Multivariate Analysis*, 93, 434–445.

Nelsen R. (1999). *An introduction to copulas*. Lecture Notes in Statistics. New York: Springer-Verlag.

Nelsen R. B. (2005). Some properties of Schur-constant survival models and their copulas. *Brazilian Journal of Probability and Statistics*, 19, 179–190.

Nelsen R. B. (2006). *An introduction to copulas.* Lecture Notes in Statistics, 2nd Edition. New York: Springer-Verlag.

Patton A. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *Journal of Financial Econometrics*, 2 (1), 130–168.

Patton A. (2006a). Estimation of copula models for time series of possibly different lengths. *Journal of Applied Econometrics*, 21, 147–173.

Patton A. (2006б). Modelling asymmetric exchange rate dependence. *International Economic Review*, 47 (2), 527–556.

Salvadori G., De Michele C. (2007). On the use of copulas in hydrology: Theory and practice. *Journal of Hydrologic Engineering*, 12 (4), 369–380.

Savu C., Trede M. (2009). Hierarchies of Archimedean copulas. *Quantitative Finance*, forthcoming.

Schmidt R. (2002). Tail dependence for elliptically contoured distributions. *Mathematical Methods of Operations Research*, 55 (2), 301–327.

Schoutens W. (2003). *Lévy processes in finance: Pricing financial derivatives.* Wiley.

Schweizer B., Sklar A. (1983). *Probabilistic metric spaces*. New York: Elsevier.

Schweizer B., Wolff E. F. (1976). Sur une mesure de dependence pour les variables aleatoires. *Comptes Rendus de l'Academie des Sciences de Paris. Ser. A*, 283, 659–661.

Schweizer B., Wolff E. (1981). On non-parametric measures of dependence for random variables. *Annals of Statistics*, 9, 879–885.

Sklar A. (1959). Fonctions de répartition á n dimensions et leurs marges. *Publ. Inst. Statis. Univ. Paris*, 8, 229–231.

Sklar A. (1996). Random variables, distribution functions, and copulas: Personal look backward and forward. *Lecture notes. Monograph series,* 28, 1–14.

Zhang L., Singh V. (2006). Bivariate flood frequency analysis using the copula method. *Journal of Hydrologic Engineering*, 11 (2), 150–164.

Wang S. (1998). Aggregation of correlated risk portfolios: Model and algorithms. *Proceedings of the Casualty Actuarial Society*, LXV, 848–893.

Whelan N. (2004). Sampling from Archimedean copulas. *Quantitative Finance*, 4 (3), 339–352.

Appendix 7 "Analysis of multidimensional probability distributions with copula functions - Part II."

Analysis of multidimensional probability distributions with copula functions - Part II. *Applied Econometrics*, 23(3), 98-132. http://pe.cemi.rssi.ru/pe_2011_3_98-132.pdf

Permisson to post the Publisher's version/PDF: according to http://appliedeconometrics.cemi.rssi.ru/AppEc.files/contents_en.html, "*full texts are available for downloading with approximately one year lag*". In some cases, this 1-year lag can be shorter.

Д. Фантаццини

Моделирование многомерных распределений с использованием копула-функций. II[1]

Статья содержит вторую часть консультации, посвященной копула-функциям и их использованию в моделировании многомерных распределений вероятностей. В ней описываются парные копула-функции (включая понятия канонической и D-ветвизации), различные характеристики зависимости анализируемых случайных величин (в том числе меры хвостовой зависимости, особенно актуальные в случае несимметричных распределений), а также параметрические, полупараметрические и непараметрические методы статистического оценивания распределений, представленных с помощью копула-функций.

Ключевые слова: парные копула-функции, ветвизация, меры зависимости, хвостовая зависимость, метод максимального правдоподобия (одношаговый, двухшаговый, канонический, трехшаговый), полупараметрические и непараметрические методы статистического оценивания.

JEL classification: C69, C49.

Содержание

4. Парные копула-функции
4.1. Введение
4.2. Каноническая ветвизация
4.3. D-ветвизация
4.4. Эмпирические приложения в статистическом пакете *R*: построение парных копула-функций

5. Меры зависимости
5.1. Коэффициент корреляции
5.2. Коэффициенты ранговой корреляции: ρ-Спирмена и τ-Кендалла
5.3. Хвостовая зависимость
5.4. Эмпирические приложения в статистическом пакете *R*: меры зависимости

6. Процедуры оценивания: параметрические методы
6.1. Метод максимального правдоподобия (ММП или одношаговый метод)
6.2. Двухшаговый метод максимального правдоподобия
6.3. Эмпирические приложения в статистическом пакете *R*: процедуры параметрического оценивания

7. Процедуры оценивания: полупараметрические и непараметрические методы
7.1. Канонический метод максимума правдоподобия (КММП)
7.2. Трехшаговый канонический метод максимального правдоподобия (KME–CML метод)
7.3. Методы непараметрического оценивания
7.4. Эмпирические приложения в статистическом пакете *R*: процедуры полупараметрического оценивания

[1] Продолжение. Начало консультации см. в (Фантаццини, 2011). Перевод на русский язык осуществлен А. В. Кудровым под научной редакцией С. А. Айвазяна. Сохранена сквозная нумерация разделов, рисунков, формул, определений и теорем, начатая в первой части. Окончание консультации, посвященное описанию подходов к экспериментальному подбору копула-функции и методов статистической проверки гипотез, связанных с моделями копула-функций, предполагается опубликовать в следующем номере нашего журнала.

4. ПАРНЫЕ КОПУЛА-ФУНКЦИИ

4.1. Введение

Детальное изучение парных копула-функций было начато в работе (Joe, 1997), а позднее продолжено в (Bedford, Cooke, 2001, 2002; Kurowicka, Cooke, 2006) (моделирование) и (Aas et al., 2009) (статистические выводы). Их использование дает возможность получить значительно более гибкую структуру зависимости, чем структура перестановочных или иерархически вложенных архимедовых копула-функций. При помощи парных копула-функций можно разложить многомерную плотность на произведение $n(n-1)/2$ двумерных копула-функций, из которых $n-1$ являются безусловными, а остальные — условными. Важно отметить, что используемые двумерные копула-функции не обязаны принадлежать одному классу.

Рассмотрим n случайных величин $(X_1,\ldots,X_n)$ с совместной функцией распределения $H(x_1,\ldots,x_n)$, частными распределениями $F_i(x_i)$, $i=1,\ldots,n$, совместной плотностью $f(x_1,\ldots,x_n)$ и плотностями частных распределений $f_i(x_i)$, $i=1,\ldots,n$. Тогда для точек, в которых все плотности непрерывны и положительны[2]:

$$f(x_1,\ldots,x_n)=f(x_n\mid x_1,\ldots,x_{n-1})\cdot f(x_1,\ldots,x_{n-1})=\prod_{t=2}^{n} f(x_t\mid x_1,\ldots,x_{t-1})\cdot f_1(x_1)\,. \qquad (11)$$

Из теоремы Склара известно, что для абсолютно непрерывного n-мерного распределения $H(\cdot)$, имеющего строго возрастающие непрерывные частные распределения $F_1(\cdot),\ldots,F_n(\cdot)$, имеем:

$$f(x_1,\ldots,x_n)=c_{1,\ldots,n}(F_1(x_1),\ldots,F_n(x_n))\cdot f_1(x_1)\cdot\ldots\cdot f_n(x_n),$$

где $c_{1,\ldots,n}(\cdot,\ldots,\cdot)$ — n-мерная плотность копула-функции, описывающей анализируемое распределение $H(x_1,\ldots,x_n)$, а $F_1(x_1),\ldots,F_n(x_n)$ — значения частных функций распределения рассматриваемых случайных величин в точках $x_1,\ldots,x_n$ соответственно.

В двумерном случае

$$f(x_1,x_2)=c_{12}(F_1(x_1),F_2(x_2))\cdot f_1(x_1)\cdot f_2(x_2),$$

откуда следует, что:

$$f(x_1\mid x_2)=c_{12}(F_1(x_1),F_2(x_2))\cdot f_1(x_1).$$

Для трех случайных величин X_1, X_2 и X_3 имеем:

$$f(x_1\mid x_2,x_3)=c_{12|3}\big(F(x_1\mid x_3),F(x_2\mid x_3)\big)f(x_1\mid x_3),$$

где $c_{12|3}(\cdot)$ — копула-функция для условных функций распределений $((X_1\mid X_3),(X_2\mid X_3))$. Кроме того,

$$f(x_1\mid x_2,x_3)=c_{13|2}\big(F(x_1\mid x_2),F(x_3\mid x_2)\big)f(x_1\mid x_2).$$

[2] В правой части формулы (11) и некоторых других местах далее для того, чтобы избежать усложнения обозначений, функции условной плотности и условного распределения случайной величины X_i при условиях на переменные с номерами $J_1,J_2,\ldots,J_k$ не будут помечаться соответствующими индексами $i\mid J_1,J_2,\ldots,J_k$.

Последнее соотношение может быть обобщено на n-мерный случай:

$$f(x \mid \mathbf{v}) = c_{xv_j|\mathbf{v}_{-j}}\left(F(x \mid \mathbf{v}_{-j}), F(v_j \mid \mathbf{v}_{-j})\right) \cdot f(x \mid \mathbf{v}_{-j}), \tag{12}$$

где $\mathbf{v}$ — вектор значений компонент, на которые наложены условия, v_j — произвольно выбранная компонента вектора $\mathbf{v}$, а через $\mathbf{v}_{-j}$ обозначен вектор $\mathbf{v}$, из которого исключена j-ая компонента.

Теперь, если воспользоваться (12), то из (11) получим следующее представление для $f(x_t \mid x_1, \ldots, x_{t-1})$:

$$f(x_t \mid x_1, \ldots, x_{t-1}) = \frac{f(x_t, x_{t-1} \mid x_1, \ldots, x_{t-2})}{f(x_{t-1} \mid x_1, \ldots, x_{t-2})} = c_{t-1,t|1,\ldots,t-2} \cdot f(x_t \mid x_1, \ldots, x_{t-2}), \tag{13}$$

где для различающихся индексов $i, j, i_1, \ldots, i_k$ таких, что $i < j$ и $i_1 < \ldots < i_k$, можно выразить двумерную условную копула-функцию при значениях аргументов, равных соответствующим условным функциям распределения, а именно:

$$c_{i,j|i_1,\ldots,i_k} = c_{i,j|i_1,\ldots,i_k}\left(F(x_i \mid x_{i_1}, \ldots, x_{i_k}), (F(x_j \mid x_{i_1}, \ldots, x_{i_k})\right).$$

Используя (13), можно записать (11) в виде:

$$\begin{aligned} f(x_1, \ldots, x_n) &= f(x_1) \cdot \prod_{t=2}^{n} \prod_{k=1}^{t-1} c_{t-k,t|1,\ldots,t-k-1} \cdot f(x_t) = \prod_{r=1}^{n} f(x_r) \cdot \prod_{t=2}^{n} \prod_{k=1}^{t-1} c_{t-k,t|1,\ldots,t-k-1} = \\ &= \prod_{r=1}^{n} f(x_r) \cdot \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{j,j+1|1,\ldots,j-1}, \quad \text{где} \quad j = t-k, \quad j+i = t. \end{aligned} \tag{14}$$

Представление (14) называют разложением парных копула-функций. Для многомерных распределений существует много возможностей разложения на парные копула-функции. В (Bedford, Cooke, 2001, 2002) представлена графическая модель, называемая *моделью регулярной ветвизации*. Основной идеей этого метода является представление разложения на парные копула-функции типа (14) в виде последовательности вложенных деревьев с неориентированными ребрами, названными ветвями. В n-мерном случае ветвизация представлена $n-1$ *деревом* так, что j-ое дерево имеет $n+1-j$ *узлов* и $n-j$ *ребер*. Каждое ребро отвечает плотности некоторой парной копула-функции, а ребра j-го дерева становятся узлами $(j+1)$-го дерева. Два узла $(j+1)$-го дерева связаны ребром в том случае, если соответствующие ребра j-го дерева имеют общий узел. Разложение на парные копула-функции определяется $n(n-1)/2$ ребрами, а также плотностями частных распределений.

Сосредоточим внимание на двух примерах регулярной ветвизации, которые на сегодняшний день привлекли большое внимание исследователей. Речь пойдет о *канонической ветвизации* и *D-ветвизации*. Для более детальной информации относительно общей теории регулярной ветвизации и ее графического представления рекомендуется работа (Kurowicka, Cooke, 2006).

Рассмотрение упомянутых выше примеров регулярной ветвизации начнем с двумерных условных распределений, которые являются ключевыми элементами этой процедуры.

Условные распределения для наиболее часто используемых двумерных копула-функций

В работе (Joe, 1997) показано, что:

$$F(x\,|\,\mathbf{v}) = \frac{\partial C_{x,v_j|\mathbf{v}_{-j}}\left(F(x\,|\,\mathbf{v}_{-j}), F(v_j\,|\,\mathbf{v}_{-j})\right)}{\partial F(v_j\,|\,\mathbf{v}_{-j})} \text{ для любого } j, \tag{15}$$

где $C_{x,v_j|\mathbf{v}_{-j}}$ — условная двумерная копула-функция (формальное определение см. выше). Если $\mathbf{v}$ имеет размерность 1 (т. е. $\mathbf{v} = v$), то (15) принимает вид:

$$F(x\,|\,v) = \frac{\partial C_{x,v}\left(F_1(x), F_2(v)\right)}{\partial F(v)}.$$

Пусть $F_1(x) = P(X \le x) = x$ и $F_2(v) = P(V \le v) = v$ (т. е. $f(x) = f(v) = 1$). Введем обозначение:

$$h(x,v;\Theta) = F(x\,|\,v) = \frac{\partial C_{x,v}(x,v;\Theta)}{\partial v}, \tag{16}$$

где через Θ обозначены параметры копула-функции совместного распределения $P(X \le x,\ V \le v)$. Согласно терминологии работы (Aas et al., 2009) функцию $h(x,v;\Theta)$ назовем *h*-функцией, а $h^{-1}(u,v;\Theta)$ — обратной *h*-функцией относительно первого аргумента *u*.

Для двумерной *нормальной копула-функции* *h*-функция имеет вид:

$$h(u_1,u_2;\rho_{12}) = \Phi\left(\frac{\Phi^{-1}(u_1) - \rho_{12}\Phi^{-1}(u_2)}{\sqrt{1-\rho_{12}^2}}\right),$$

а обратная *h*-функция относительно первого аргумента равна:

$$h^{-1}(u_1,u_2;\rho_{12}) = \Phi\left(\Phi^{-1}(u_1)\sqrt{1-\rho_{12}^2} + \rho_{12}\Phi^{-1}(u_2)\right),$$

где $\Phi(z)$ — значение функции распределения стандартного нормального закона в точке *z,* а ρ_{12} — коэффициент корреляции между анализируемыми случайными величинами.

Для двумерной *копула-функции Стьюдента* с ν_{12} числом степеней свободы *h*-функция равна:

$$h(u_1,u_2;\rho_{12},\nu_{12}) = t_{\nu_{12}+1}\left(\left(t_{\nu_{12}}^{-1}(u_1) - \rho_{12}t_{\nu_{12}}^{-1}(u_2)\right)\left(\frac{\left(\nu_{12} + \left(t_{\nu_{12}}^{-1}(u_2)\right)^2\right)(1-\rho_{12}^2)}{\nu_{12}+1}\right)^{-1/2}\right),$$

а обратная *h*-функция относительно первого аргумента равна

$$h^{-1}(u_1,u_2;\rho_{12},\nu_{12}) = t_{\nu_{12}}\left(t^{-1}_{\nu_{12}+1}\left(\sqrt{\frac{\left(\nu_{12}+\left(t^{-1}_{\nu_{12}}(u_2)\right)^2\right)(1-\rho^2_{12})}{\nu_{12}+1}}\right)+\rho_{12}t^{-1}_{\nu_{12}}(u_2)\right).$$

В приведенных формулах используются следующие обозначения: $t_v(z)$ — значение функции распределения Стьюдента с v степенями свободы в точке z, а $t_v^{-1}(u)$ — значение обратной к $t_v(z)$ функции распределения в точке u.

Для двумерной *копула-функции Клейтона* h-функция имеет вид:

$$h(u_1,u_2;\alpha_{12}) = u_2^{-\alpha_{12}-1}\left(u_1^{-\alpha_{12}}+u_2^{-\alpha_{12}}-1\right)^{-1-1/\alpha_{12}},$$

а обратная h-функция относительно первого аргумента равна:

$$h^{-1}(u_1,u_2;\alpha_{12}) = \left(\left(u_1\cdot u_2^{\alpha_{12}+1}\right)^{-\frac{\alpha_{12}}{\alpha_{12}+1}}+1-u_2^{-\alpha_{12}}\right)^{-1/\alpha_{12}}.$$

Для двумерной *копула-функции Гумбеля* h-функция принимает вид:

$$h(u_1,u_2;\alpha_{12}) = C_{12}(u_1,u_2)\cdot\frac{1}{u_2}\cdot(-\ln u_2)^{\alpha_{12}-1}\cdot\left\{(-\ln u_1)^{\alpha_{12}}+(-\ln u_2)^{\alpha_{12}}\right\}^{1/\alpha_{12}-1},$$

где $C_{12}(u_1,u_2)$ — двумерное распределение Гумбеля. К сожалению, в случае копула-функции Гумбеля обратная h-функция может быть вычислена лишь с использованием численных методов. В условиях большой размерности целесообразнее использовать копула-функцию Клейтона, см. (Joe, 1997) или (Cherubini et al., 2004).

4.2. Каноническая ветвизация

Регулярную ветвизацию, при которой каждое дерево имеет единственный узел, называют *канонической ветвизацией*. Разложение n-мерной плотности, отвечающее канонической ветвизации, представлено в формуле (14), т. е.

$$f(x_1,x_2,\ldots,x_n) = \prod_{k=1}^{n} f(x_k)\prod_{j=1}^{n-1}\prod_{i=1}^{n-j} c_{j,j+i|1,\ldots,j-1}(F(x_j\mid x_1,\ldots,x_{j-1}),F(x_{j+i}\mid x_1,\ldots,x_{j-1})). \quad (17)$$

На рисунке 11 показаны канонические ветвизации в 5-мерном случае:

Представления вида (17) обладают преимуществом в ситуациях, когда можно выделить одну ключевую переменную, влияющую на другие переменные. Такая переменная может быть использована в качестве корня канонической ветвизации (переменная 1 на рис. 11).

Генерирование многомерных наблюдений

В работе (Aas et al., 2009) показано, что алгоритм генерирования n зависимых случайных величин является общим как для канонической, так и для D-ветвизации:

- смоделируем независимые равномерно распределенные на [0,1] случайные величины $w_1,\ldots,w_n$;

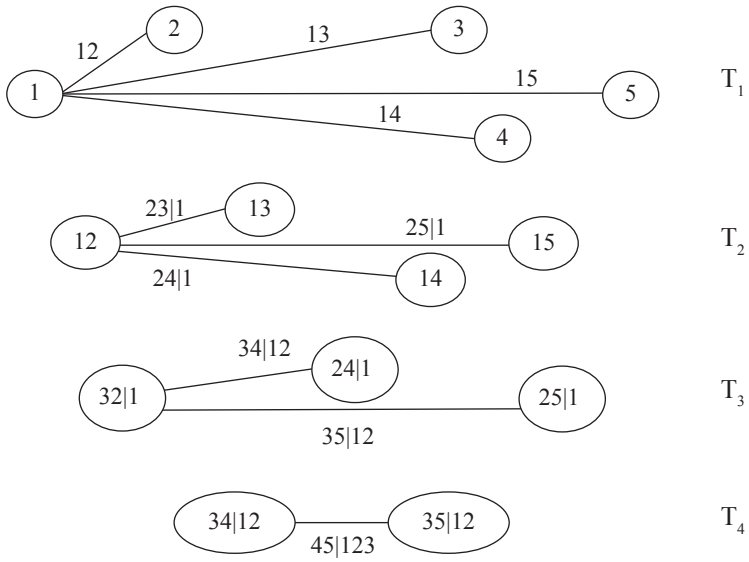


Рис. 11. Примеры канонических ветвизаций в пятимерном случае

- возьмем

$$x_1 = w_1;\ x_2 = F^{-1}(w_2 \mid x_1);\ x_3 = F^{-1}(w_3 \mid x_1, x_2);\ \ldots;\ x_n = F^{-1}(w_n \mid x_1, \ldots, x_{n-1}).$$

Для того чтобы при любом j определить распределение случайной величины $(X_j \mid X_1, X_2, \ldots, X_{j-1})$, в работе (Aas et al., 2009) предлагается использовать h-функцию, заданную соотношением (16), и соотношение (15). Отметим, что отличие канонической ветвизации от D-ветвизации состоит в выборе переменной v_j для (15). Кроме того,

$$F(x_j \mid x_1, \ldots, x_{j-1}) = \frac{\partial C_{j,j-1|1,\ldots,j-2}(F(x_j \mid x_1, \ldots, x_{j-2}), F(x_{j-1} \mid x_1, \ldots, x_{j-2}))}{\partial F(x_{j-1} \mid x_1, \ldots, x_{j-2})}.$$

На основании этих результатов в работе (Aas et al., 2009) предложена следующая процедура генерирования наблюдений, имеющих плотность в форме канонической ветвизации:

```
Sample w_1,...,w_n independent uniform on [0,1]
Set x_1 = v_{1,1} = w_1
for i ← 2,...,n
v_{i,1} = w_i
for k ← i−1, i−2,...,1
v_{i,1} = h^{-1}(v_{i,1}, v_{k,k}; Θ_{k,i−k})
end for
x_i = v_{i,1}
if i == n then
Stop
end if
for j ← 1,...,i−1
v_{i,j+1} = h(v_{i,j}, v_{j,j}; Θ_{j,i−j})
end for
end for
```

Внешний цикл указанной выше процедуры работает по порядковым номерам компонент моделируемых многомерных наблюдений. Он включает в себя два подцикла: первый из них позволяет вычислить смоделированное значение i-ой компоненты наблюдения, второй — вычислить условные распределения, необходимые для вычисления $(i+1)$-ой компоненты. Для вычисления этих условных распределений в работе (Aas et al., 2009) используется h-функция, определенная соотношением (16), для которой в качестве аргументов берутся рассчитанные $v_{i,j} = F(x_i \mid x_1, \ldots, x_{j-1})$. Параметры $\Theta_{j,i}$ для h-функции — это параметры плотности соответствующей копула-функции $c_{j,j+i|1,\ldots,j-1}(\cdot \mid \cdot)$.

Оценивание параметров

Копула-функции с канонической ветвизацией могут быть оценены с помощью метода максимального правдоподобия. Для простоты положим, как и в (Aas et al., 2009), что $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,T})$, $i = 1, \ldots, n$ независимы во времени. Такое предположение не является ограничительным, поскольку при наличии зависимости можно на первом этапе оценить одномерные модели временных рядов, а затем для стандартизированных остатков использовать копула-функцию с канонической ветвизацией. Такой подход к оцениванию можно считать расширением *метода максимального псевдоправдоподобия*. Для копула-функций он впервые был предложен в работе (Oakes, 1994), а позднее в работах (Genest et al., 1995; Shih, Louis, 1995) были представлены такие асимптотические свойства получаемых оценок, как состоятельность и асимптотическая нормальность. В (Kim et al., 2007) при помощи методов симуляционного моделирования показано, что в условиях, когда частные распределения неизвестны (это довольно распространенная ситуация в прикладных задачах), метод максимального псевдоправдоподобия дает более точные оценки, чем метод максимального правдоподобия. Кроме того, в (Kim et al., 2008) показано, что в случае многомерных моделей с гетероскедастичностью этот метод дает оценки, являющиеся состоятельными и асимптотически нормальными.

Однако в случае, когда вместо многомерной копула-функции используется разложение на парные копула-функции, методология оценивания будет отличаться от представленной выше (Aas et al., 2009).

Для копула-функции с канонической ветвизацией логарифм правдоподобия имеет следующий вид:

$$\sum_{j=1}^{n-1} \sum_{i=1}^{n-j} \sum_{t=1}^{T} \ln\left[c_{j,j+i|1,\ldots,j-1}(F(x_{j,t} \mid x_{1,t}, \ldots, x_{j-1,t}), F(x_{j+i,t} \mid x_{1,t}, \ldots, x_{j-1,t}))\right]. \qquad (18)$$

В работе (Aas et al., 2009) предложен следующий алгоритм вычисления логарифмической функции правдоподобия (18)

```
log-likelihood = 0.
for i←1,...,n
v_{0,i} = x_i
end for
for j←1,...,n−1
for i←1,...,n−j
log-likelihood = log-likelihood + L(v_{j−1,1}, v_{j−1,i+1}; Θ_{j,i})
end for
```

```
if j==n−1 then
Stop
end if
for i←1,...,n−j
```
$\mathbf{v}_{j,i} = h(\mathbf{v}_{j-1,i+1}, \mathbf{v}_{j-1,1}; \Theta_{j,i})$
```
end for
end for
```

Здесь $L(\mathbf{x}, \mathbf{v}; \Theta)$ — логарифмическая функция правдоподобия выбираемой двумерной копула-функции, параметры которой определены вектором Θ, а наблюдения $\mathbf{x}$ и $\mathbf{v}$ заданы. Кроме того, в упомянутой работе представлена процедура нахождения начальных значений оценок параметров, вычисляемых при помощи численной максимизации логарифмического правдоподобия. Эмпирический анализ показывает, что начальные значения оценок, вычисленных при помощи процедуры из (Aas et al., 2009), и итоговые значения оценок, полученных при одновременном оценивании всех параметров, весьма близки, в то время как значение функции правдоподобия увеличивается незначительно. Это может означать, что процедура вычисления начальных значений оценок параметров дает состоятельную оценку всех параметров, что значительно облегчает вычислительную процедуру. Однако этот вопрос требует более детального исследования.

4.3. D-ветвизация

Регулярная ветвизация, при которой у любого дерева T_j не существует узла, соединенного с более чем двумя ребрами, называется *D-ветвизацией*. Многомерная (*n*-мерная) плотность, отвечающая D-ветвизации, имеет вид:

$$\prod_{k=1}^{n} f(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{i,i+j|i+1,\ldots,i+j-1}(F(x_i \mid x_{i+1},\ldots,x_{i+j-1}), F(x_{i+j} \mid x_{i+1},\ldots,x_{i+j-1})).$$

На рисунке 12 показана 5-мерная D-ветвизация:

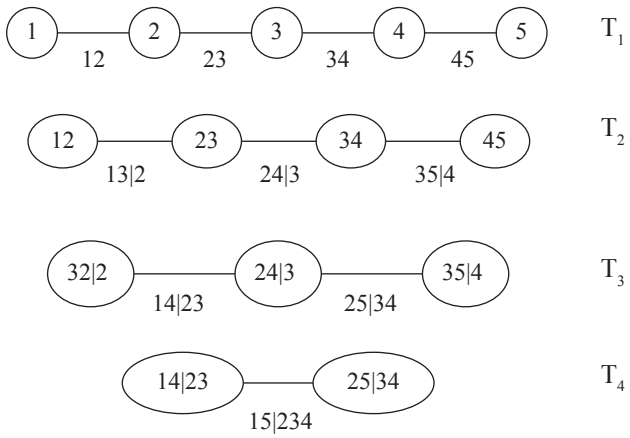


Рис. 12. D-ветвизация для пяти переменных

Генерирование многомерных наблюдений

В работе (Aas et al., 2009) предложена следующая процедура моделирования наблюдений, закон распределения которых представлен с помощью D-ветвизации:

Sample $w_1,\ldots,w_n$ independent uniform on [0,1]
Set $x_1 = v_{1,1} = w_1$
Set $x_2 = v_{2,1} = h^{-1}(w_2, v_{1,1}; \Theta_{1,1})$
$v_{2,2} = h(v_{1,1}, v_{2,1}; \Theta_{1,1})$
for $i \leftarrow 3,\ldots,n$
$v_{i,1} = w_i$
for $k \leftarrow i-1, i-2,\ldots,2$
$v_{i,1} = h^{-1}(v_{i,1}, v_{i-1,2k-2}; \Theta_{k,i-k})$
end for
$v_{i,1} = h^{-1}(v_{i,1}, v_{i-1,1}; \Theta_{1,i-1})$
$x_i = v_{i,1}$
if $i = n$ then
Stop
end if
$v_{i,2} = h(v_{i-1,1}, v_{i,1}; \Theta_{1,i-1})$
$v_{i,3} = h(v_{i,1}, v_{i-1,1}; \Theta_{1,i-1})$
for $i > 3$ then
for $j \leftarrow 2,\ldots,i-2$
$v_{i,2j} = h(v_{i-1,2j-2}, v_{i,2j-1}; \Theta_{j,i-j})$
$v_{i,2j+1} = h(v_{i,2j-1}, v_{i-1,2j-2}; \Theta_{j,i-j})$
end for
end if
$v_{i,2i-2} = h(v_{i-1,2i-4}, v_{i,2i-3}; \Theta_{i-1,1})$
end for

Как и в случае канонической ветвизации, процедура моделирования D-ветвизации состоит из одного внешнего цикла, включающего один подцикл для моделирования переменных, а также один подцикл для вычисления необходимых условных распределений. Тем не менее, с вычислительной точки зрения этот алгоритм является менее эффективным, чем алгоритм моделирования канонической ветвизации, поскольку число условных распределений, которые необходимо вычислить, для D-ветвизации равно $(n-2)^2$, а для канонической ветвизации — $(n-2)(n-1)/2$. Отметим, что в случае D-ветвизации параметры $\Theta_{j,i}$ в h-функции — это множество параметров соответствующей копула-функции $c_{i,i+j|i+1,\ldots,i+j-1}(\cdot|\cdot)$.

Оценивание параметров

В случае D-ветвизации логарифмическая функция правдоподобия имеет следующий вид:

$$\sum_{j=1}^{n-1}\sum_{i=1}^{n-j}\sum_{t=1}^{T}\ln\Big[c_{i,i+j|i+1,\ldots,i+j-1}\Big(F(x_{i,t}\,|\,x_{i+1,t},\ldots,x_{i+j-1,t}),F(x_{i+j,t}\,|\,x_{i+1,t},\ldots,x_{i+j-1,t})\Big)\Big].$$

В работе (Aas et al., 2009) предлагается следующий алгоритм вычисления логарифмической функции правдоподобия:

log-likelihood = 0.
for $i=1,\ldots,n$
$\mathbf{v}_{0,i}=\mathbf{x}_i$
end for
for $i=1,\ldots,n-1$
log-likelihood = log-likelihood + $l(\mathbf{v}_{0,i},\mathbf{v}_{0,i+1};\Theta_{1,i})$
end for
$\mathbf{v}_{1,1}=h(\mathbf{v}_{0,1},\mathbf{v}_{0,2};\Theta_{1,1})$
for $k=1,\ldots,n-3$
$\mathbf{v}_{1,2k}=h(\mathbf{v}_{0,k+2},\mathbf{v}_{0,k+1};\Theta_{1,k+1})$
$\mathbf{v}_{1,2k+1}=h(\mathbf{v}_{0,k+1},\mathbf{v}_{0,k+2};\Theta_{1,k+1})$
end for
$\mathbf{v}_{1,2n-4}=h(\mathbf{v}_{0,n},\mathbf{v}_{0,n-1};\Theta_{1,n-1})$
for $j=2,\ldots,n-1$
for $i=1,\ldots,n-j$
log-likelihood = log-likelihood + $l(\mathbf{v}_{j-1,2i-1},\mathbf{v}_{j-1,2i};\Theta_{j,i})$
end for
if $j=n-1$ then
Stop
end if
$\mathbf{v}_{j,1}=h(\mathbf{v}_{j-1,1},\mathbf{v}_{j-1,2};\Theta_{j,1})$
if $n>4$ then
for $i=1,\ldots,n-j-2$
$\mathbf{v}_{j,2i}=h(\mathbf{v}_{j-1,2i+2},\mathbf{v}_{j-1,2i+1};\Theta_{j,i+1})$
$\mathbf{v}_{j,2i+1}=h(\mathbf{v}_{j-1,2i+1},\mathbf{v}_{j-1,2i+2};\Theta_{j,i+1})$
end for
end if
$\mathbf{v}_{j,2n-2j-2}=h(\mathbf{v}_{j-1,2n-2j},\mathbf{v}_{j-1,2n-2j-1};\Theta_{j,n-j})$
end for

Здесь $l(\mathbf{x},\mathbf{v};\Theta)$ — логарифмическая функция правдоподобия для выбираемой копула-функции, параметры которой определяются вектором Θ, а векторы наблюдений $\mathbf{x}$ и $\mathbf{v}$ заданы. Отметим, что $\Theta_{j,i}$ — параметры плотности копула-функции $c_{i,i+j|i+1,\&,i+j-1}(\cdot|\cdot)$.

4.4. Эмпирические приложения в статистическом пакете *R*: построение парных копула-функций

Если нужно смоделировать и оценить копула-функцию, используя ее разложение на парные копула-функции, необходимо воспользоваться модулем процедур `copulaGOF` из пакета *R*, разработанного Даниэлем Бергом. Так, например, D-ветвизация в четырехмерном случае может быть смоделирована и оценена следующим образом:

```
x = SimulateCopulae(n=1000,d=4,construction=list(type="dpcc",copula=c("clayton",
"gumbel","frank","gumbel","clayton","gumbel")),param=list(c(2,3,6,1.3,1,1.4),
rep(0,6)))
pairs(x)
```

```
dpcc.par = EstimateCopulaParameter(x,construction=list(type="dpcc",copula=
c("clayton","gumbel","frank","gumbel","clayton","gumbel")))
dpcc.par
$t [1] 1.951387 2.991580 6.052075 1.343258 0.949556 1.410108
$nu [1] 0
$loglik [1] 1899.111
```

Биплоты (диаграммы рассеяния пар компонент) для наблюдений из четырехмерной копула-функции с D-ветвизацией представлены на рис. 13.

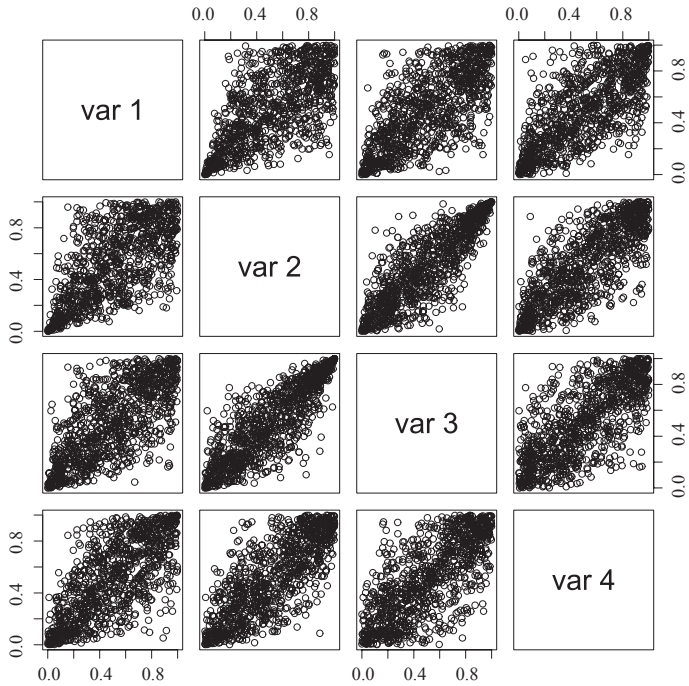


Рис. 13. Диаграмма рассеяния пар компонент для наблюдений из четырехмерной копула-функции с D-ветвизацией

5. Меры зависимости

Меры зависимости являются весьма полезными инструментами для описания структуры двумерной зависимости. В этой главе рассмотрим три возможные меры: *коэффициент корреляции*, *коэффициент ранговой корреляции* и *коэффициент хвостовой зависимости*.

Достаточно хорошей мерой зависимости в классе эллиптических распределений является коэффициент корреляции. Этот класс включает в себя, например, нормальное распределение, смеси нормальных распределений. Однако коэффициент корреляции как мера зависимости для многомерных распределений вне класса эллиптических распределений обладает рядом недостатков. В качестве альтернативы предлагаются две другие меры зависимости, которые в некоторых случаях оказываются более приемлемыми.

5.1. Коэффициент корреляции

Коэффициент корреляции между случайными величинами X и Y, как известно, определяется следующим образом (см., например, (Айвазян, 2010, п. 3.2.2)):

$$\rho(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}},$$

где $\mathrm{Cov}(X,Y)$ — ковариация между X и Y, $\mathrm{Var}(X)$ и $\mathrm{Var}(Y)$ — дисперсии X и Y соответственно. Ниже представлены основные свойства коэффициента корреляции.

1. $|\rho(X,Y)|\le 1$.
2. Если X и Y независимы, то $\rho(X,Y)=0$.
3. $|\rho(X,Y)|=1$ тогда и только тогда, когда $P(X=a+bY)=1$ для некоторых a и $b\neq 0$.
4. $\rho(\alpha X+\beta,\gamma Y+\delta)=\mathrm{sign}(\alpha\gamma)\rho(X,Y)$.
5. Пусть (X,Y) имеет совместное двумерное нормальное распределение с частными стандартными нормальными распределениями. Тогда коэффициент корреляции ρ между X и Y однозначно определяет совместное распределение (X,Y).

Обобщение коэффициента корреляции на многомерный случай можно найти во многих учебниках по статистике, например, (Mardia et al., 1997).

Пусть $\hat{\rho}$ — выборочный коэффициент корреляции. Хорошо известно, что для проверки гипотезы $H_0:\rho=0$ против альтернативы $H_1:\rho\neq 0$ может быть использована следующая тестовая статистика:

$$\hat{t}=\frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}},$$

которая при справедливости нулевой гипотезы имеет t_{n-2}-распределение — см., например, (Айвазян, 2010, с. 78).

Если воспользоваться z-преобразованием Фишера, то получим:

$$z=\frac{1}{2}\log\left(\frac{1-\hat{\rho}}{1+\hat{\rho}}\right),\quad \zeta=\frac{1}{2}\log\left(\frac{1-\rho}{1+\rho}\right),$$

$$\sqrt{n-3}(z-\zeta)\xrightarrow{d}N(0,1),$$

где знаком $\xrightarrow{d}$ обозначена асимптотическая сходимость по распределению при бесконечно растущем объеме выборки n $(n\to\infty)$. Более того, хорошо известно, что если $\rho\neq 0$, то $\sqrt{n}(\hat{\rho}-\rho)\xrightarrow{d}N\left(0,(1-\rho^2)^2\right)$.

Если X и Y имеют распределения (не обязательно нормальные), для которых конечен четвертый момент и $\rho(X,Y)\neq 0$, то

$$\sqrt{n}(\hat{\rho}-\rho)\xrightarrow{d}N\left(0,\gamma^2\right)$$

при некотором γ^2 (см. (Fang et al., 1987; Rodgers, Nicewander, 1988; Embrechts et al., 2002) и ссылки в этих работах).

К сожалению, для многомерных распределений неэллиптического типа коэффициент корреляции обладает рядом недостатков.

1. Равенство коэффициента корреляции нулю эквивалентно независимости в случае многомерного нормального распределения. Однако уже для многомерного распределения Стьюдента это не так.

2. Коэффициент корреляции инвариантен относительно линейных преобразований, но, как правило, не инвариантен относительно более общих преобразований $T(\cdot)$:

$$\rho(T(X),T(Y)) \neq \rho(X,Y).$$

Например, две случайные величины, имеющие логнормальное распределение, имеют коэффициент корреляции, отличный от коэффициента корреляции между их лог-преобразованиями.

3. Частные распределения и корреляционная матрица однозначно определяют лишь только совместное распределение эллиптического типа, но это неверно в общем случае.

4. Дисперсии случайных величин X и Y должны быть конечными, иначе корреляция не определена. Это свойство указывает на то, что коэффициент корреляции — далеко не идеальная мера зависимости, которая может быть неопределенной для распределений с «тяжелыми хвостами».

Рассмотрим пример, иллюстрирующий один из вышеупомянутых случаев. Пусть X_1, X_2 — нормально распределенные случайные величины с нулевым средним и дисперсией $\sigma^2 > 0$, а коэффициент корреляции между ними равен ρ. Тогда коэффициент корреляции случайных величин $Y_i = \exp(X_i)$, $i = 1,2$, имеющих логнормальное распределение, равен

$$\mathrm{Corr}(Y_1,Y_2) = \frac{e^{\rho\sigma^2} - 1}{e^{\sigma^2} - 1}.$$

При $\rho = 1$ всегда получим, что $\mathrm{Corr}(Y_1,Y_2)=1$, но наименьшее возможное (при $-1 \le \rho \le 1$) значение $\mathrm{Corr}(Y_1,Y_2)$ всегда будет больше –1. Например, если $\sigma = 1$, то $\mathrm{Corr}(Y_1,Y_2) \in [-0.368, 1]$. Кроме того, интервал достижимых значений $\mathrm{Corr}(Y_1,Y_2)$ будет тем меньше, чем больше σ.

В качестве другого примера, иллюстрирующего недостатки коэффициента корреляции, рассмотрим величину X_1 со стандартным нормальным распределением $N(0,1)$ и $X_2 = X_1^2$. Тогда

$$\mathrm{Cov}(X_1,X_2) = \mathrm{E}\left(X_1 \cdot (X_1^2 - 1)\right) = \mathrm{E}(X_1^3) - \mathrm{E}(X_1) = 0.$$

Таким образом, с одной стороны, имеем функциональную зависимость между X_1 и X_2, а с другой стороны, нулевую корреляцию между ними.

5.2. Коэффициенты ранговой корреляции: ρ-Спирмена и τ-Кендалла

Для того чтобы преодолеть указанные выше недостатки, необходимо обратиться к идее, возникшей в теории непараметрических статистик, которая предлагает сосредоточить внимание не на самих данных, а на их рангах. В результате были предложены две важные меры зависимости: ρ-Спирмена и τ-Кендалла (определение и основные свойства этих мер пар-

ной зависимости см., например, в (Айвазян, 2010, с. 96–107)). Перед тем, как приступить к обсуждению этих мер зависимости, коротко представим понятия согласованности и рассогласованности.

Определение 5. *Наблюдения* (x_i, y_i) *и* (x_j, y_j) *называют согласованными, если* $x_i < x_j$ *и* $y_i < y_j$, *или если* $x_i > x_j$ *и* $y_i > y_j$. *Аналогично,* (x_i, y_i) *и* (x_j, y_j) *называют рассогласованными наблюдениями, если* $x_i < x_j$ *и* $y_i > y_j$ *или если* $x_i > x_j$ *и* $y_i < y_j$. *Это эквивалентно тому, что (*x_i, y_i*)* *и* (x_j, y_j) *являются согласованными наблюдениями, если* $(x_i - x_j)(y_i - y_j) > 0$, *и рассогласованными, если* $(x_i - x_j)(y_i - y_j) < 0$.

Другими словами, согласованность наблюдений возникает в случае, если бóльшие значения одного наблюдения соответствуют бóльшим значениям другого наблюдения, а малые значения также соответствуют малым значениям. Если это не так, то наблюдения называют рассогласованными.

Коэффициент ранговой корреляции Спирмена (ρ-Спирмена)

Коэффициент ρ-Спирмена — мера согласованности между двумя случайными величинами, основанная на использовании понятий согласованности и рассогласованности.

Определение 6. *Для двух непрерывных случайных величин X и Y, совместное распределение которых имеет копула-функцию* $C(\cdot)$, *ρ-Спирмена определяется как*

$$\rho_S(X,Y) = 3Q(C,P) = 12\int\int uv\,dC(u,v) - 3 = 12\int\int C(u,v)\,du\,dv - 3, \qquad (19)$$

где $u = F_X(x)$, $v = F_Y(y)$, *а* $Q(\cdot)$ *— разность между вероятностями согласованности и рассогласованности* (*для более детальной информации см.* (*Nelsen, 2006*)).

Эквивалентное выражение для ρ-Спирмена может быть дано в форме

$$\rho_s(X,Y) = \mathrm{Corr}\big(F(X), G(Y)\big),$$

где корреляция (Corr) понимается в классическом смысле, а F и G — функции частных распределений случайных величин X и Y. Этот результат вытекает из того факта, что ранги u и v являются наблюдениями из равномерно распределенных на [0; 1] случайных величин $U = F(X)$ и $V = G(Y)$, совместное распределение которых определяется копула-функцией C.

Поскольку U и V имеют среднее 1/2 и дисперсию 1/12, выражение для $\rho_S(X,Y)$ в (18) может быть представлено в следующем виде:

$$\rho_S(X,Y) = 12\int\int uv\,dC(u,v) - 3 = 12\mathrm{E}(UV) - 3 = \frac{\mathrm{E}(UV) - 1/4}{1/12} = \frac{\mathrm{E}(UV) - \mathrm{E}(U)\mathrm{E}(V)}{\sqrt{\mathrm{Var}(U)\,\mathrm{Var}(V)}}.$$

Таким образом, ρ-Спирмена для пары непрерывных случайных величин X и Y равно корреляции между рангами X и Y.

В многомерном случае ρ-Спирмена (*матрица*) определяется как

$$\rho_S(\mathbf{X}) = \mathbf{Corr}\big(F_1(X_1), \ldots, F_n(X_n)\big),$$

где $\mathbf{Corr}(\cdot)$ — корреляционная матрица; компоненты матрицы $\rho_S(\mathbf{X})$ определяются как

$$\rho_S(\mathbf{X})_{ij} = \mathrm{Corr}\left(F_i(X_i), F_j(X_j)\right).$$

Перечислим основные свойства коэффициента ρ-Спирмена.

1. ρ_S симметричен.

2. $|\rho_S| \le 1$. В частности, для $Y = T(X)$, где $T(\cdot)$ — строго возрастающая функция, $\rho_S(X,Y) = 1$; а для $Y = T(X)$, где $T(\cdot)$ — строго убывающая функция, $\rho_S(X,Y) = -1$.

3. $\rho_S(T_x(X), T_y(Y)) = \rho_S(X,Y)$ для любых строго возрастающих отображений $T_x(\cdot)$ и $T_y(\cdot)$.

4. Если X и Y независимы, то $\rho_S(X,Y) = 0$.

5. Если (X,Y) имеет нормальное распределение с коэффициентом корреляции ρ и стандартными нормальными частными распределениями, то

$$\rho = 2\sin\left[(\pi/6)\rho_S\right] \quad \text{или} \quad \rho_S = \frac{6}{\pi}\arcsin\left(\frac{1}{2}\rho\right).$$

Последний результат справедлив также в случае, когда совместное распределение случайных величин X и Y имеет нормальную копула-функцию, а частные распределения непрерывны. К сожалению, соотношение между коэффициентом ρ-Спирмена и коэффициентом корреляции не выполняется для всех распределений из эллиптического семейства — см., например, (Hult, Lindskog, 2002; Nelsen 2006).

Можно показать, что непараметрическая оценка для ρ_S, построенная по наблюдениям (X_i, Y_i), $i = 1,\ldots,T$, имеет вид

$$\hat{\rho}_S = \frac{6}{T(T-1)(T-2)} \sum_{1 \le i < j < k \le T} 3 \cdot \mathrm{sign}(X_i - X_j) \cdot \mathrm{sign}(Y_i - Y_k)$$

и является несмещенной.

Если X_i и Y_i независимы, $i = 1,\ldots,T$, то распределение $\sqrt{T}\,\hat{\rho}_S$ асимптотически сходится к $N(0,1)$. Более детальную информацию см. в (Айвазян, 2010; Hollander, Wolfe, 1973; Statistics with confidence, 1989; Conover, 1999; van de Wiel, Bucchianico, 2001).

Коэффициент ранговой корреляции Кендалла (τ-Кендалла)

Коэффициент τ-Кендалла — это также мера связи между двумя случайными величинами, основанная на использовании понятий согласованности и рассогласованности и определяемая следующим образом.

Определение 7 (заимствовано из (Kruskal, 1958), см. также (Айвазян, 2010, с. 102–107)). *Пусть (X_1, Y_1) — случайный вектор с распределением $H(x,y)$, а (X_2, Y_2) — случайный вектор, такой, что X_2 и Y_2 независимы, распределение X_1 совпадает с распределением X_2, распределение Y_1 совпадает с распределением Y_2. Коэффициент τ-Кендалла для вектора (X, Y), компоненты которого есть непрерывные случайные величины, определяется как разность между вероятностью согласованности и вероятностью рассогласованности, т. е.*

$$\tau(X,Y) = \mathrm{P}\left\{(X_1 - X_2)(Y_1 - Y_2) > 0\right\} - \mathrm{P}\left\{(X_1 - X_2)(Y_1 - Y_2) < 0\right\} = \mathrm{E}[\mathrm{sign}(X_1 - X_2)\,\mathrm{sign}(Y_1 - Y_2)]$$
$$= \mathrm{P}\left\{X_1 > X_2,\, Y_1 > Y_2\right\} + \mathrm{P}\left\{X_1 < X_2,\, Y_1 < Y_2\right\} - \mathrm{P}\left\{X_1 < X_2,\, Y_1 > Y_2\right\} - \mathrm{P}\left\{X_1 > X_2,\, Y_1 < Y_2\right\}.$$

Тогда для n-мерного случайного вектора $\mathbf{X}$ и случайного вектора $\tilde{\mathbf{X}}$, компоненты которого независимы, а их частные распределения равны частным распределениям соответствующих компонент вектора $\mathbf{X}$, матрица τ-Кендалла $\rho_\tau(\mathbf{X})$ определяется следующим образом:

$$\rho_\tau(\mathbf{X}) = \mathrm{Cov}\left[\mathrm{sign}(\mathbf{X} - \tilde{\mathbf{X}})\right],$$

где (i, j)-ый элемент этой матрицы определяется формулой

$$\rho_\tau(\mathbf{X})_{ij} = \mathrm{Cov}[\mathrm{sign}(X_i - \tilde{X}_i), \mathrm{sign}(X_j - \tilde{X}_j)].$$

В частности, в двумерном случае ($n = 2$) матрицей τ-Кендалла будет

$$\rho_\tau(\mathbf{X}) = \begin{pmatrix} 1 & \tau(X,Y) \\ \tau(X,Y) & 1 \end{pmatrix}.$$

Следует отметить, что существуют некоторые другие n-мерные обобщения τ-Кендалла. Более детальную информацию можно найти, например, в (Clemen, Jouini, 1996) и (Barbe et al., 1996). Основные свойства этой меры зависимости, а также соответствующие доказательства представлены в работах (Embrechts et al., 2002) и (Nelsen 2006).

Как правило, если τ-Кендалла положителен, то с большой вероятностью мы имеем положительную зависимость. Если этот коэффициент отрицателен, то следует ожидать отрицательную зависимость. Кроме того, τ-Кендалла может быть представлен в терминах копула-функции, что облегчает вычислительную работу.

Теорема 5. Коэффициент *τ-Кендалла для двумерной случайной величины, характеризующейся копулой* $C(u,v)$, *определяется формулой*

$$\tau(X,Y) = 4\int_0^1\int_0^1 C(u,v)dC(u,v) - 1.$$

Доказательство см. в (Nelsen, 1999, p. 127–129).

Основные свойства коэффициента τ-Кендалла:

1. $\tau(X,Y)$ симметричен.
2. $-1 \le \tau(X,Y) \le 1$.
3. Если X и Y независимы, то $\tau(X,Y) = 0$.
4. $\tau(T_x(X), T_y(Y)) = \tau(X,Y)$ для любых строго возрастающих отображений $T_x(\cdot)$ и $T_y(\cdot)$.

Оценка коэффициента τ-Кендалла требует вычисления двойного интеграла, что для распределений эллиптического типа представляет собой довольно непростую задачу. Однако можно показать (см. (Lindskog et al., 2003)), что *τ-Кендалла для эллиптических распределений* определяется следующим образом:

$$\tau(X,Y) = \frac{2}{\pi}\arcsin\rho,$$

где ρ — обычный парный коэффициент корреляции.

Для архимедовых копула-функций ситуация значительно проще, т. к. $\tau(X,Y)$ может быть оценен с использованием генератора копула-функции. А именно, в работе (Genest, MacKay,

1986) показано, что τ-*Кендалла для архимедовых копула-функций* определяется соотношением:

$$\tau(X,Y)=1+4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}dt,$$

где $\varphi(\cdot)$ — генератор копула-функции. Например, можно показать, что для копула-функций Клейтона или Гумбеля мы имеем следующие результаты:

- для копула-функции Клейтона: $\tau(X,Y)=\dfrac{\alpha}{\alpha+2}$;
- для копула-функции Гумбеля: $\tau(X,Y)=1-\dfrac{1}{\alpha}$,

где α — параметр копула-функции.

Можно показать, что непараметрическая оценка для τ-Кендалла, построенная по наблюдениям (X_i,Y_i), $i=1,\ldots,T$ и имеющая вид

$$\hat{\tau}=\frac{2}{T(T-1)}\sum_{1\le i<j\le T}\operatorname{sign}(X_i-X_j)\operatorname{sign}(Y_i-Y_j)$$

является несмещенной.

Если X_i и Y_i независимы, то распределение случайной величины $\sqrt{T}\hat{\tau}$ сходится к $N(0,4/9)$. Более подробная информация содержится, например, в (Ferguson et al., 2000).

5.3. Хвостовая зависимость

Определим понятия верхней и нижней хвостовой зависимости (Joe, 1997).

Определение 8. *Пусть (X, Y) — случайный вектор, компоненты которого есть непрерывные случайные величины с частными распределениями F_X и F_Y. Тогда коэффициент λ^U верхней хвостовой зависимости X и Y определяется соотношением*:

$$\lambda^U=\lim_{u\to 1}\mathrm{P}[Y>F_Y^{-1}(u)\,|\,X>F_X^{-1}(u)]=\lim_{u\to 1}\mathrm{P}[X>F_X^{-1}(u)\,|\,Y>F_Y^{-1}(u)]=\lim_{u\to 1}\frac{1-2u+C(u,u)}{1-u}$$

при условии, что пределы существуют, где $C(u,v)$ — двумерная копула-функция случайного вектора (X,Y).

При этом случайные величины X и Y называют асимптотически зависимыми на верхнем хвосте, если $\lambda^U\in(0,1]$, X и Y называют асимптотически независимыми, если $\lambda^U=0$.

Другими словами, верхняя хвостовая зависимость существует тогда, когда имеется положительная вероятность одновременного возникновения положительные выбросов. λ^U широко используется в теории экстремальных значений и представляет собой вероятность того, что одна переменная примет экстремальные значения при условии, что другая переменная также принимает экстремальные значения. Коэффициент λ^U можно рассматривать как меру зависимости квантилей — см., например, (Coles et al., 1999).

Аналогично можно определить нижнюю хвостовую зависимость.

Определение 9. *Пусть* (X,Y) *— двумерный случайный вектор, компонентами которого являются непрерывные случайные величины, частные распределения которых равны* F_X *и* F_Y. *Тогда коэффициент* λ^L *нижней хвостовой зависимости между X и Y равен*:

$$\lambda^L = \lim_{u\to 0} \mathrm{P}[Y \le F_Y^{-1}(u) \mid X \le F_X^{-1}(u)] = \lim_{u\to 0} \mathrm{P}[X \le F_X^{-1}(u) \mid Y \le F_Y^{-1}(u)] = \lim_{u\to 0} \frac{C(u,u)}{u}$$

при условии, что пределы существуют.

Случайные величины X и Y называются асимптотически зависимыми на нижнем хвосте, если $\lambda^L \in (0,1]$ *и асимптотически независимыми на нижнем хвосте, если* $\lambda^L = 0$.

Таким образом, нижняя хвостовая зависимость существует тогда, когда существует положительная вероятность возникающих одновременно отрицательных выбросов.

Поскольку эллиптические распределения являются радиально симметричными, для них коэффициенты верхней и нижней хвостовой зависимости будут равны между собой. Например, для *нормальной* копула-функции коэффициенты верхней и нижней хвостовой зависимости определяются соотношением

$$\lambda^L = \lambda^U = 2\lim_{x\to\infty}[\,\mathrm{P}(Y > x \mid X = x)] = 2\lim_{x\to\infty}\left[1 - \Phi\left(x\frac{\sqrt{1-\rho}}{\sqrt{1+\rho}}\right)\right] = 0.$$

Таким образом, вне зависимости от величины парного коэффициента корреляции ρ $(\rho \ne 1)$, если рассматривать экстремально большие положительные значения X и Y на правом хвосте, то для них получим асимптотическую независимость. Этот факт хорошо известен — см., например, (Sibuya, 1961; Resnick, 1987).

Для *копула-функции Стьюдента* коэффициенты верхней и нижней хвостовой зависимости также совпадают и равны

$$\lambda^L = \lambda^U = 2\lim_{x\to\infty}[\,\mathrm{P}(Y > x \mid X = x)] = 2 - 2\mathrm{t}_{\nu+1}\left(\sqrt{\nu+1}\cdot\frac{\sqrt{1-\rho}}{\sqrt{1+\rho}}\right).$$

Очевидно, они возрастают по ρ и убывают по ν. В этой формуле через $t_{\nu+1}$, как и ранее, обозначена одномерная функция распределения Стьюдента с $\nu+1$ степенями свободы. Если число степеней свободы устремить к бесконечности, то λ^U будет стремиться к 0 при $\rho < 1$.

Рассмотрим коэффициенты λ^L, λ^U для архимедовых копула-функций. Так, для копула-функции *Гумбеля* лишь верхняя хвостовая зависимость положительна: $\lambda^U = 2 - 2^{1/\alpha}$, а $\lambda^L = 0$; тогда как для копула-функции *Клейтона* положительна лишь нижняя хвостовая зависимость: $\lambda^U = 0$, $\lambda^L = 2^{-1/\alpha}$.

Следует отметить, что хвостовая зависимость для архимедовых копула-функций может быть представлена в терминах генераторов, обратных к ним функций и первых производных, см. (Joe, 1997; Nelsen, 2006). В работах (Coles et al., 1999; Poon et al., 2004) представлены робастные непараметрические оценки для коэффициентов хвостовой зависимости.

5.4. Эмпирические приложения в статистическом пакете *R*: меры зависимости

В пакете *R* использование модулей `copula` и `fcopulae` дает возможность вычислять упомянутые выше меры зависимости для множества типов копула-функций. Ниже представлены некоторые примеры (заинтересованным читателям рекомендуется ознакомиться с руководством к пакету *R*).

```
# EXAMPLES WITH ARCHIMEDEAN COPULAS AND THE "copula" PACKAGE
gumbel.cop <- gumbelCopula(3)
kendallsTau(gumbel.cop)
[1] 0.6666667
spearmansRho(gumbel.cop)
[1] 0.848167
tailIndex(gumbel.cop)
lower upper
0.000000 0.740079
# let us compute the sample versions
x <- rcopula(gumbel.cop, 200)
cor(x, method = "kendall")

[,1] [,2]
[1,] 1.0000000 0.6592965
[2,] 0.6592965 1.0000000
cor(x, method = "spearman")

[,1] [,2]
[1,] 1.0000000 0.8483507
[2,] 0.8483507 1.0000000
# compare with the true parameter value 3
calibKendallsTau(gumbel.cop, cor(x, method="kendall")[1,2])
[1] 2.935103
calibSpearmansRho(gumbel.cop, cor(x, method="spearman")[1,2])
[1] 3.001893
calibSpearmansRho(gumbel.cop, cor(x, method="spearman")[1,2])
[1] 3.001893
# EXAMPLES WITH ARCHIMEDEAN COPULAS AND THE "fcopulae" PACKAGE
# ellipticalTau Computes Kendall's tau for elliptical copulae
# ellipticalRho computes Spearman's rho for elliptical copulae
ellipticalTau(rho = -0.5)

Tau
-0.3333333 attr(,"control")

rho
-0.5
```

```
ellipticalRho(rho = 0.75, type = "t", subdivisions = 100)

Rho
0.71
attr(,"control")

rho   type   tau

"0.75"   "t"   "0.5399"
# ellipticalTailCoeff
# Student-t Tail Coefficient:
ellipticalTailCoeff(rho = 0.25, param = 3, type = "t")

lambda
0.1962612
attr(,"control")

rho   type   param.nu

"0.25"   "t"   "3"
```

6. Процедуры оценивания: параметрические методы

Пусть $x_{1t},\ldots,x_{nt}$, $t=1,\ldots,T$ — многомерные наблюдения, где n — размерность наблюдаемой случайной величины, а T — число имеющихся наблюдений.

6.1. Метод максимального правдоподобия (ММП или одношаговый метод)

Пусть $f(\cdot)$ — плотность совместного распределения случайного вектора $(X_1, X_2,\ldots,X_n)$.

$$f(\tilde{x}_{1.},\ldots,\tilde{x}_{n.};\alpha_1,...,\alpha_n,\gamma)=c(F_1(\tilde{x}_{1.};\alpha_1),...,F_1(\tilde{x}_{n.};\alpha_n);\gamma)\cdot\prod_{i=1}^{n} f_i(\tilde{x}_{i.};\alpha_i),$$

где f_i — плотность одномерного частного распределения X_i, $\tilde{x}_1,\tilde{x}_2,\ldots,\tilde{x}_n$ — некоторые заданные (текущие) значения случайных величин $X_1, X_2,\ldots,X_n$ соответственно, а $c(\cdot)$ — плотность копула-функции, определяемая следующим соотношением:

$$c(u_1,...,u_n;\gamma)=\frac{\partial^n C(u_1,u_2,...,u_n;\gamma)}{\partial u_1 \partial u_2 ... \partial u_n}.$$

Пусть $\theta = (\alpha_1, ..., \alpha_n; \gamma)$ — вектор оцениваемых параметров, где α_i, $i = 1, ..., n$ — параметры частных распределений F_i, γ — вектор параметров копула-функции. Логарифмическая функция правдоподобия имеет следующий вид:

$$l(\theta) = \sum_{t=1}^{T} \ln(c(F_1(x_{1,t}; \alpha_1), \ldots, F_n(x_{n,t}; \alpha_n); \gamma)) + \sum_{t=1}^{T} \sum_{i=1}^{n} \ln f_i(x_{i,t}; \alpha_{i,t}) .$$

Оценка максимального правдоподобия $\hat{\theta}_{ML}$ для параметров определяется как

$$\hat{\theta}_{ML} = \arg\max_{\theta} l(\theta) .$$

Ниже приводится результат, описывающий асимптотические свойства оценки максимального правдоподобия.

Теорема 6 (*состоятельность и асимптотическое распределение ММП-оценок). Пусть* $\hat{\theta}_{ML}$ *— ММП-оценка. Тогда при выполнении некоторых стандартных условий регулярности, см.* (*White, 1994*)*, справедливо следующее:* $\hat{\theta}_{ML} \xrightarrow{P} \theta_0$*, где* $\theta_0 = (\alpha_1, \ldots, \alpha_n, \gamma)'$*. Кроме того,*

$$\sqrt{T}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N\left(0, I^{-1}(\theta_0)\right),$$

где $I(\theta_0)$ — информационная матрица Фишера.

Доказательство см. в (White, 1994) или (Gourieroux, Monfort, 1995).

6.2. Двухшаговый метод максимального правдоподобия[3]

В соответствии с этим методом параметры частных распределений оцениваются независимо от параметров копула-функции. Другими словами, процедура оценивания разделена на следующие два шага:

Шаг 1: *Используя ММП, оцениваются параметры* α_i, $i = 1, ..., n$, *частных одномерных распределений* $F_i(\cdot)$*:*

$$\hat{\alpha}_i = \arg\max_{\alpha} l^i(\alpha) = \arg\max_{\alpha} \sum_{t=1}^{T} \ln f_i(x_{i,t}; \alpha),$$

где l^i *— логарифмическая функция правдоподобия для частного распределения* $F_i(\cdot)$.

Шаг 2: *Оцениваются параметры* γ *копула-функции при заданных значениях оценок, полученные на шаге 1:*

$$\hat{\gamma} = \arg\max_{\gamma} l^c(\gamma) = \arg\max_{\gamma} \sum_{t=1}^{T} \ln(c(F_1(x_{1,t}; \hat{\alpha}_1), \ldots, F_n(x_{n,t}; \hat{\alpha}_n); \gamma)),$$

где l^c *— логарифмическая функция правдоподобия для копула-функции.*

[3] В англоязычном оригинале этот метод назван The Inference Functions for Margins method (IFM-method).

Ниже (см. теоремы 7–9) описаны асимптотические свойства оценок, получаемых IFM-методом. Эти результаты представляют собой простые обобщения результатов, полученных для ММП-процедуры, см. (Newey, McFadden, 1994; White, 1994; Patton, 2006a, b).

Представленные ниже результаты получены в предположении некоторых стандартных условий регулярности, которые мы не будем конкретизировать, заинтересованный читатель может ознакомиться с ними в работе (White, 1994). Из соображений простоты рассматривается двумерный случай. Отметим, что обобщения на многомерный случай получаются достаточно просто.

Теорема 7 (состоятельность $\hat{\alpha}_1$ и $\hat{\alpha}_2$). *Если* $n \to \infty$, *то* $\hat{\alpha}_1 \xrightarrow{P} \alpha_1$ *и* $\hat{\alpha}_2 \xrightarrow{P} \alpha_2$.

Доказательство см. в (White, 1994, Theorem 3.13).

Теорема 8 (состоятельность параметра γ). *Пусть функция*

$$n^{-1}\sum_{t=1}^{T}\ln c(F_1(x_{1,t};\hat{\alpha}_1),F_2(x_{2,t};\hat{\alpha}_2);\gamma)$$

имеет единственный максимум в точке $\gamma_0 \in \operatorname{int}(\Gamma)$. *Тогда* $\hat{\gamma} \xrightarrow{P} \gamma_0$ *при* $n \to \infty$, *где* Γ — *область допустимых значений параметра* γ, *а* $\operatorname{int}(\Gamma)$ — *внутренность этого множества.*

Доказательство см. в (White, 1994, Theorem 3.10).

Перед формулировкой теоремы 9 введем ряд обозначений. Определим штрафы для оценки $\hat{\theta}_{IFM} = \left(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\gamma}\right)'$:

$$s_{1t} = \frac{\partial}{\partial\alpha_1}\ln f_1(x_{1,t};\alpha_1), \quad s_{2t} = \frac{\partial}{\partial\alpha_2}\ln f_2(x_{2,t};\alpha_2), \quad s_{3t} = \frac{\partial}{\partial\gamma}\ln c(F_1(x_{1,t};\alpha_1),F_2(x_{2,t};\alpha_2);\gamma).$$

Матрица Гесса (гессиан) $Hess(\theta)$ для штрафов равна:

$$\begin{pmatrix} T^{-1}\sum_{t=1}^{T}\left(\frac{\partial^2 \ln f_1(x_{1,t};\alpha_1)}{\partial\alpha_1^2}\right) & 0 & 0 \\ 0 & T^{-1}\sum_{t=1}^{T}\left(\frac{\partial^2 \ln f_2(x_{2,t};\alpha_2)}{\partial\alpha_2^2}\right) & 0 \\ T^{-1}\sum_{t=1}^{T}\left(\frac{\partial^2 \ln c(F_1(\cdot),F_2(\cdot);\gamma)}{\partial\alpha_1\partial\gamma}\right) & T^{-1}\sum_{t=1}^{T}\left(\frac{\partial^2 \ln c(F_1(\cdot),F_2(\cdot);\gamma)}{\partial\alpha_2^2}\right) & T^{-1}\sum_{t=1}^{T}\left(\frac{\partial^2 \ln c(F_1(\cdot),F_2(\cdot);\gamma)}{\partial\gamma^2}\right) \end{pmatrix}.$$

Матрица OPG определяется как

$$\begin{pmatrix} T^{-1}\sum_{t=1}^{T} s_{1t}\cdot s_{1t}' & T^{-1}\sum_{t=1}^{T} s_{1t}\cdot s_{2t}' & T^{-1}\sum_{t=1}^{T} s_{1t}\cdot s_{3t}' \\ T^{-1}\sum_{t=1}^{T} s_{2t}\cdot s_{1t}' & T^{-1}\sum_{t=1}^{T} s_{2t}\cdot s_{2t}' & T^{-1}\sum_{t=1}^{T} s_{2t}\cdot s_{3t}' \\ T^{-1}\sum_{t=1}^{T} s_{3t}\cdot s_{1t}' & T^{-1}\sum_{t=1}^{T} s_{3t}\cdot s_{2t}' & T^{-1}\sum_{t=1}^{T} s_{3t}\cdot s_{3t}' \end{pmatrix}.$$

Теорема 9 (асимптотика оценок, получаемых IFM-методом). *Для IFM-оценок выполняется свойство асимптотической нормальности, а именно*

$$\sqrt{T}(\hat{\theta}_{IFM} - \theta_0) \xrightarrow{d} N\left(0, H_0^{-1} B_0 \left(H_0^{-1}\right)'\right),$$

где $H_0 = \mathbf{E}\left[Hess(\theta_0)\right]$, $B_0 = \mathbf{E}\left[OPG(\theta_0)\right]$, *т. е.* H_0 *и* B_0 *— средние значения для гессиана штрафов и матрицы OPG. Матрица* $H_0^{-1} B_0 (H_0^{-1})'$ *известна в теории копула-функций как информационная матрица Годамби* (*Joe, Xu, 1996; Joe, 1997*).

Из теорем 7–9 вытекает следующий весьма интересный факт.

Равная эффективность IFM-метода и ММП. Пусть x_{1t} и x_{2t} независимы (т. е. совместное распределение (x_{1t}, x_{2t}) имеет копула-функцию $C(u_1, u_2; \gamma) = u_1 \cdot u_2$, плотность которой $c(u_1, u_2; \gamma) = 1$), и для плотностей частных распределений выполнено информационное равенство. Тогда IFM-метод и ММП дают оценки, имеющие одинаковые асимптотические распределения.

6.3. Эмпирические приложения в статистическом пакете *R*: процедуры параметрического оценивания

Модуль `copula` в статистическом пакете *R* позволяет получать как ММП-, так и IFM-оценки. Ниже представлен пример, взятый из работы (Yan, 2007). Вначале генерируется выборка из двумерного распределения, частные распределения которого — гамма-распределения с параметрами $\theta_\gamma = (2,1,3,2)$, а копула-функция — нормальная копула-функция с параметром $\alpha = 0.5$; затем по сгенерированной выборке строятся ММП- и IFM-оценки параметров.

```
# One step EML example from Yan (2007)
myMvd <- mvdc(copula=ellipCopula(family="normal", param=0.5),

margins=c("gamma", "gamma"),

paramMargins=list(list(shape=2, scale=1),

list(shape=3, scale=2)))
n <- 200
dat <- rmvdc(myMvd, n)
loglikMvdc(c(2,1,3,2,0.5), dat, myMvd)
[1] -778.3114
mm <- apply(dat, 2, mean)
vv <- apply(dat, 2, var)
b1.0 <- c(mm[1]^2 / vv[1], vv[1] / mm[1])
b2.0 <- c(mm[2]^2 / vv[2], vv[2] / mm[2])
a.0 <- sin(cor(dat[,1], dat[,2], method = "kendall") * pi / 2)
```

```
start <- c(b1.0, b2.0, a.0)
fit.eml <- fitMvdc(dat, myMvd, start=start, optim.control=list(trace=TRUE,
maxit=2000))
fit.eml
The Maximum Likelihood estimation is based on 200 observations.
Margin 1 :

Estimate Std. Error
m1.shape 2.1279360 0.19842701
m1.scale 0.9246803 0.09716286
Margin 2 :

Estimate Std. Error
m2.shape 3.213331 0.3061706
m2.scale 1.778994 0.1835066
Copula:

Estimate Std. Error
rho.1 0.4671964 0.0552821
The maximized loglikelihood is -777.1381
The convergence code is 0 see ?optim.
# IFM method
loglik.marg <- function(b, x) sum(dgamma(x, shape=b[1], scale=b[2],
log=TRUE))
ctrl <- list(fnscale = -1)
b1hat <- optim(b1.0, fn=loglik.marg, x=dat[,1], control=ctrl)$par
b2hat <- optim(b2.0, fn=loglik.marg, x=dat[,2], control=ctrl)$par
udat <- cbind(pgamma(dat[,1], shape=b1hat[1], scale=b1hat[2]),
pgamma(dat[,2], shape=b2hat[1], scale=b2hat[2]))
fit.ifm <- fitCopula(myMvd, udat, method="ml",start=a.0)
c(b1hat, b2hat, fit.ifl@estimate) fit.ifl
[1] 2.3494821 0.9395788 3.1645483 2.0540858 0.4782428
fit.ifm
The estimation method is Maximum Likelihood based on 200 observations.

Estimate Std. Error z value Pr(>|z|)
rho.1 0.4782428 0.04920972 9.718461 0
The maximized loglikelihood is 25.95508
The convergence code is 0
```

7. Процедуры оценивания: полупараметрические и непараметрические методы

7.1. Канонический метод максимума правдоподобия (КММП)[4]

КММП отличается от описанных выше методов тем, что в нем не делается никаких предположений относительно параметрической формы частных распределений. Процесс оценивания в КММП состоит из следующих двух шагов.

1. Преобразуем наблюдения $(x_{1t}, x_{2t}, \ldots, x_{nt})$, $t = 1, \ldots, T$ следующим образом:

$$\hat{u}_{it} = F_{i,T}(x_{it}),$$

где $F_{i,T}(\cdot) = \frac{1}{T+1}\sum_{t=1}^{T}\mathbf{1}_{\{x_{it} \le \cdot\}}$, $i = 1, \ldots, n$, а $\mathbf{1}_{\{\cdot\}}$ — индикаторная функция. Другими словами, $F_{i,T}(x)$ — это непараметрическая оценка частной функции распределения переменной X_i.

2. Оценим параметры копула-функции, максимизируя логарифмическую функцию правдоподобия[5]:

$$\hat{\gamma}_{CML} = \arg\max_{\gamma} \sum_{t=1}^{T} \ln c(F_{1,T}(x_{1t}), \ldots, F_{n,T}(x_{nt}); \gamma).$$

Для простоты при описании асимптотических свойств КММП-оценки снова будем рассматривать только двумерный случай. Пусть $l(u_1, u_2, \gamma) = \ln c(u_1, u_2; \gamma)$. Далее индексами $1, 2, \gamma$ будем обозначать частные производные функции $l(u_1, u_2, \gamma)$ относительно u_1, u_2, γ соответственно. Кроме того, положим

$$S_T = \frac{1}{T}\sum_{t=1}^{T} l_{\gamma}\left(F_{1,T}(x_{1t}), F_{2,T}(x_{2t}); \gamma\right), \qquad H_T = \frac{1}{T}\sum_{t=1}^{T} l_{\gamma\gamma}\left(F_{1,T}(x_{1t}), F_{2,T}(x_{2t}); \gamma\right).$$

Поскольку $\hat{\gamma}_{CML}$ является решением следующего дифференциального уравнения:

$$S_T = \frac{1}{T}\sum_{t=1}^{T} l_{\gamma}\left(F_{1,T}(x_{1t}), F_{2,T}(x_{2t}); \gamma\right) = 0,$$

то, используя разложение функции S_T в ряд Тейлора, получим:

$$0 = \frac{1}{T}\sum_{t=1}^{T} l_{\gamma}\left(F_{1,T}(x_{1t}), F_{2,T}(x_{2t}); \gamma\right)\Bigg|_{\gamma=\hat{\gamma}_{CML}} \simeq S_T + H_T(\hat{\gamma}_{CML} - \gamma),$$

и, следовательно,

$$\sqrt{T}(\hat{\gamma}_{CML} - \gamma) \simeq \sqrt{T}\,\frac{S_T}{-H_T}.$$

---

[4] В англоязычном оригинале *The Canonical Maximum Likelihood (CML)*.

[5] В дальнейшем предполагается, что анализируемое семейство копул зависит от единственного параметра g.

Из работ (Ruymgaart et al., 1972; Genest et al., 1995; Ruud, 2000) отметим, в частности, следующие результаты, справедливые при соблюдении некоторых условий регулярности:

а) $\hat{\gamma}_{CML} \to \gamma_0$ (п. н.) при $T \to \infty$;

б) $\sqrt{T}(\hat{\gamma}_{CML} - \gamma_0) \xrightarrow{d} N\left(0, \sigma^2 h^{-2}\right)$, где

$\sigma^2 = \mathrm{Var}\left[l_\gamma\left(F_1(x_{1t}), F_2(x_{2t}); \gamma\right) + W_1(x_{1t}) + W_2(x_{2t})\right]$, $h = \mathrm{E}\left[l_{\gamma\gamma}\left(F_1(x_{1t}), F_2(x_{2t}); \gamma\right)\right]$;

выражение для σ^2 дано здесь для *двумерного* случая ($n = 2$ и, соответственно, $i = 1, 2$), а $W_i(x_{it})$ определяются соотношениями:

$$W_i(x_{it}) = -\int \mathbf{1}_{\{F_i(x_{it}) \le u_i\}}\, l_{\gamma i}\left(u_{1t}, u_{2t}; \gamma\right) c\left(u_{1t}, u_{2t}; \gamma\right) du_1 du_2;$$

в) полупараметрическая оценка $\hat{\gamma}_{CML}$ менее эффективна, чем оценка, полученная с помощью одношагового метода максимального правдоподобия $\hat{\gamma}_{ML}$ (за исключением случая, когда x_{1t} и x_{2t} статистически независимы).

7.2. Трехшаговый канонический метод максимального правдоподобия (KME–CML метод)[6]

Начиная с момента выхода работы (Genest et al., 1995), использование полупараметрических методов для копула-функций эллиптического типа стало обычной практикой, см., например, (Cherubini et al., 2004; McNeil et al., 2005). В частности, для копула-функций Стьюдента, после того, как на первом шаге получены непараметрические оценки частных функций распределения, можно получить оценку корреляционной матрицы, вычисленную с использованием метода моментов и оценок ранговых коэффициентов корреляции τ-Кендалла. При этом степени свободы оцениваются с использованием метода максимального правдоподобия. Отметим, что оценка частных распределений на первом шаге не является необходимым условием для оценки корреляционной матрицы, поскольку оценка τ-Кендалла основана на количестве согласованных наблюдений, которое не зависит от монотонных преобразований. Однако оценивание частных распределений является необходимым при оценке числа степеней свободы.

Ниже дается описание трехшагового KME–CML метода (см. также (Bouyé et a1., 2000; McNeil et a1., 2005; Fantazzini, 2010)).

Трехшаговый KME–CML метод оценивания параметров копула-функций

1. Преобразуем $(x_{1t}, x_{2t}, \ldots, x_{nt})$ *к нормированным рангам* $(F_{1T}(x_{1t}), F_{2T}(x_{2t}), \ldots, F_{nT}(x_{nt}))$, *используя эмпирическую функцию распределения для каждой компоненты вектора наблюдений* (*здесь и далее n — размерность анализируемой случайной величины, T — общее число ее наблюдений*).

2. Для всех пар (j, k) *компонент вектора наблюдений оценим* τ*-Кендалла:*

$$\hat{R}^{\tau}_{jk} = \tau\left[F_{jT}(X_j), F_{kT}(X_k)\right] = (C_T^2)^{-1} \sum_{1 \le t < s \le T} \mathrm{sign}\left((x_{1t} - x_{1s})(x_{2t} - x_{2s})\right).$$

[6] В англоязычном оригинале метод называется *Kendall-t Moment — Estimator of Canonical Maximum Likelihood Method*.

Возьмем оценку корреляции для (j,k) *компонент вектора наблюдений* $\hat{\Sigma}_{jk} = \sin(\frac{\pi}{2}\hat{R}^{\tau}_{jk})$. *Поскольку покомпонентное преобразование, осуществленное на первом шаге, не гарантирует положительной определенности матрицы* $\hat{\Sigma} = \left(\hat{\Sigma}_{jk}\right)$, *то для ее обеспечения обычно вносятся некоторые поправки* (*Rousseeuw, Molenberghs, 1993*).

3. Находим оценку степеней свободы $\hat{\nu}_{CML}$ *копула-функций Стьюдента с использованием метода максимального правдоподобия:*

$$\hat{\nu}_{CML} = \arg\max_{\nu} \sum_{t=1}^{T} \ln c_{T-copula}\left(F_{1T}(x_{1t}),\ldots,F_{nT}(x_{nt});\hat{\Sigma},\nu\right).$$

Второй шаг описанной процедуры реализует метод моментов, основанный на оценках $q = n(n-1)/2$ моментов и такого же количества коэффициентов τ-Кендалла, оцененных с использованием эмпирических распределений (эта оценка известна как моментная оценка τ-Кендалла). Следует еще раз отметить, что нет необходимости использовать оценку τ-Кендалла, полученную с использованием именно эмпирических распределений, т. к. τ-Кендалла сохраняет одно и то же значение при всех монотонных преобразованиях. Таким образом, можно взять q моментов от некоторой векторнозначной функции ψ, зависящей от корреляций $\theta_0 = (\rho_1,\ldots,\rho_q)'$, а именно:

$$\psi\left(F_1,\ldots,F_n;\theta_0\right) = \begin{pmatrix} \mathrm{E}\left[\psi_1(F_1,F_2);\ \rho_1(\theta)\right] \\ \vdots \\ \mathrm{E}\left[\psi_q(F_{n-1},F_n);\ \rho_q(\theta)\right] \end{pmatrix}.$$

Тогда теоретические моментные тождества могут быть представлены в виде

$$\psi(F_1,\ldots,F_n;\theta_0) = (0,0,\ldots,0)', \tag{20}$$

где θ_0 — истинное значение оцениваемого векторного параметра. Соответственно, оценка $\hat{\theta}$ определится из (20) с заменой теоретических моментов их выборочными аналогами.

Для оценки, полученной с помощью метода KME–CML, справедливы следующие теоремы (все доказательства представлены в приложении А работы (Fantazzini, 2010)).

Теорема 10 (состоятельность $\hat{\theta}$). *Предположим, что* $(x_{1t},\ldots,x_{nt})$ *— независимые* (*по t*) *наблюдения n-мерной случайной величины* $(X_1,\ldots,X_n)$ *со структурой зависимости, заданной плотностью копула-функции* $c(u_{1,t},\ldots,u_{n,t};\Sigma_0,\nu_0)$. *Предположим, что:*

а) пространство параметров Θ *является компактным подмножеством* $\mathbb{R}^q$;

б) q-мерный вектор моментов $\psi\left(F_1(X_1),\ldots,F_n(X_n);\theta_0\right)$ *непрерывен относительно* θ_0 *для любых* X_j;

в) $\psi\left(F_1(X_1),\ldots,F_n(X_n);\theta\right)$ *измеримы относительно* X_j *для всех* $\theta \in \Theta$;

г) $\psi\left(F_1(X_1),\ldots,F_n(X_n);\theta\right) \neq 0$ *для всех* $\theta \neq \theta_0, \theta \in \Theta$;

д) $\sup_{\theta\in\Theta}\left\|\psi\left(F_1(X_1),\ldots,F_n(X_n);\theta\right)\right\| < \infty$, *где* $\|A\|$ *обозначает евклидову норму матрицы A.*

Тогда $\hat{\theta} \xrightarrow{P} \theta_0$ *при* $T \to \infty$.

При условиях теоремы 10, дополненных некоторыми условиями регулярности (Genest et al., 1995), доказана и состоятельность оценки $\hat{\nu}_{CML}$ (т. е. $\hat{\nu}_{CML} \xrightarrow{P} \nu_0$ при $T \to \infty$).

Асимптотическая нормальность оценок, полученных методом KME–CML, не очевидна, т. к. используется трехшаговая процедура, при которой на втором и третьем шагах используются разные методы оценивания. Возможным решением является представление оценки на третьем шаге КММП, как оценки специального вида, полученной методом моментов. Отметим, что КММП-оценка определяется путем приравнивания к нулю производной (относительно параметра ν) логарифма функции правдоподобия:

$$\frac{\partial l(\cdot;\nu)}{\partial \nu} = \sum_{t=1}^{T} l_\nu \left(F_{1T}(x_{1t}),\ldots,F_{nT}(x_{nt});\hat{\Sigma},\hat{\nu} \right) = 0.$$

Разделив обе части на T, получим определение оценки по методу моментов:

$$\frac{1}{T}\sum_{t=1}^{T} l_\nu \left(F_{1T}(x_{1t}),\ldots,F_{nT}(x_{nt});\hat{\Sigma},\hat{\nu} \right) = \frac{1}{T}\sum_{t=1}^{T} \psi_\nu (F_{1T}(x_{1t}),\ldots,F_{nT}(x_{nt});\hat{\Sigma},\hat{\nu}) = 0 .$$

Таким образом, КММП-оценка может быть представлена как оценка, получаемая методом моментов (ММ).

Оценка методом моментов — это то значение θ, которое обеспечивает равенство между выборочными моментами (выборочные средние) и их теоретическими аналогами. Поскольку число параметров равно q, то требуются q моментных уравнений с неизвестными $\hat{\theta}$ (более подробно см. (Greene, 2002)). Таким образом, можно использовать известные асимптотические результаты, справедливые для метода моментов.

Определим выборочные моменты вектора $\Psi_{KME-CML}$, зависящего от параметров $\hat{\Xi} = (\hat{\rho}_1,\ldots\hat{\rho}_q,\hat{\nu})'$:

$$\Psi_{KME-CML}\left(F_{1T}(x_{1t}),\ldots,F_{nT}(x_{nt});\hat{\Xi} \right) = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} \psi_1 \left(F_{1T}(x_{i1}), F_{2T}(x_{i2});\ \hat{\rho}_1 \right) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \frac{1}{T}\sum_{t=1}^{T} \psi_q \left(F_{(n-1),T}(x_{(n-1)t}), F_{nT}(x_{nt});\ \hat{\rho}_q \right) \\ \frac{1}{T}\sum_{t=1}^{T} \psi_\nu \left(F_{1T}(x_{1t}),\ldots,F_{nT}(x_{nt});\hat{\Sigma},\hat{\nu} \right) \end{pmatrix} = 0. \qquad (21)$$

Для определения ковариационной матрицы предельного распределения оценки $\hat{\Xi}$ необходимо подкорректировать (21) с учетом непараметрических оценок частных распределений и их дисперсий, см. (Genest et al., 1995, §4), а именно

$$\Delta_0 = \begin{pmatrix} \psi_1 \left(F_1(X_1), F_2(X_2);\ \rho_1 \right) \\ \vdots \\ \psi_q \left(F_{n-1}(X_{n-1}), F_n(X_n);\ \rho_q \right) \\ \psi_\nu \left(F_1(X_1),\ldots,F_n(X_n);\Sigma_0,\nu_0 \right) + \sum_{j=1}^{n} W_{j,\nu}(X_j) \end{pmatrix} = 0 ,$$

$$\Upsilon_0 \equiv \mathrm{Var}\left[\Delta_0\right] = \mathbf{E}\left[\Delta_{KME-CML}\, \Delta'_{KME-CML} \right],$$

где

$$W_{j,\nu}(X_j) = \int 1_{\{F_j(X_j)\le u_j\}} \frac{\partial^2}{\partial \nu \partial u_j} \ln c(u_1,\ldots,u_n) dC(u_1,\ldots,u_n).$$

С учетом вышеизложенного справедлива следующая теорема.

Теорема 11 (асимптотическое распределение KME–CML оценок). *Пусть выполнены предположения предыдущей теоремы и дополнительные условия из* (*Genest et al., 1995*). *Кроме того, предположим, что матрица* $\frac{\partial \Psi_{KME-CML}(\cdot;\Xi)}{\partial \Xi'}$ *имеет ограниченные (по T) элементы и является отрицательно определенной, тогда как* Υ_0 *также имеет ограниченные (по T) элементы, но положительно определена.*

Тогда при $T \to \infty$:

$$\sqrt{T}(\hat{\Xi} - \Xi_0) \xrightarrow{d} N\left\{0, \mathbf{E}\left[\frac{\partial \Psi_{KME-CML}}{\partial \Xi'}\right]^{-1} \Upsilon_0 \left(\mathbf{E}\left[\frac{\partial \Psi_{KME-CML}}{\partial \Xi'}\right]^{-1}\right)'\right\}. \quad (22)$$

Отметим, что последнее асимптотическое свойство выполняется и для многомерных моделей с *гетероскедастичностью*. Для таких моделей на первом этапе получают состоятельные оценки параметров одномерных частных распределений, а затем рассчитывают соответствующие оценки остатков (так называемые «невязки») которые используются для оценки совместного распределения многомерного вектора остатков при помощи копула-функций. Этот результат является прямым следствием теорем 1 и 2 из работы (Kim et al., 2008), которая, в свою очередь, использует работы (Koul, Ling, 2006; Koul, 2002). Отметим, что похожие результаты представлены также и в (Chen, Fan, 2006), но без доказательств. Соответствующий результат формулируется в теореме 12.

Теорема 12 (асимптотика KME–CML оценок для многомерных моделей с гетероскедастичностью). *Предположим, что выполнены условия а)–д) теоремы 10, а также A.1–A.9 из (Kim et al., 2008). Тогда для KME–CML оценки имеет место асимптотическая сходимость* (22).

Условия (A.1)–(A.4) справедливы для моделей достаточно общего вида. В частности, условия (A.1)–(A.2) требуют, чтобы плотность копула-функций имела непрерывные частные производные до третьего порядка включительно. Кроме того, они должны быть конечными вместе с их вторым моментом. Условие (A.3) — это техническое условие на частные производные, а условие (A.4) требует, чтобы были выполнены условия (A.1) из (Genest et al., 1995). Условия (A.5)–(A.8) также являются техническими и взяты из работ (Koul, 2002) и (Koul, Ling, 2006). Условие (A.9) не является слишком ограничительным. Например, если условное среднее является функцией исторических значений строго стационарных и эргодических временных рядов, то каждое слагаемое суммы является строго стационарным и эргодическим процессом с нулевым средним, а значит, условие (A.9) выполняется (см. (Taniguchi, Kakizawa; 2000, теоремы 1.3.3–1.3.5), а также (Kim et a1., 2008)).

Свойства и вычислительные аспекты в условиях малых выборок

Поскольку основные свойства предлагаемого полупараметрического метода являются асимптотическими, то на примерах смоделированных выборок в работе (Fantazzini, 2010) было проведено исследование, показавшее, что в условиях малых выборок и больших зна-

чений ν численная максимизация логарифмической функции правдоподобия не сходилась значительно чаще при использовании КМЕ–CML оценок, чем при использовании ММП оценок. В то время как ММП-оценка параметра попадала в 95% доверительный интервал примерно для 95% случаев, КМЕ–CML оценка попадала в соответствующий доверительный интервал лишь в 30% случаев. Однако такое снижение доли случаев со сходимостью более значительно для двумерных копула-функций Стьюдента по сравнению с копула-функциями Стьюдента большей размерности, которые типичны для финансовых портфелей. Кроме того, как ММП, так и КМЕ–CML метод показали весьма существенные значения средних и медиан смещений для оцененных корреляций, когда фактические значения последних принимают значения, близкие к нулю.

Наконец, в работе (Fantazzini, 2010) показано, что метод собственных значений, представленный в (Rousseeuw, Molenberghs, 1993), должен быть использован для получения положительно определенной корреляционной матрицы не только в условиях малых выборок ($T < 100$), но и в условиях, когда описываемому процессу отвечает наименьшее собственное значение, близкое к нулю. Эта поправка влечет положительное смещение оценки параметра ν, но ее влияние на сходимость при максимизации для получения параметра ν весьма ограничено.

Таким образом, предыдущие результаты показывают, что КМЕ–CML метод может быть использован в случае малых выборок и при относительно небольших значениях числа степеней свободы, однако ММП является более привлекательной альтернативой. Возможной стратегией при оценке параметров является следующее: на первом этапе использовать КМЕ–CML метод; если оценка числа степеней свободы больше 20, то следует использовать ММП-оценку, если последняя обеспечивает сходимость при максимизации соответствующей функции правдоподобия. Иначе, в качестве альтернативного решения следует использовать нормальную копула-функцию, к которой стремится копула-функция Стьюдента при $\nu \to \infty$ (нормальная и стьюдентовская копула-функции достаточно близки уже при $\nu > 20$).

7.3. Методы непараметрического оценивания

В работе (Fermanian, Scaillet, 2003) предлагается непараметрическая оценка для копула-функций многомерных стационарных процессов, удовлетворяющих условиям сильного перемешивания.

Перед тем, как обратиться к асимптотическим свойствам этой оценки, определим понятие ядерной функции.

Определение 10. *Ядерная функция — это действительнозначная функция* $K(u)$, $-\infty < u < +\infty$, *которая удовлетворяет следующим условиям:*

1) *симметричность:* $K(u) = K(-u)$;
2) *в точке* $u = 0$ *достигается максимум*;
3) $\int_{-\infty}^{+\infty} K(u)du = 1$;
4) *неотрицательность:* $K(u) \geq 0$;
5) $K(u)$ *ограничена*;
6) $\int_{-\infty}^{\infty} u^2 K(u)du < \infty$.

Свойства 5 и 6 необходимы для вывода асимптотических свойств оценок, получаемых на основе ядерных функций. Возьмем n-мерное ядро и его кумулятивную функцию:

$$k(\mathbf{x}) = \prod_{j=1}^{n} k_j(x_j),\ K(\mathbf{X}) = \prod_{j=1}^{n} \int_{-\infty}^{x_j} k_j(x)dx = \prod_{j=1}^{n} K_j(x_j),$$

где для простоты взяты произведения одномерных ядерных функций. Кроме того, возьмем

$$k(\mathbf{x};\mathbf{h}) = \prod_{j=1}^{n} k_j\left(\frac{x_j}{h_j(T)}\right),\ K(\mathbf{X};\mathbf{h}) = \prod_{j=1}^{n} K_j\left(\frac{x_j}{h_j(T)}\right),$$

где $\mathbf{h}(T)$ — диагональная матрица с элементами $h_j(T),\ j = 1,\ldots,n$ на главной диагонали и детерминантом $|\mathbf{h}(T)|$. При этом $h_j(T)$ положительна и $h_j \to 0$ при $T \to \infty$. $\mathbf{h}(T)$ обычно называют «шириной окна» ядерной функции.

Определение 11. *Ядерная оценка $\hat{f}_j(\cdot)$ плотности $f_j(\cdot)$ распределения случайной величины Y_{jt} в точке y_j определяется как*

$$\hat{f}_j(y_j) = \frac{1}{Th_j}\sum_{i=1}^{T} k_j\left(\frac{y_j - Y_{jt}}{h_j(T)}\right),$$

а ядерная оценка плотности случайного вектора $\mathbf{Y}_t$ в точке $\mathbf{y} = (y_1,\ldots,y_n)'$ равна

$$\hat{f}(\mathbf{y}) = \frac{1}{T\,|\,h(T)\,|}\sum_{i=1}^{T} k\left(\mathbf{y} - \mathbf{Y}_t;\mathbf{h}(T)\right).$$

Аналогичным образом, ядерная оценка кумулятивной функции распределения случайной величины Y_{jt} в точке y_j равна

$$\hat{F}_j(y_j) = \int_{-\infty}^{y_j} \hat{f}_j(x)dx,$$

а ядерная оценка кумулятивной функции распределения случайного вектора $\mathbf{Y}_t$ в точке $\mathbf{y}$ *равна:*

$$\hat{F}(\mathbf{y}) = \int_{-\infty}^{y_1}\ldots\int_{-\infty}^{y_n} \hat{f}(\mathbf{x})d\mathbf{x}.$$

Например, для гауссовской ядерной функции $k_j(x) = \varphi(x)$ получаются следующие ядерные оценки функций распределения:

$$\hat{F}_j(y_j) = \frac{1}{T}\sum_{i=1}^{T}\Phi\left(\frac{y_j - Y_{jt}}{h_j}\right),\qquad \hat{F}(\mathbf{y}) = \frac{1}{T}\sum_{i=1}^{T}\prod_{j=1}^{n}\Phi\left(\frac{y_j - Y_{jt}}{h_j}\right),$$

где через $\varphi(\cdot)$ и $\Phi(\cdot)$ обозначены соответственно плотность и функция распределения стандартного нормального распределения.

Определение 12 (ядерная оценка копула-функции). *В силу следствия из теоремы Склара* $C(u) = C(u_1,\ldots,u_n) = H\left(F_1^{(-1)}(u_1),\ldots,F_n^{(-1)}(u_n)\right)$, *где* $H(\cdot)$ *— многомерная функция распределения. Тогда ядерная оценка* $\widehat{C}(\mathbf{u})$ *копула-функции* $C(\mathbf{u})$ *определяется следующим образом:*

$$\widehat{C}(\mathbf{u}) = \widehat{H}(\hat{\xi}) = \widehat{H}\left(\hat{\xi}_1,\ldots,\hat{\xi}_n\right),$$

где $\hat{\xi}_j = \inf_{y\in\mathbb{R}}\left\{y:\ \widehat{F}_j(y) \geq u_j\right\}$ *соответствует ядерной оценке квантиля уровня* u_j *для функции распределения случайной величины* Y_{jt}.

В работах (Fermanian, Scaillet, 2003; van der Vaart, Wellner, 1996) доказывается состоятельность и асимптотическая нормальность ядерных оценок $\widehat{C}(\mathbf{u})$.

7.4. Эмпирические приложения в статистическом пакете *R*: процедуры полупараметрического оценивания

Продолжим рассмотрение примера, начатое в п. 6.3 (см. также (Фантаццини, 2011)). Можно оценить параметр зависимости α при помощи канонического метода максимума правдоподобия (КММП) без спецификации частных распределений.

```
# CML method
eu <- cbind((rank(dat[,1]) - 0.5)/n, (rank(dat[,2]) - 0.5)/n)
fit.cml <- fitCopula(myMvd, eu, method="mpl", start=a.0)
fit.cml
The estimation method is Maximum Pseudo-Likelihood based on 200
observations.
Estimate Std. Error z value Pr(>|z|)
rho.1 0.4806303 0.05066396 9.48663 0
The maximized loglikelihood is 26.02639
The convergence code is 0
```

Ниже представлены другие примеры, взятые из `copula` пакета.

```
# Example: Gumbel copula
gumbel.cop <- gumbelCopula(3, dim=2)
n <- 200 x <- rcopula(gumbel.cop, n) ## true observations
u <- apply(x, 2, rank) / (n + 1) ## pseudo-observations
## inverting Kendall's tau
fit.tau <- fitCopula(gumbel.cop, u, method="itau")
fit.tau
## inverting Spearman's rho
fit.rho <- fitCopula(gumbel.cop, u, method="irho")
fit.rho
```

```
## maximum pseudo-likelihood
fit.mpl <- fitCopula(gumbel.cop, u, method="mpl")
fit.mpl
## maximum likelihood
fit.ml <- fitCopula(gumbel.cop, x, method="ml")
fit.ml
## A multiparameter example
normal.cop <- normalCopula(c(0.6,0.36, 0.6),dim=3,dispstr="un")
x <- rcopula(normal.cop, n) ## true observations
u <- apply(x, 2, rank) / (n + 1) ## pseudo-observations
## inverting Kendall's tau
fit.tau <- fitCopula(normal.cop, u, method="itau")
fit.tau
## inverting Spearman's rho
fit.rho <- fitCopula(normal.cop, u, method="irho")
fit.rho
## maximum pseudo-likelihood
fit.mpl <- fitCopula(normal.cop, u, method="mpl")
fit.mpl
## maximum likelihood
fit.ml <- fitCopula(normal.cop, x, method="ml")
fit.ml
```

Список литературы

Айвазян С. А. (2010). *Методы эконометрики*. М.: Магистр.

Фантаццини Д. (2011). Моделирование многомерных распределений с использованием копула-функций. I. *Прикладная эконометрика*, 2 (22), 98–134.

Aas K., Czado C., Frigessi A., Bakken H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44 (2), 182–198.

Barbe P., Genest C., Ghoudi K., Rremillard B. (1996). On Kendall's process. *Journal of Multivariate Analysis*, 58, 197–229.

Bedford T., Cooke R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32, 245–268.

Bedford T., Cooke R. M. (2002). Vines — a new graphical model for dependent random variables. *Annals of Statistics*, 30, 1031–1068.

Bouyé E., Durrleman V., Nikeghbali A., Riboulet G., Roncalli T. (2000). Copulas for finance: A reading guide and some applications. Groupe de Recherche Operationnelle, Credit Lyonnais, *Working Paper*.

Chen X., Fan Y. (2006). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of Econometrics*, 135, 125–154.

Cherubini U., Vecchiato W., Luciano E. (2004). *Copula methods in finance*. Wiley.

Clemen M. N., Jouini R. T. (1996). Copula models for aggregating expert opinions. *Operations Research*, 44, 444–457.

Coles S., Heffernan J., Tawn J. (1999). Dependence measures for extreme value analyses. *Extremes*, 2 (4), 339–365.

Conover W. J. (1999). *Practical nonparametric statistics*. 3rd ed. Wiley.

Embrechts P., McNeil A., Straumann D. (2002). Correlation and dependence in risk management: properties and pitfalls. In: M. A. H. Dempster (ed.), *Risk Management: Value at Risk and Beyond*. Cambridge: Cambridge University Press.

Fang K., Kotz S., Hg K. (1987). *Symmetric multivariate and related distributions*. London: Chapman Hall.

Fantazzini D. (2010). Three-stage semi-parametric estimation of t-copulas: Asymptotics, finite-sample properties and computational aspects. *Computational Statistics and Data Analysis*, forthcoming.

Ferguson S. T., Genest C., Hallin M. (2000). Kendall's tau for serial dependence. *Canadian Journal of Statistics*, 28, 587–604.

Fermanian J., Scaillet O. (2003). Nonparametric estimation of copulas for time series. *Journal of Risk*, 5, 25–54.

Genest C., MacKay J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. *American Statistics*, 40, 280–285.

Genest C., Ghoudi K., Rivest L. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82, 543–552.

Gourieroux C., Monfort A.. (1995). *Statistics and econometric models*. Cambridge: Cambridge University Press.

Greene W. (2002). *Econometric analysis*. Prentice Hall.

Hollander M., Wolfe D. A. (1973). *Nonparametric statistical inference*. New York: Wiley.

Hult H., Lindskog F. (2002). Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied Probability*, 34, 587–608.

Joe H., Xu J. (1996). The estimation method of inference functions for margins for multivariate models. *Technical Report* No. 166. Department of Statistics, University of British Columbia.

Joe H. (1997). *Multivariate models and dependence concepts*. London: Chapman Hall.

Kim G., Silvapulle M. J., Silvapulle P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics and Data Analysis*, 51 (6), 2836–2850.

Kim G., Silvapulle M. J., Silvapulle P. (2008). Estimating the error distribution in multivariate heteroscedastic time series models. *Journal of Statistical Planning and Inference*, 138 (5), 1442–1458.

Koul H. L. (2002). *Weighted empirical processes in dynamic nonlinear models*. Lecture Notes in Statistics, V. 166. Springer.

Koul H. L., Ling S. (2006). Fitting an error distribution in some heteroscedastic time series model. *Annals of Statistics*, 34, 994–1012.

Kruskal W. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53, 814–861.

Kurowicka D., Cooke R. M. (2006). *Uncertainty analysis with high dimensional dependence modelling*. New York: Wiley.

Lindskog F., McNeil A., Schmock U. (2003) Kendall's tau for elliptical distributions. In: G. Bol, G. Nakhaeizaden, S. Rachev, T. Ridder, K.-H. Vollmer (eds.), *Credit Risk. Measurement, Evaluation and Management*, Physica-Verlag, A Springer-Verlag Company, Heidelberg, 149–156.

Mardia K. V., Kent J., Bibby J. M. (1997). *Multivariate analysis*. San Diego: Academic Press.

McNeil A., Frey R., Embrechts P.. (2005). *Quantitative risk management: Concepts, techniques and tools*. Princeton Series in Finance, New Jersey.

Nelsen R. B. (1999). *An introduction to copulas*. Lecture Notes in Statistics. New York: Springer.

Nelsen R. B. (2006). *An introduction to copulas*. Lecture Notes in Statistics, 2nd Edition. New York: Springer.

Newey W. K., McFadden D. (1994) Large sample estimation and hypothesis testing. *Handbook of Econometrics*, V. 4, 2111–2245, Amsterdam: Elsevier Science B. V.

Oakes D. (1994). Multivariate survival distributions. *Journal of Nonparametric Statistics*, 3, 343–354.

Patton A. (2006a). Estimation of copula models for time series of possibly different lengths. *Journal of Applied Econometrics*, 21, 147–173.

Patton A. (2006b). Modelling asymmetric exchange rate dependence. *International Economic Review*, 47 (2), 527–556.

Poon S., Rockinger M., Tawn J. (2004). Extreme value dependence in financial markets: Diagnostics, models, and financial implications. *Review of Financial Studies*, 17, 581–610.

Resnick S. I. (1987). *Extreme values, regular variation and point processes.* New York: Springer.

Rodgers J. L., Nicewander W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42, 59–66.

Rousseeuw P., Molenberghs G. (1993). Transformation of non positive semidefinite correlation matrices. *Communications in Statistics — Theory and Methods*, 22, 965–984.

Ruud P. (2000). *An introduction to classical econometric theory*. Oxford University Press.

Ruymgaart F. H., Shorack G. R., Van Zwet W. R. (1972). Asymptotic normality of nonparametric tests for independence. *The Annals of Statistics*, 43, 1122–1135.

Shih J. H., Louis T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51, 1384–1399.

Sibuya M. (1961). Bivariate extreme statistics. *Annals of Mathematical Statistics*, 11, 195–210.

Statistics with confidence. (1989). Eds.: Gardner M. J., Altman D. G. London: BMJ.

Taniguchi M., Kakizawa Y. (2000). *Asymptotic theory of statistical inference for time series*. New York: Springer.

Van der Vaart A., Wellner J. (1996). *Weak convergence and empirical processes*. New York: Springer.

Van De Wiel M. A., Bucchianico A. (2001). Fast computation of the exact null distribution of Spearman's rho and Page's L statistic for samples with and without ties. *Journal of Statistical Planning and Inference*, 92, 133–145.

White H. (1994). *Estimation, inference and specification analysis*. Cambridge University Press.

Yan J. (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, 21 (4), 1–21.

Appendix 8 "Analysis of multidimensional probability distributions with copula functions - Part III."

Fantazzini, D. (2011d). Analysis of multidimensional probability distributions with copula functions - Part III. *Applied Econometrics*, 24(4), 100-130. http://pe.cemi.rssi.ru/pe_2011_4_100-130.pdf

Permisson to post the Publisher's version/PDF: according to http://appliedeconometrics.cemi.rssi.ru/AppEc.files/contents_en.html, "*full texts are available for downloading with approximately one year lag*". In some cases, this 1-year lag can be shorter.

Д. Фантаццини

Моделирование многомерных распределений с использованием копула-функций. III[1]

Заключительная часть консультации посвящена описанию подходов к эмпирическому подбору подходящей копула-функции и методов статистической проверки гипотез, связанных с моделями копула-функций.

Ключевые слова: критерий согласия, критическая статистика, копула-функция, байесовский выбор, информационный критерий качества подгонки, эмпирический анализ.

JEL classification: C69, C49.

Содержание

8. ВЫБОР КОПУЛА-ФУНКЦИИ
8.1. Информационный критерий Акаике (AIC)
8.2. Тесты отношения правдоподобия
8.3. Байесовский выбор копула-функции
8.4. Эмпирические приложения: пример выбора копула-функции
9. КРИТЕРИИ СОГЛАСИЯ ДЛЯ КОПУЛА-ФУНКЦИЙ
9.1. Тесты, основанные на преобразовании Розенблатта
9.2. Тесты, основанные на использовании эмпирических копула-функций
9.3. Другие подходы к построению критериев согласия
9.4. Тесты, основанные на усреднении критических статистик
9.5. Эмпирические приложения с пакетом *R*: пример реализации подхода $\mathcal{A}_2$

8. Выбор копула-функции

Из нескольких моделей копула-функций (определение копула-функции см. в (Фантаццини, 2011a)) следует выбрать ту, которая наилучшим образом описывает рассматриваемые данные. В разделах 8.1 – 8.3 описываются различные критерии, руководствуясь которыми исследователи обычно осуществляют такой выбор.

8.1. Информационный критерий Акаике (AIC)

Этот метод выбора копула-функций является достаточно простым, см., например, (Breymann et al., 2003; Dias, Embrechts, 2004).

[1] Окончание. Первые две части консультации см. в журнале «*Прикладная эконометрика*», №№ 22 (3) и 23 (2), 2011. Перевод с английского А. В. Кудрова под научной редакцией С. А. Айвазяна. Сохранена сквозная нумерация разделов, рисунков, формул, определений и теорем, начатая в первой части.

AIC (Akaike Information Criterion) — это критерий, позволяющий выбрать копула-функцию, которая отвечает наименьшему значению из следующих:

$$\mathrm{AIC}(C_k,\hat{\alpha}_k) = -2\sum_{t=1}^{T}\log c_k(u_{1t},...,u_{nt};\hat{\alpha}_k) + 2q_k, \quad k=1,...,K, \tag{23}$$

где C_k — k-ая модель копула-функции, $c_k(\cdot)$ — плотность для C_k; $\hat{\alpha}_k$ — вектор параметров копула-функции C_k, q_k — число параметров, от которых зависит функция C_k. Функция $\mathrm{AIC}(C_k,\hat{\alpha}_k)$ в определенном смысле штрафует модели с большим числом параметров.

AIC предполагает, что наиболее адекватная модель находится среди рассматриваемых моделей. Если сравнивать «невложенные» модели, то слагаемое $2q_k$ является не вполне корректным штрафом, и указанная выше формула становится неприменимой. Для того чтобы разрешить эту проблему, Такеучи предложил в (Takeuchi, 1976) информационный критерий TIC. Однако TIC редко используется на практике, поскольку для него требуется наличие весьма больших (по размеру) выборок наблюдений.

В применении AIC некоторые вопросы остаются открытыми. Например, до сих пор неизвестно, что происходит с AIC в случае, если в качестве частных одномерных распределений использовать эмпирическую функцию распределения. Использование эмпирической функции распределения как оценки соответствующей частной функции распределения на практике является стандартной процедурой, но последствия такого использования для AIC до сих пор не изучены.

8.2. Тесты отношения псевдоправдоподобия

В работе (Chen, Fan, 2006) предлагаются так называемые тесты отношения псевдоправдоподобия, которые учитывают случайность, возникающую в AIC при использовании нормированных рангов. Эти тесты позволяют удостовериться в том, что среди рассматриваемых моделей нет такой, которая была бы значительно лучше выбранной (проверяемой) модели. При этом не требуется, чтобы рассматриваемый класс моделей содержал истинную модель. Таким образом можно сравнивать невложенные модели.

Основной идеей тестов является сравнение каждой из рассматриваемых моделей-кандидатов с остальными моделями с точки зрения меры их правдоподобия, по результатам сравнения выбирается та модель, которая оказывается наиболее правдоподобной. А именно, пусть $C_1(u_1,...,u_n;\alpha_1)$ — выбранная из числа рассматриваемых *модель-образец* копула-функции, $C_i(u_1,...,u_n;\alpha_i)$, $i=2,...,M$ — остальные модели-кандидаты, $C^0(u_1,...,u_n)$ — истинная (и *ненаблюдаемая*) копула-функция, в действительности описывающая зависимости между компонентами многомерных наблюдений. В работе (Chen, Fan, 2006) предлагается, используя расстояние Кульбака–Лейблера, тестировать гипотезу о том, является ли модель-образец наиболее правдоподобной среди всех рассматриваемых моделей. Таким образом, нулевой гипотезой является:

$$\mathrm{H}_0^M : \max_{i=2,...,M} \mathrm{E}^0\left[\ln\frac{c_i\left(F_1^0(\varepsilon_{1t}),...,F_n^0(\varepsilon_{nt});\alpha_i^*\right)}{c_1\left(F_1^0(\varepsilon_{1t}),...,F_n^0(\varepsilon_{nt});\alpha_1^*\right)}\right] \le 0,$$

где α_i^*, $i=1,...,M$ — параметры рассматриваемых моделей копула-функций; $F_j^0(\cdot)$ — истинные (неизвестные) частные распределения для ε_{jt}, $j=1,...,n$ [2]. Указанная нулевая гипотеза означает, что среди всех рассматриваемых моделей модель-образец оказывается, в определенном смысле, ближе к истинной модели, чем все остальные из рассматриваемых. Альтернативная гипотеза

$$\mathrm{H}_1^M : \max_{i=2,..,M} \mathbf{E}^0\left[\ln\frac{c_i\left(F_1^0(\varepsilon_{1t}),...,F_n^0(\varepsilon_{nt});\alpha_i^*\right)}{c_1\left(F_1^0(\varepsilon_{1t}),...,F_n^0(\varepsilon_{nt});\alpha_1^*\right)}\right]>0$$

означает, что среди рассматриваемых моделей существует модель-кандидат, которая ближе к истинной модели, чем модель-образец. Тест из работы (Chen, Fan, 2006) основан на следующей статистике отношения псевдоправдоподобия ($i=2,...,M$):

$$LR_n(F_{1T},...,F_{nT};\tilde{\alpha}_i,\tilde{\alpha}_1)=\frac{1}{T}\sum_{t=1}^{T}\left[\ln\frac{c_i\left(F_{1T}(\varepsilon_{1t}),...,F_{nT}(\varepsilon_{nt});\tilde{\alpha}_i\right)}{c_1\left(F_{1T}(\varepsilon_{1t}),...,F_{nT}(\varepsilon_{nt});\tilde{\alpha}_1\right)}\right],$$

где F_{jT}, $j=1,...,n$ — нормированные ранги; знак ~ над буквой означает, что соответствующий параметр α оценен при помощи квази метода максимального правдоподобия (КММП). Возьмем $\Omega=(\sigma_{ik})_{i,k=2}^M$, где

$$\sigma_{ik}=\mathrm{Cov}^0\left[\ln\frac{c_i(U_{1t},...,U_{nt};\alpha_i^*)}{c_1(U_{1t},...,U_{nt};\alpha_1^*)}+\sum_{j=1}^{n}\left\{Q_{i,j}(U_{jt};\alpha_i^*)-Q_{1,j}(U_{jt};\alpha_1^*)\right\},\right.$$
$$\left.\ln\frac{c_k(U_{1t},...,U_{nt};\alpha_k^*)}{c_1(U_{1t},...,U_{nt};\alpha_1^*)}+\sum_{j=1}^{n}\left\{Q_{k,j}(U_{jt};\alpha_k^*)-Q_{1,j}(U_{jt};\alpha_1^*)\right\}\right]$$

при $U_{jt}=F_j^0(\varepsilon_{jt})$ и $Q_{i,j}(U_{jt};\alpha_i^*)\equiv\mathbf{E}^0\{l_{ij}(U_s;\alpha_i^*)[\mathbf{1}_{\{U_{jt}\le U_{js}\}}-U_{js}]\,|\,U_{jt}\}$ для $i=1,...,M$, $j=1,...,d$, добавочные слагаемые $Q_{i,j}(\cdot)$ введены в связи с необходимостью оценивать частные распределения $F_j^0(\cdot)$, $j=1,...,n$, а $\mathbf{1}_{\{x<x_0\}}$, как и прежде, индикаторная функция множества $\{x<x_0\}$. Более подробно см. (Chen, Fan, 2006). Отметим, что если частные распределения известны, то эти слагаемые исчезают. Как показано в работе (Chen, Fan, 2006), при некоторых условиях регулярности:

$$n^{1/2}\left\{LR_n(F_{1T},...,F_{nT};\tilde{\alpha}_i,\tilde{\alpha}_1,)-\mathbf{E}^0\left[\ln\frac{c_i(U_{1t},...,U_{nt};\alpha_i^*)}{c_1(U_{1t},...,U_{nt};\alpha_1^*)}\right]\right\}_{i=2,...,M}\xrightarrow{d}(Z_2,...,Z_M)',$$

где вектор $(Z_2,...,Z_M)'$ имеет нормальное распределение $N(0,\Omega)$, так что

$$\max_{i=2,..,M} n^{1/2}\left\{LR_n(F_{1T},...,F_{nT};\tilde{\alpha}_i,\tilde{\alpha}_1)-\mathbf{E}^0\left[\ln\frac{c_i(U_t;\alpha_i^*)}{c_1(U_t;\alpha_1^*)}\right]\right\}\xrightarrow{d}\max_{i=2,...,M} Z_i. \qquad (24)$$

[2] Верхний индекс ^{0} здесь и в дальнейшем означает, что соответствующая характеристика вычислена в условиях гипотезы H_0^M.

В (Chen, Fan, 2006) показано, что при так называемой «наименее благоприятной» конфигурации

$$T_T^M = \max_{i=2,\dots,M} [n^{1/2} LR_n(F_{1T},\dots,F_{nT};\tilde{\alpha}_i,\tilde{\alpha}_1,\tilde{\theta})] \xrightarrow{d} \max_{i=2,\dots,M} Z_i\,, \tag{25}$$

т. к. второе слагаемое в (24) обнуляется: $\mathbf{E}^0[\cdot]=0$. Поскольку асимптотическая ковариационная матрица Ω для $Z_2,\dots,Z_M$ зависит от $\alpha_1^*,\dots,\alpha_M^*$, распределение $\max_i Z_i$ неизвестно. Но, несмотря на это, можно использовать *p*-значения, получаемые либо методом Монте-Карло, либо бутстрап методом, см. (White, 2000; Chen, Fan, 2006).

Состоятельная оценка для $\Omega=(\sigma_{ik})_{i,k=2}^M$ определяется формулой:

$$\hat{\sigma}_{ik} = \frac{1}{T}\sum_{t=1}^{T}\left(\ln\frac{c_i(F_{tT};\tilde{\alpha}_i)}{c_1(F_{tT};\tilde{\alpha}_1)} - \frac{1}{T}\sum_{s=1}^{T}\ln\frac{c_i(F_{sT};\tilde{\alpha}_i)}{c_1(F_{sT};\tilde{\alpha}_1)} + \sum_{j=1}^{n}\left(\hat{Q}_{i,j}(F_{jt};\tilde{\alpha}_i) - \hat{Q}_{1,j}(F_{jt};\tilde{\alpha}_1)\right)\right)\times$$
$$\times\left(\ln\frac{c_k(F_{tT};\tilde{\alpha}_k)}{c_1(F_{tT};\tilde{\alpha}_1)} - \frac{1}{T}\sum_{s=1}^{T}\ln\frac{c_k(F_{sT};\tilde{\alpha}_k)}{c_1(F_{sT};\tilde{\alpha}_1)} + \sum_{j=1}^{n}\left(\hat{Q}_{k,j}(F_{jt};\tilde{\alpha}_k) - \hat{Q}_{1,j}(F_{jt};\tilde{\alpha}_1)\right)\right),$$

где $\hat{Q}_{i,j}$ — состоятельная оценка $Q_{i,j}$, которая вычисляется следующим образом:

$$\hat{Q}_{i,j}(U_{js};\hat{\alpha}_i) = \frac{1}{T}\sum_{t=1,t\neq s}^{T} l_{ij}(U_t;\hat{\alpha}_i)\left(\mathbf{1}\{U_{js}\le U_{jt}\} - U_{jt}\right)^3.$$

Затем можно вычислить разложение Холецкого $\hat{\Omega}=\hat{C}\hat{C}'$ и сформировать

$$\hat{Z}_b = (\hat{Z}_{b,2},\dots,\hat{Z}_{b,M})' = \hat{C}\eta_b\,,$$

где $\eta_b \in N(0, I_{M-1})$, т. е. имеет $(M-1)$-мерное стандартное нормальное распределение. После чего найдем наибольшую порядковую статистику $\zeta_{b,M-1} = \max_{k=2,\dots,M}\hat{Z}_{b,k}$, а для того, чтобы вычислить *p*-значение методом Монте-Карло, повторим эти итерации много раз (например, порядка $B=100000$), см. (Chen, Fan, 2006):

$$\frac{1}{B}\sum_{b=1}^{B}\mathbf{1}\{\zeta_{b,M-1}\le T_T^M\}\,.$$

8.3. Байесовский выбор копула-функций

В (Huard et al., 2006) предлагается байесовская процедура выбора наиболее вероятного двумерного семейства копула-функций среди заданного множества копула-функций. В этом подходе параметры копула-функций интерпретируются как случайные величины, и, следовательно, они не должны оцениваться. Кроме того, в указанной работе параметризация плотности копула-функции осуществляется в терминах τ-Кендалла (см. (Фантаццини, 2011b, п. 5.2)), так что априорные распределения параметров заменены априорными распределениями τ-Кендалла, которые концептуально являются более удобными. Априорное распре-

[3] Здесь и в дальнейшем $\mathbf{1}\{\cdot\}$ означает, как и раньше, индикаторную функцию множества $\{\cdot\}$.

деление τ-Кендалла берется одним и тем же для всех тестируемых копула-функций и используется как базис для сравнения.

Пусть $\mathcal{C}$ — множество всех копула-функций. Возьмем из $\mathcal{C}$ конечное подмножество $\mathcal{C}_Q \subset \mathcal{C}$ копула-функций, которые необходимо включить в предложенный метод. Каждая копула-функция из $\mathcal{C}_Q$ обозначается C_l, $l = 1,...,Q$. Метод, предложенный в работе (Huard et al., 2006), на первом шаге состоит в проверке следующих Q гипотез:

$$\mathrm{H}_l: \quad \textit{данные извлечены из копула-функции} \quad C_l, \quad l = 1,\ldots,Q.$$

Для этого вычисляются $\mathrm{P}(\mathrm{H}_l \mid D)$ — вероятности реализации гипотезы H_l при заданных наблюдениях D, которые представляют собой T независимых пар (u_t, v_t), $t = 1,...,T$, с равномерно распределенными компонентами (в работе (Huard et al., 2006) называемые квантилями). Если используются нормированные ранги, то условие независимости может нарушаться, и предложенный метод следует рассматривать как аппроксимацию и применять с некоторой осторожностью. Используя теорему Байеса, получим:

$$\mathrm{P}(\mathrm{H}_l \mid D, I) = \frac{\mathrm{P}(D \mid \mathrm{H}_l, I)\mathrm{P}(\mathrm{H}_l \mid I)}{\mathrm{P}(D \mid I)}, \tag{26}$$

где $\mathrm{P}(D \mid \mathrm{H}_l, I)$ — функция правдоподобия, $\mathrm{P}(\mathrm{H}_l \mid I)$ — априорное распределение для копула-функции, $\mathrm{P}(D \mid I)$ — нормирующая константа; I соответствует дополнительной информации. «Правильной» копула-функцией, в терминах работы (Huard et al., 2006), называют копула-функцию, отвечающую наибольшему значению апостериорной вероятности $\mathrm{P}(\mathrm{H}_l \mid D, I)$.

В (Huard et al., 2006) плотность копула-функции перепараметризируется в терминах τ-Кендалла ($\tau = g_l(\alpha)$), который становится общим параметром для всех копула-функций из $\mathcal{C}_Q$.

Таблица 3. Выражения для τ-Кендалла для некоторых копула-функций

Копула-функция	$\tau = g_l(\alpha)$	Область изменения τ
Клейтон	$1 - 2(2+\alpha)^{-1}$	$[0;1] \setminus \{0\}$
Гумбель	$1 - \alpha^{-1}$	$[0;1]$
Франк	$1 - 4\alpha^{-1}\left(1 - \alpha^{-1}\int_0^{\alpha} s/(e^s - 1)\,ds\right)$	$[-1;1] \setminus \{0\}$
Гаусс	$2\pi^{-1}\arcsin\rho$	$[-1;1]$

Следовательно, τ-Кендалла может использоваться в (26) как случайная величина

$$\mathrm{P}(\mathrm{H}_l \mid D, I) = \int_{-1}^{+1} \mathrm{P}(\mathrm{H}_l, \tau \mid D, I)\,d\tau = \int_{-1}^{+1} \frac{\mathrm{P}(D \mid \mathrm{H}_l, \tau, I)\mathrm{P}(\mathrm{H}_l \mid \tau, I)\mathrm{P}(\tau \mid I)\,d\tau}{\mathrm{P}(D \mid I)}, \tag{27}$$

где $\mathrm{P}(\mathrm{H}_l \mid \tau, I)$ — априорная вероятность гипотезы H_l для копула-функции; $\mathrm{P}(\tau \mid I)$ — априорная плотность для τ-Кендалла; правдоподобие $\mathrm{P}(D \mid \mathrm{H}_l, \tau, I)$, которое теперь зависит от τ, может быть вычислено следующим образом:

$$\mathrm{P}(D \mid \mathrm{H}_l, \tau, I) = \prod_{t=1}^{T} \mathrm{P}(u_t, v_t \mid \tau, l, I) = \prod_{t=1}^{T} c_l\left(u_t, v_t \mid g_l^{-1}(\tau)\right), \tag{28}$$

где $c_l(u_t, v_t \mid g_l^{-1}(\tau))$ — плотность l-ой копула-функции, параметризованной в терминах τ-Кендалла.

Для того чтобы выбрать априорные распределения, в работе (Huard et al., 2006) указываются некоторые требования и правила:

(I_1) τ-Кендалла принадлежит множеству Λ и каждая реализация $\tau \in \Lambda$ равновероятна;

(I_2) для заданного τ все копула-функции, для которых $\tau \in \Omega_l$, равновероятны, здесь Ω_l — область значений τ в условиях справедливости гипотезы H_l.

Область Λ необходима для того, чтобы включать дополнительную информацию относительно зависимости между переменными. Например, если известно, что τ положительно, то можно предположить, что $\Lambda = [0; 1]$; если априорная информация относительно значений τ отсутствует, то $\Lambda = [-1; 1]$. Следовательно, учитывая (I_1), априорное распределение τ есть:

$$\mathrm{P}(\tau \mid I_1) = \begin{cases} \dfrac{1}{\lambda(\Lambda)} & \text{при} \quad \tau \in \Lambda; \\ 0 & \text{иначе,} \end{cases} \tag{29}$$

где $\lambda(\cdot)$ обозначает меру Лебега, т. е. $\lambda(\Lambda)$ — длина интервала Λ. Кроме того, условие (I_2) определяет априорное распределение для копула-функции:

$$\mathrm{P}(\mathrm{H}_l \mid \tau, I_2) \propto \mathbf{1}\{\tau \in \Omega_l\}, \tag{30}$$

где предполагается, что все копула-функции равновероятны относительно τ. В случае, если известно, что τ концентрируется вокруг определенного значения, то в работе (Huard et al., 2006) предлагается использовать бета-распределение на интервале [–1; 1] с параметрами, обеспечивающими такую форму распределения, которая соответствует априорной информации.

Если подставить (28), (30) и (29) в (27), то получим:

$$\begin{aligned} \mathrm{P}(\mathrm{H}_l \mid D, I) &= \frac{1}{\mathrm{P}(D \mid I)} \int_{-1}^{+1} \prod_{t=1}^{T} c_l\left(u_t, v_t \mid g_l^{-1}(\tau)\right) \cdot \frac{\mathbf{1}\{\tau \in \Omega_l \cap \Lambda\}}{\lambda(\Lambda)} d\tau = \\ &= \frac{1}{\mathrm{P}(D \mid I)\lambda(\Lambda)} \int_{\Omega \cap \Lambda} \prod_{t=1}^{T} c_l\left(u_t, v_t \mid g_l^{-1}(\tau)\right) d\tau, \end{aligned} \tag{31}$$

где нормирующая константа $\mathrm{P}(D \mid I)$ в (31) вычисляется при помощи так называемого правила суммы, более подробно об этом см. (Jaynes, Bretthorst, 2003):

$$\mathrm{P}(D \mid I) = \sum_{l=1}^{Q} \mathrm{P}(D \mid \mathrm{H}_l, I) \mathrm{P}(\mathrm{H}_l \mid I).$$

Важно отметить, что правило суммы справедливо только тогда, когда гипотезы H_l являются взаимоисключающими и исчерпывающими (в совокупности). К сожалению, эти условия обычно не выполняются по двум причинам. Во-первых, если данные сгенерированы из копула-функции, не входящей в $\mathcal{C}_Q$ (что является довольно распространенной ситуацией), то гипотезы H_l в совокупности не являются исчерпывающими. Во-вторых, если множество гипотез H_l содержит две или более похожих копула-функции, то в этом случае гипотезы H_l не являются исчерпывающими. В работе (Huard et al., 2006) весьма кратко рассматриваются возможные решения, которые бы обеспечили полноту, но более детальное изучение этого вопроса оставлено для дальнейшего исследования.

Этот метод представляется весьма интересным, поскольку не требует предварительного оценивания параметров копула-функций. Более того, отметим, что рассматриваемые копула-функции не обязаны быть вложенными.

8.4. Эмпирические приложения: пример выбора копула-функций

Критерий AIC не вычисляется напрямую в пакете *R*, но несложно написать небольшую функцию, которая бы вычисляла значение AIC при заданных значении максимума логарифмической функции правдоподобия и числе параметров. Эта задача остается в качестве упражнения.

В пакете MATLAB реализован код для байесовского выбора копула-функции из работы (Huard et al., 2006), он доступен на сайте http://code.google.com/p/copula/.

9. Критерии согласия для копула-функций

После того, как, согласно указанным выше процедурам, выбрана «наиболее подходящая» копула-функция, следующим шагом является проверка того, насколько хорошо она согласуется с имеющимися данными.

Одним из вариантов такой проверки является *неформальная графическая диагностика*, в соответствии с которой сравниваются нормированные ранги со случайными выборками, сгенерированными в соответствии с используемой копула-функцией. А именно, графически сравнивается эмпирическая оценка копула-функции с параметрической моделью. Однако эту процедуру можно рассматривать в качестве предварительного и приблизительного анализа.

С недавнего времени начали появляться работы, в которых изучаются статистические тесты, используемые в критериях согласия. Вот лишь некоторые из них: (Genest, Rivest, 1993; Shih, 1998; Breymann et al., 2003; Malevergne, Sornette, 2003; Scaillet, 2005; Fermanian, 2005; Panchenko, 2005; Genest et al., 2006a; Berg, Bakken, 2007; Dobrić, Schmid, 2007; Quessy et al., 2007; Genest et al., 2009; Genest, Rémillard, 2008; Berg, 2009).

Пусть $C(\cdot)$ — n-мерная копула-функция для рассматриваемой выборки. Необходимо протестировать следующую гипотезу:

$$\mathrm{H}_0 : C \in \mathcal{C} = \{C_\alpha ; \alpha \in \Theta\} \text{ против } \mathrm{H}_1 : C \notin \mathcal{C},$$

где $C_\alpha(\cdot)$ — параметрическая копула-функция, Θ — пространство параметров.

В случае, когда выбирается наиболее приемлемая *одномерная* модель, обычно используются такие критерии согласия, как критерий Колмогорова–Смирнова или критерий Андерсона–Дарлинга.

Возможным вариантом построения критерия согласия для копула-функций, используемых для описания многомерных распределений, являются подходы, основанные на многомерном группировании, см., например, (Dobrić, Schmid, 2005). К сожалению, при этом, как и во всех подходах, основанных на группировании, используется дискретизация вероятностного пространства, которая практически нереализуема в случае задач большой размерности. Более того, группирование данных в некотором смысле произвольно и весьма нетривиально. Аналогично, подходы, основанные на использовании многомерных ядерных оценок копула-функций, представленные, например, в работах (Fermanian, 2005; Scaillet, 2005), являются в случае большой размерности (характерной для задач из области финансов и страхования) чрезмерно затратными с вычислительной точки зрения. Поэтому далее, характеризуя многомерный случай, будут рассматриваться некоторые другие подходящие подходы.

Перед тем как приступить к анализу различных тестов, кратко определим, что представляет собой *преобразование Розенблатта*. Изначально оно было предложено в работе (Rosenblatt, 1952) и стало известным как *условно-вероятностное интегральное преобразование* (Conditional Probability Integral Transform, CPIT). Оно позволяет преобразовать набор зависимых случайных величин, описываемых некоторым многомерным распределением, в набор независимых равномерно распределенных на [0; 1] случайных величин. Таким образом, если для многомерных случайных величин имеется тест на независимость и равномерность, то, используя преобразование Розенблатта, можно протестировать адекватность любой модели.

Если взять случайную величину $\nu = F(X)$, где $F(\cdot)$ — функция распределения случайной величины X, то ν будет иметь равномерное распределение на [0; 1]. Если в качестве $F(\cdot)$ взять иную функцию распределения, скажем $G(\cdot)$, то $\nu = G(X)$ уже не будет равномерно распределенной. Это простое свойство лежит в основе большинства работ, в которых изучаются вопросы, связанные с тестированием распределений и копула-функций.

Дадим формальное определение преобразования Розенблатта.

Определение 13 (преобразование Розенблатта). *Пусть* $X = (X_1, ..., X_n)$ *— случайный вектор с функцией распределения* $F_X(x_1, x_2, \ldots, x_n)$; $F_{X_i} = P(X_i \le x_i)$ *— функция распределения случайной величины* X_i, $i = 1, \ldots, n$. *Рассмотрим следующее вероятностно-интегральное преобразование, определяющее n случайных величин* $V_i = T(X_i)$:

$$
\begin{aligned}
T(x_1) &= P(X_1 \le x_1) = F_{X_1}(x_1);\\
T(x_2) &= P(X_2 \le x_2 \mid X_1 = x_1) = F_{X_2|X_1}(x_2 \mid x_1);\\
&\vdots\\
T(x_n) &= P(X_n \le x_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) =\\
&= F_{X_n|X_1,\ldots,X_{n-1}}(x_n \mid x_1, \ldots, x_{n-1}).
\end{aligned}
$$

Тогда случайные величины $V_i = T(X_i)$ $(i = 1, 2, \ldots, n)$ *являются равномерно распределенными на отрезке* [0; 1] *и взаимно независимыми, а сами преобразования* $T(X_i)$ *называют-*

ся преобразованиями Розенблатта или условно-вероятностными интегральными преобразованиями (CPIT).

Предположим, что многомерная функция распределения $F_X(\cdot)$ имеет копула-функцию $C(\cdot)$, т. е.

$$F_X(x_1, x_2, \ldots, x_n) = C(F_{X_1}(x_1), \ldots, F_{X_n}(x_n)).$$

Обозначим через $C_i(u_1, \ldots, u_i)$ совместное *i*-мерное распределение, такое, что:

$$C_i(u_1, \ldots, u_i) = C(u_1, \ldots, u_i, 1, \ldots, 1), \quad i = 1, 2, \ldots, n-1.$$

При этом $C_1(u_1) = u_1$ и $C_n(u_1, \ldots, u_n) = C(u_1, \ldots, u_n)$. Пусть случайный вектор $(U_1, \ldots, U_i)$ имеет распределение $C_i(u_1, \ldots, u_n)$. Тогда плотность условного распределения U_i в точке u_i при заданных значениях $u_1, u_2, \ldots, u_{i-1}$ случайных величин $U_1, \ldots, U_{i-1}$ соответственно равна:

$$c_i(u_i \mid u_1, \ldots, u_{i-1}) = \frac{\partial^{i-1} C_i(u_1, \ldots, u_i)}{\partial u_1 \ldots \partial u_{i-1}} \Bigg/ \frac{\partial^{i-1} C_{i-1}(u_1, \ldots, u_i)}{\partial u_1 \ldots \partial u_{i-1}}$$

при $i = 1, \ldots, n$. Теперь с использованием условных распределений $C_i(\cdot)$ можно определить случайные величины V_i при $i = 2, \ldots, n$:

$$V_i = C_i(F_{X_i}(X_i) \mid F_{X_1}(X_1), \ldots, F_{X_{i-1}}(X_{i-1})).$$

Как указано в работе (Berg, 2009), преобразование Розенблатта дает значительное преимущество при тестировании критериев согласия. В (Hong, Li, 2005) представлены результаты многомерных тестов с использованием имитационного моделирования и преобразованных исходных случайных величин, которые оказались лучше, чем результаты с использованием *непреобразованных* случайных величин. В работе (Chen et al., 2004) высказывается предположение, что аналогичный вывод справедлив и для критериев согласия для копула-функций.

В качестве недостатка тестов, основанных на преобразовании Розенблатта, следует отметить отсутствие инвариантности относительно принятого порядка нумерации рассматриваемых случайных величин. Общее число возможных преобразований случайных величин тогда будет равно *n*!. В (Berg, 2009) показано, что для некоторых подходов, основанных на преобразовании Розенблатта, оценки *p*-значений критической статистики зависят от того, какой порядок преобразований выбран для рассматриваемых случайных величин. Однако в этой работе также указывается, что по мере роста числа наблюдений различия в *p*-значениях уменьшаются, причем эти *p*-значения асимптотически сходятся к некоторым общим критическим уровням. Таким образом, в работе (Berg, 2009) высказывается предположение, что выбор порядка преобразований лишь незначительным образом влияет на результаты теста.

Следуя схеме, представленной в (Berg, 2009), критерии согласия для копула-функций могут быть классифицированы в девять семейств:

$\mathcal{A}_1$: *Тесты, основанные на преобразовании Розенблатта.* Предложены в работе (Berg, Bekken, 2007). Этот подход в качестве частных случаев включает в себя подходы, предложенные в работах (Malevergne, Sornette, 2003; Breymann et al., 2003; Chen et al., 2004).

$\mathcal{A}_2$: *Тесты, основанные на эмпирических оценках копула-функций.* Предложены в работах (Kole et al., 2007; Genest, Rémillard, 2008).

$\mathcal{A}_3$: *Тесты, основанные на подходе* $\mathcal{A}_2$ *и преобразовании Розенблатта.* Предложены в (Genest et al., 2009).

$\mathcal{A}_4$: *Тесты, основанные на эмпирических оценках копула-функций и их функций распределения.* Предложены в работах (Genest, Rivest, 1993; Wang, Wells, 2000; Savu, Trede, 2008; Genest et al., 2006a).

$\mathcal{A}_5$: *Тесты, основанные на функциях спирменовской меры зависимости* (*см. (Фантаццини, 2001b, п. 5.2*)). Предложены в (Quessy et al., 2007).

$\mathcal{A}_6$: *Тесты, основанные на критерии Шиха* (*Shih, 1998*) *для двумерной модели Клейтона.* Обобщены на случай произвольной размерности в работе (Berg, 2009).

$\mathcal{A}_7$: *Тесты, основанные на внутреннем произведении векторов, как мере расстояния между ними.* Предложены в работе (Panchenko, 2005).

$\mathcal{A}_8$: *Тесты, основанные на подходе* $\mathcal{A}_7$ *и преобразовании Розенблатта.* Предложены в (Berg, 2009).

$\mathcal{A}_9$: *Тесты, основанные на объединении указанных выше подходов.* Предложены в (Berg, 2009).

Подходы $\mathcal{A}_1 - \mathcal{A}_5$ основаны на снижении размерности, поскольку статистики этих тестов являются одномерными, тогда как статистика подхода $\mathcal{A}_6$ строится с использованием моментов; критерии $\mathcal{A}_7$, $\mathcal{A}_8$ представляют собой многомерные подходы. В следующих разделах 9.1 – 9.4 будем придерживаться структуры, предложенной в работах (Berg, 2009; Berg, Bekken, 2007).

9.1. Тесты, основанные на преобразовании Розенблатта

Подход $\mathcal{A}_1$ (CPIT-подход)

Подход $\mathcal{A}_1$ основан на преобразовании Розенблатта, примененном к нормированным рангам $\hat{\mathbf{u}}_1 = (\hat{u}_{11},\ldots,\hat{u}_{n1}),\ldots,\hat{\mathbf{u}}_T = (\hat{u}_{1T},\ldots,\hat{u}_{nT})$, где компоненты $\hat{u}_{it}$ вектора $\hat{\mathbf{u}}_t$ вычисляются в предположении справедливости нулевой гипотезы о копула-функции $C_{\hat{\alpha}}$ по следующей формуле (см. п. 1 из раздела 7.1 работы (Фантаццини, 2001b)):

$$\hat{u}_{it} = F_{1,T}(x_{it}), \text{ где } F_{i,T}(x) = \frac{1}{T+1}\sum_{t=1}^{T}\mathbf{1}\{x_{it} < x\}, \quad i = 1,2,\ldots,n. \tag{32}$$

В соотношении (32) правая часть определяет непараметрическую оценку частной функции распределения переменной X_i в точке x_{it}. При справедливости нулевой гипотезы, выборка $\mathbf{V} = (\mathbf{v}_1,\ldots,\mathbf{v}_T)$ представляет собой выборку из копула-функции P, отвечающей случаю независимости (также обозначаемой как $C_\perp$). К сожалению, эмпирический аналог преобразования Розенблатта предполагает использование рангов, что индуцирует зависимость эмпирических значений V_i, $i = 1,2,\ldots,n$. Таким образом, V_i, $i = 1,\ldots,n$, не являются независимыми одинаково распределенными случайными величинами. Тем не менее, как показа-

но в работе (Genest et al., 2006a), можно получать достаточно надежные оценки *p*-значений критических статистик при помощи параметрической бутстрап процедуры. Теоретические свойства этих оценок представлены в работе (Genest, Rémillard, 2008)[4].

Общий вид критической статистики критериев согласия для подхода $\mathcal{A}_1$ следующий:

$$W_{1t} = \sum_{i=1}^{n} \Gamma\{v_{it}\}, \quad t = 1, \ldots, T, \tag{33}$$

где $\Gamma\{\cdot\}$ — весовая функция, используемая для взвешивания информации из $(\mathbf{v}_1, \ldots, \mathbf{v}_T)$. В качестве весовой функции может быть использована, например, $\Gamma\{v_{it}\} = (\Phi^{-1}(v_{it}))^2$, что соответствует подходу из (Breymann et al., 2003). Более того, если нулевая гипотеза соответствует гауссовской копула-функции, тогда имеем подход, предложенный в работе (Malevergne, Sornette, 2003). В двух последних упомянутых работах используется статистика Андерсона–Дарлинга (Anderson, Darling, 1954). В работе (Berg, Bekken, 2007) показано, что статистика Андерсона–Дарлинга с $\Gamma\{v_{it}\} = |v_{it} - 0.5|$ дает достаточно хорошие результаты при тестировании нулевой гипотезы, соответствующей гауссовской копула-функции. Следовательно, для двух упомянутых выше вариантов выбора весовой функции имеем два подхода:

$$\mathcal{A}_{1t}^{(a)} = \sum_{i=1}^{n} \left(\Phi^{-1}(v_{it})\right)^2 \quad \text{и} \quad \mathcal{A}_{1t}^{(b)} = \sum_{i=1}^{n} |v_{it} - 0.5|.$$

Можно показать, что для подхода $\mathcal{A}_{1t}^{(a)}$ функция распределения $F_1(\cdot)$ статистики W_{1t} для любых t представляет собой χ_n^2-распределение. Однако, как говорилось выше, если использовать нормированные ранги, распределение W_{1t} неизвестно. Но для аппроксимации функции распределения $F_1(\cdot)$ при справедливости нулевой гипотезы необходимо обратиться к бутстрап процедуре.

Так называемый «тестовый датчик» (test observator) S_1 подхода $\mathcal{A}_1$ определяется как функция распределения для $F_1(W_{1t})$, см. (Berg, 2009):

$$S_{1t}(w) = \mathrm{P}\left(F_1(W_{1t}) \le w\right), \quad w \in [0;1].$$

При справедливости нулевой гипотезы имеем $S_{1t}(w) = w$ при всех *t*. Эмпирическая версия тестового датчика может быть вычислена в виде:

$$\hat{S}_1(w) = \frac{1}{T+1} \sum_{t=1}^{T} \mathbf{1}\{F_1(W_{1t}) \le w\}, \quad w = \frac{1}{T+1}, \ldots, \frac{T}{T+1}. \tag{34}$$

Для всех подходов, которые предполагают снижение размерности, в работе (Berg, 2009) рассматривается статистика Крамера–фон Мизеса, но следует иметь в виду, что могут быть использованы и другие статистики (например, Андерсона–Дарлинга, Колмогорова–Смир-

[4] Асимптотические свойства этой бутстрап процедуры до сих пор были получены лишь для подходов $\mathcal{A}_2$ и $\mathcal{A}_4$. Однако результаты, полученные в работах (Berg, Bekken, 2007; Dobrić, Schmid, 2007; Berg, 2009), указывают на справедливость упомянутых асимптотических свойств и для других подходов.

нова, см. (Berg, Bekken, 2007; Marsaglia, Marsaglia, 2004)). Если использовать статистику Крамера–фон Мизеса, то критическая статистика имеет вид:

$$\hat{T}_1 = \frac{T}{3} + \frac{T}{T+1}\sum_{t=1}^{T}\hat{S}_1\left(\frac{t}{T+1}\right)^2 - \frac{T}{(T+1)^2}\sum_{t=1}^{T}(2t+1)\hat{S}_1\left(\frac{t}{T+1}\right)^2 \tag{35}$$

(см. доказательство в (Berg, 2009)).

Некоторое расширение: CPIT-2 подход. В работе (Berg, Bekken, 2007) показано, что описанный выше подход, в котором используется весовая функция $(\Phi^{-1}(\cdot))^2$, не всегда является адекватным, поскольку данным, расположенным на границе n-мерного единичного гиперкуба, присваиваются бóльшие веса: при небольших выборках такой способ взвешивания делает подход менее устойчивым и менее мощным, поскольку будут присутствовать несколько наблюдений в граничных областях. Более того, в (Berg, Bekken, 2007) показано, что CPIT-подход может давать плохие результаты в случаях, когда рассматриваемые данные расположены радиально асимметрично.

Учитывая указанные недостатки, в (Berg, Bekken, 2007) был предложен новый подход, названный CPIT-2, который позволяет использовать любую весовую функцию в (33) и при помощи дополнительного интегрального условно-вероятностного преобразования, построенного с использованием порядковых статистик, идентифицировать радиальную асимметрию. По существу, в работе (Berg, Bekken, 2007) предлагается на первом шаге применить условное интегрально-вероятностное преобразование к нормированным рангам $\hat{\mathbf{U}}$, а на втором шаге применить условное интегрально-вероятностное преобразование к $\mathbf{V}$, полученному на первом шаге.

Пусть $\mathbf{V} = (V_1,\ldots,V_n)$, где $V_1,\ldots,V_n$ — независимые равномерно распределенные на [0; 1] случайные величины (будем обозначать это $\mathbf{V} \in U(0;1)^n$). Вектор V получен путем применения интегрального условно-вероятностного преобразования к $\hat{\mathbf{U}}$. В работе (Berg, Bekken, 2007) соответствующие порядковые статистики обозначаются как:

$$V_{(1)} \le V_{(2)} \le \ldots \le V_{(n-1)} \le V_{(n)}.$$

Если $V_{(1)},\ldots,V_{(n)}$ — порядковые статистики выборки независимых одинаково распределенных из $U(0;1)$ случайных величин, тогда $V_{(i)}$ имеет бета-распределение с параметрами $(i, n-(i-1))$, см. (D'Agostino, Stephens, 1986, глава 8). Для того чтобы вычислить интегральное условно-вероятностное преобразование для порядковых статистик, Берг и Беккен используют результаты теоремы 2.7 из (David, 1981). С использованием результатов теоремы 1 из (Deheuvels, 1984) и того факта, что в условиях справедливости нулевой гипотезы $\mathbf{V} \in U(0;1)^n$, в работе (Berg, Bekken, 2007) получено следующее выражение для интегрального условно-вероятностного преобразования порядковой статистики $\mathbf{V}$:

$$H_1 = F_{V_{(i)}|V_{(i-1)}}(v_{(i)}) = 1 - \left(\frac{1-v_{(i)}}{1-v_{(i-1)}}\right)^{n-(i-1)}, \quad i = 1,\ldots,n;\ v_{(0)} = 0. \tag{36}$$

Слишком большое или слишком малое значение H_i может означать ошибочность нулевой гипотезы, см. (Glen et al., 2001). Позднее в работе (Berg, Bekken, 2007) получена статистика, основанная на одновременном использовании $\mathbf{V}$ и $\mathbf{H}$:

$$W_{2t} = \sum_{i=1}^{n} \Gamma_V(V_{(i,t)}) \cdot \Gamma_H(H_{it}), \quad t = 1, \ldots, T, \tag{37}$$

где $\Gamma_V(\cdot)$ и $\Gamma_H(\cdot)$ — весовые функции, используемые для взвешивания информации из $\mathbf{V}$ и $\mathbf{H}$ соответственно. Выбор весовых функций зависит от того, какой области определения копула-функции уделяется наибольшее внимание. Например, если $\Gamma_V = (\Phi^{-1}(\cdot))^2$ и $\Gamma_H = 1$, то получим исходный CPIT-подход (33) из (Breymann et al., 2003).

В работе (Berg, Bekken, 2007) указывается на то, что во многих тестах могут быть получены достаточно хорошие результаты при использовании следующей комбинации двух весовых функций:

а) $\Gamma_V(X) = |X - 0.5|$, $\Gamma_H(X) = 1$,

б) $\Gamma_V(X) = (X - 0.5)^2$, $\Gamma_H(X) = 1$.

Более того, в случае сильной асимметрии рекомендуется выбрать $\Gamma_V(X) = |X - 0.5|$, $\Gamma_H(X) = |X - 0.5|$.

Наконец, важно отметить, что в целом распределение W_{2t} неизвестно, и для его оценки необходимо использовать бутстрап приближение. Если вычислить W_{2t} с использованием некоторых весовых функций Γ_V и Γ_H, то можно смоделировать n независимых одинаково распределенных $U(0;1)$ случайных величин $\tilde{V}$, вычислить $\tilde{W}_{2t}$, используя те же самые весовые функции, что и для W_{2t}. Повторяя эту процедуру большое число раз в предположении справедливости нулевой гипотезы, можно аппроксимировать функцию распределения F_{2t} для W_{2t}.

Реализация тестов

Далее представлено подробное описание тестовых процедур, включая вычисление оценок соответствующих p-значений в предположении справедливости нулевой гипотезы о параметрическом виде копула-функции. Среди описанных процедур — CPIT и CPIT-2 подходы.

Процедура 9.1. ***Реализация теста в подходе*** $\mathcal{A}_1$

1. Получить псевдо-наблюдения $(\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_T)$ путем преобразования выборочных данных $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ в нормированные ранги с использованием (32).

2. Оценить параметры копула-функции α с использованием состоятельной ММП-оценки $\hat{\alpha} = \arg\max_{\alpha} l(\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_T; \alpha)$, где l — логарифмическая функция правдоподобия наблюдений $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_T$, т. е. полупараметрического метода.

3. Вычислить $(\mathbf{v}_1, \ldots, \mathbf{v}_T)$, полученные путем применения интегрального условно-вероятностного преобразования к $(\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_T)$, в предположении справедливости нулевой гипотезы о том, что копула-функция равна $C_{\hat{\alpha}}$.

4. Вычислить CPIT-2 выборку $(\mathbf{h}_1, \ldots, \mathbf{h}_T)$, применяя интегральное условно-вероятностное преобразование (36) к $(\mathbf{v}_1, \ldots, \mathbf{v}_T)$.

5. Используя весовые функции Γ_V и Γ_H (если $\Gamma_H = 1$, то W_{1t} и W_{2t} совпадают), вычислить W_{1t} или W_{2t} по формулам (33) или (37) соответственно.

6. Если при справедливости нулевой гипотезы для W_{1t} или W_{2t} функция распределения известна, вычислить $F_1(W_1)$ или $F_2(W_2)$ и перейти к шагу 8. Если распределение неизвестно, аппроксимировать $F_1(W_1)$ или $F_2(W_2)$ следующим образом: для некоторого достаточно большого целого N_b итеративно повторять нижеприведенную процедуру для каждого $l \in \{1,\ldots,N_b\}$.

а) сгенерировать случайную выборку $\mathbf{v}_l^* = (v_{1,l}^*,\ldots,v_{n,l}^*)$, компоненты которой представляют собой компоненты $U(0;1)^n$ случайного вектора;

б) применяя интегральное условно-вероятностное преобразование (36) к $(v_{1,l}^*,\ldots,v_{n,l}^*)$, вычислить $\mathbf{h}_l^* = (h_{1,l}^*,\ldots,h_{n,l}^*)$;

в) в соответствии с (33) или (37), используя те же весовые функции, что и на шаге 5, и беря $(v_{1,l}^*,\ldots,v_{n,l}^*)$ и $(h_{1,l}^*,\ldots,h_{n,l}^*)$, вычислить $\hat{W}_{1,l}^*$ или $\hat{W}_{2,l}^*$ соответственно.

7. Вычислить $F_1(W_1) = \dfrac{1}{N_b+1}\sum_{l=1}^{N_b} \mathbf{1}\{\hat{W}_{1,l}^* > W_1\}$. $F_2(W_2)$ вычисляется аналогичным образом.

8. В соответствии с (34) и (35) вычислить $\hat{T}_1$ (относительно статистики Андерсона–Дарлинга см. (Berg, Bekken, 2007)). $\hat{T}_2$ вычисляется аналогично.

9. Для некоторого достаточно большого целого K повторить следующие шаги для каждого $k = 1,\ldots,K$:

а) получить псевдо-наблюдения $(\hat{\mathbf{u}}_{1,k}^0,\ldots,\hat{\mathbf{u}}_{T,k}^0)$ путем преобразования выборочных данных $(\mathbf{x}_{1,k}^0,\ldots,\mathbf{x}_{T,k}^0)$ в нормированные ранги с использованием (32);

б) по полученным таким образом псевдо-наблюдениям оценить параметры копула-функции α^0 с использованием состоятельной ММП-оценки $\hat{\alpha}_k^0 = \arg\max_{\alpha} l(\hat{\mathbf{u}}_{1,k}^0,\ldots,\hat{\mathbf{u}}_{T,k}^0;\alpha)$, т. е. полупараметрического метода;

в) вычислить $(\mathbf{v}_{1,k}^0,\ldots,\mathbf{v}_{T,k}^0)$, полученные путем применения интегрального условно-вероятностного преобразования к $(\hat{\mathbf{u}}_{1,k}^0,\ldots,\hat{\mathbf{u}}_{T,k}^0)$ в предположении справедливости нулевой гипотезы о том, что копула-функция равна $C_{\hat{\alpha}_k^0}$;

г) вычислить CPIT-2 выборку $(\mathbf{h}_{1,k}^0,\ldots,\mathbf{h}_{T,k}^0)$ путем применения интегрального условно-вероятностного преобразования (36) к $(\mathbf{v}_{1,k}^0,\ldots,\mathbf{v}_{T,k}^0)$;

д) используя такие же весовые функции, как на шаге 5, примененные к $(\mathbf{v}_{1,k}^0,\ldots,\mathbf{v}_{T,k}^0)$ и $(\mathbf{h}_{1,k}^0,\ldots,\mathbf{h}_{T,k}^0)$, вычислить $W_{1,k}^0$ или $W_{2,k}^0$ по формулам (33) или (37) соответственно;

е) если при справедливости нулевой гипотезы для $W_{1,k}^0$ или $W_{2,k}^0$ функция распределения известна, вычислить $F_1(W_{1,k}^0)$ или $F_2(W_{2,k}^0)$ и перейти к шагу (з);

- если распределение неизвестно, аппроксимировать $F_1(\cdot)$ или $F_2(\cdot)$, итеративно повторяя (для некоторого достаточно большого целого N_b) нижеприведенную процедуру для каждого $l \in \{1,\ldots,N_b\}$:
- сгенерировать случайную выборку $\mathbf{v}_{l,k}^{0*} = (v_{1,l,k}^{0*},\ldots,v_{n,l,k}^{0*})$ из $U(0;1)^n$ случайного вектора;
- применяя интегральное условно-вероятностное преобразование (36) к $(v_{1,l,k}^{0*},\ldots,v_{n,l,k}^{0*})$, вычислить $\mathbf{h}_l^{0*} = (h_{1,l,k}^{0*},\ldots,h_{n,l,k}^{0*})$;
- в соответствии с (33) или (37), используя те же весовые функции, что и на шаге 5, и беря $(v_{1,l,k}^{0*},\ldots,v_{n,l,k}^{0*})$ и $(h_{1,l,k}^{0*},\ldots,h_{n,l,k}^{0*})$, вычислить $\hat{W}_{1,l,k}^{0*}$ или $\hat{W}_{2,l,k}^{0*}$ соответственно;

ж) вычислить $F_1(W_{1,k}^0) = \dfrac{1}{N_b+1}\sum_{l=1}^{N_b} \mathbf{1}\{\hat{W}_{1,l,k}^{0*} > W_{1,k}^0\}$. $F_2(W_{2,k}^0)$ вычисляется аналогичным образом;

з) в соответствии с (34) и (35) вычислить $\hat{T}_{1,k}^0$ (относительно статистики Андерсона–Дарлинга см. (Berg, Bekken, 2007)); $\hat{T}_{2,k}^0$ вычисляется аналогично.

10. Для оценки p-значения для подхода $\mathcal{A}_1$ воспользоваться формулой:

$$\hat{p} = \frac{1}{K+1}\sum_{k=1}^{K}\mathbf{1}\{\hat{T}_{1,k}^{0} > \hat{T}_1\}.$$

9.2. Тесты, основанные на использовании эмпирических копула-функций

Подход $\mathcal{A}_2$

Подход $\mathcal{A}_2$, описанный в (Berg, 2009), основан на тесте, предложенном в работе (Genest, Rémillard, 2008). А именно, этот подход основан на использовании эмпирической копула-функции, впервые представленной в (Deheuvels, 1979):

$$\hat{\mathfrak{C}}(\mathbf{u}) = \frac{1}{T+1}\sum_{t=1}^{T}\mathbf{1}\{\hat{u}_{1t} \le u_1,\ldots,\hat{u}_{nt} \le u_n\}, \tag{38}$$

где $\mathbf{u} = (u_1,\ldots,u_n) \in [0;1]^n$. Идея, лежащая в основе этого теста, состоит в сравнении $\hat{\mathfrak{C}}(\cdot)$ с *параметрической* оценкой копула-функции $C_{\hat{\alpha}}(\cdot)$. Это классический подход в статистике, который состоит в построении одномерного критерия согласия, основанного на расстоянии между эмпирической функцией распределения и распределением, предполагаемым при нулевой гипотезе. С использованием статистики Крамера–фон Мизеса, в работе (Genest et al., 2009) получена следующая критическая статистика:

$$\hat{T}_2 = T\int_{[0;1]^d}\left(\hat{\mathfrak{C}}(\hat{\mathbf{u}}) - \mathfrak{C}_{\hat{\alpha}}(\hat{\mathbf{u}})\right)^2 d\hat{\mathfrak{C}}(\hat{\mathbf{u}}) = \sum_{t=1}^{T}\left(\hat{\mathfrak{C}}(\hat{\mathbf{u}}_t) - C_{\hat{\alpha}}(\hat{\mathbf{u}}_t)\right)^2. \tag{39}$$

Далее более подробное представлено описание параметрической бутстрап процедуры вычисления оценок p-значений при тестировании параметрической нулевой гипотезы для копула-функции.

Процедура 9.2. ***Реализация теста в подходе $\mathcal{A}_2$***

1. Получить псевдо-наблюдения $(\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_T)$ путем преобразования выборочных данных $(\mathbf{x}_1,\ldots,\mathbf{x}_T)$ в нормированные ранги с использованием (32).

2. Оценить параметры α копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha} = \arg\max_{\alpha} l(\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_T;\alpha)$, т. е. полупараметрического метода.

3. В соответствии с формулой (38) вычислить $\hat{\mathfrak{C}}(\hat{\mathbf{u}})$.

4. Если имеется аналитическое выражение для $C_\alpha(\cdot)$, то, подставляя $\hat{\mathfrak{C}}(\hat{\mathbf{u}})$ и $C_{\hat{\alpha}}(\hat{\mathbf{u}})$ в (39), вычислить оценку $\hat{T}_2$ и перейти к шагу 5.

Если аналитического выражения для $C_\alpha(\cdot)$ нет, то выбрать достаточно большое натуральное $N_b \ge T$ и реализовать следующие шаги:

а) согласно предположению о справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}_1^*,\ldots,\mathbf{x}_{N_b}^*)$ и, согласно (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}_1^*,\ldots,\hat{\mathbf{u}}_{N_b}^*)$;

б) вычислить аппроксимацию $C_{\hat{\alpha}}$ при помощи

$$C^*_{\hat{\alpha}}(\mathbf{u}) = \frac{1}{N_b+1}\sum_{l=1}^{N_b} \mathbf{1}\{\hat{\mathbf{u}}^*_l \le \mathbf{u}\}, \quad \mathbf{u} \in [0;1]^n;$$

в) по формуле (39) вычислить приближение статистики Крамера–фон Мизеса:

$$\hat{T}_2 = \sum_{t=1}^{T}\left(\hat{C}(\hat{\mathbf{u}}_t) - C^*_{\hat{\alpha}}(\hat{\mathbf{u}}_t)\right)^2.$$

5. При некотором достаточно большом натуральном K для $k = 1,\ldots,K$ последовательно повторить следующие шаги:

а) согласно предположению о справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}^0_{1,k},\ldots,\mathbf{x}^0_{T,k})$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}^0_{1,k},\ldots,\hat{\mathbf{u}}^0_{T,k})$;

б) по полученным таким образом псевдо-наблюдениям оценить параметры α^0 копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha}^0_k = \arg\max_{\alpha^0} l\left(\hat{\mathbf{u}}^0_{1.k},\ldots,\hat{\mathbf{u}}^0_{T.k};\alpha^0\right)$, т.е. с использованием полупараметрического метода;

в) определить $\hat{C}^0_k(u)$ по формуле $\hat{C}^0_k(\mathbf{u}) = \frac{1}{T+1}\sum_{t=1}^{T}\mathbf{1}\{\hat{\mathbf{u}}^0_{t,k} \le \mathbf{u}\}, \quad \mathbf{u} \in [0;1]^n$;

г) если определено аналитическое выражение для $C_\alpha(\cdot)$, то положить:

$$\hat{T}^0_{2,k} = \sum_{t=1}^{T}\left(\hat{C}^0_k(\hat{\mathbf{u}}^0_{t,k}) - C_{\hat{\alpha}^0_k}(\hat{\mathbf{u}}^0_{t,k})\right)^2$$

и перейти к шагу 6; если аналитическое выражение для $C_\alpha(\cdot)$ не определено, то выбрать достаточно большое натуральное $N_b \ge T$ и реализовать следующие шаги:

- согласно предположению о справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}^{0*}_{1,k},\ldots,\mathbf{x}^{0*}_{N_b,k})$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}^{0*}_{1,k},\ldots,\hat{\mathbf{u}}^{0*}_{N_b,k})$;
- вычислить аппроксимацию $C_{\hat{\alpha}^0_k}$ при помощи:

$$C^{0*}_{\hat{\alpha}^0_k}(\mathbf{u}) = \frac{1}{N_b+1}\sum_{l=1}^{N_b}\mathbf{1}\{\hat{\mathbf{u}}^{0*}_{l,k} \le \mathbf{u}\};$$

- по формуле (39) вычислить приближение статистики Крамера–фон Мизеса:

$$\hat{T}^*_{2,k} = \sum_{t=1}^{T}\left(\hat{C}^0_k(\hat{\mathbf{u}}^0_{t,k}) - C^{0*}_{\hat{\alpha}^0_k}(\hat{\mathbf{u}}^0_{t,k})\right)^2.$$

6. Для оценки p-значения в подходе $\mathcal{A}_2$ воспользоваться формулой

$$\hat{p} = \frac{1}{K+1}\sum_{k=1}^{K}\mathbf{1}\{\hat{T}^0_{2,k} > \hat{T}_2\}.$$

Некоторое расширение. В работе (Kole et al., 2007) предложен тест, который можно рассматривать в качестве расширения подхода $\mathcal{A}_2$ на случаи, когда аналитическое выражение (гауссовское, стьюдентовское и т. п.) для копула-функции отсутствует.

В основе этого теста лежит та же идея, что и в (Genest et al., 2009), состоящая в вычислении расстояния между эмпирической и параметрической копула-функциями. Однако формальное представление функции распределения эллиптического типа, как правило, невозможно, и вычисление их значений с вычислительной точки зрения очень трудоемко в условиях большой размерности. В работе (Kole et al., 2007) используется свойство постоянства функции плотности эллиптических распределений на эллипсоидах. Каждой случайной величине, имеющей эллиптическое распределение, отвечает одномерная случайная величина со специфическим распределением, которое соответствует радиальным эллипсоидам, для которых плотность постоянна (см. (Fang et al., 1990)). Вместо того чтобы рассматривать исходные наблюдения, в работе (Kole et al., 2007) рассматриваются квадраты радиусов эллипсоида, отвечающего постоянной плотности. Как следствие, авторы последней упомянутой работы сравнивают эмпирические распределения квадратов радиусов с их теоретическими распределениями, которые представляют собой некоторые стандартные распределения в случае гауссовской и стьюдентовской копула-функции.

Пусть имеется случайный вектор $U = (u_{1t},...,u_{nt})'$, $t = 1,...,T$, с частными равномерными на [0; 1] распределениями, зависимости между компонентами которого определяются *гауссовской копула-функцией* с корреляционной матрицей Σ. В работе (Kole et al., 2007) строится квадрат радиуса следующим образом:

$$Z_{\Phi} = \zeta' \Sigma^{-1} \zeta, \tag{40}$$

где $\zeta = (\Phi^{-1}(u_{1t}),...,\Phi^{-1}(u_{nt}))'$ — вектор значений обратной функции гауссовского одномерного распределения в точках $u_{1t},...,u_{nt}$. Легко показать, что случайная величина Z_{Φ} имеет χ^2_n-распределение.

Пусть $V = (v_1,...,v_n)'$ — случайный вектор, в котором v_i имеет равномерное распределение на [0; 1], а зависимости между компонентами совместного распределения определяются копула-функцией Стьюдента с корреляционной матрицей Σ и числом степеней свободы ν. В (Kole et al., 2007) квадрат радиуса строится следующим образом:

$$Z_{\nu} = \zeta_{\nu}' \Sigma^{-1} \zeta_{\nu} / n, \tag{41}$$

где $\zeta_{\nu} = (t_{\nu}^{-1}(v_1),...,t_{\nu}^{-1}(v_n))'$ — вектор значений обратной функции одномерного распределения Стьюдента с числом степеней свободы ν. Случайная величина Z_{ν} имеет F-распределение с параметрами n и ν, т. к. ζ_{ν} имеет распределение Стьюдента и может быть представлено как $\zeta_{\nu} = W / \sqrt{S / n}$, где W — n-мерная нормальная случайная величина с ковариационной матрицей Σ, а S — одномерная случайная величина с χ^2_{ν}-распределением. Как следствие, для Z_{ν} из (41) имеем:

$$Z_{\nu} = \frac{W' \Sigma^{-1} W / n}{S / \nu}.$$

Это представление показывает, что Z_ν есть отношение двух независимых χ^2-распределенных случайных величин, каждая из которых делится на соответствующее ей количество степеней свободы. Тем самым Z_ν имеет $F_{n,\nu}$-распределение.

Практическая реализация этого расширения подхода $\mathcal{A}_2$ аналогична реализации самого $\mathcal{A}_2$ и состоит из четырех шагов.

1. При помощи IFM-метода (Фантаццини, 2011b) оценить параметры α копула-функции следующим образом:

а) для моделирования частных распределений воспользоваться полупараметрическим методом из работы (Danielsson, de Vries, 2000), который состоит в том, чтобы моделировать центральную часть функции распределения при помощи эмпирического распределения, а хвосты — при помощи одномерных распределений из теории экстремальных значений; в частности, для моделирования хвостов можно использовать распределение Парето; соответственно, хвосты определяются левым эмпирическим 0.01-квантилем и правым эмпирическим 0.99-квантилем;

б) при помощи метода максимума правдоподобия оценить параметры α копула-функции.

2. Оценить качество выбранной параметрической копула-функции, вычисляя ее расстояние до эмпирической копула-функции; если копула-функция принадлежит эллиптическому семейству (например, гауссовская или стьюдентовская), то с использованием (40) или (41) воспользоваться одним из стандартных критериев согласия. В работе (Kole et. al., 2007) используются следующие меры расстояния:

$$D_{KS}^m = \max_t |\Im_E - F_H|,$$

$$D_{KS}^a = \int_x |\Im_E - F_H| \, dF_H,$$

$$D_{AD}^m = \max_t \frac{|\Im_E - F_H|}{\sqrt{F_H(1-F_H)}},$$

$$D_{AD}^a = \int_x \frac{|\Im_E - F_H|}{\sqrt{F_H(1-F_H)}} dF_H,$$

где $\Im_E$ — эмпирическая функция распределения, а F_H — гипотетическое распределение. Первая мера расстояния — расстояние Колмогорова–Смирнова, второе расстояние — среднее абсолютное отклонение, третья мера — расстояние Андерсона–Дарлинга, четвертое — аналог расстояния Андерсона–Дарлинга с усреднением. Для того чтобы уменьшить влияние выбросов при использовании расстояний Андерсона–Дарлинга, в работе (Kote et al., 2007), следуя (Malevergne, Sornette, 2003), исходное $|\mathfrak{F}_E - F_H|$ заменено на $(\mathfrak{F}_E - F_H)^2$.

3. Если форма меры расстояния известна, а параметры копула-функции оценены, для вычисления распределения случайных величин, отвечающих указанным выше четырем расстояниям, в работе (Kole et al., 2007) используется параметрическая бутстрап процедура с K итерациями, это отвечает в точности шагу 5 в реализации подхода $\mathcal{A}_2$. Отме-

тим, что двухшаговый IFM-метод также должен быть использован в упомянутой бутстрап процедуре.

4. Подобно шагу 6 в реализации подхода $\mathcal{A}_2$, для оценки p-значений использовать полученное на предыдущем шаге распределение для мер расстояний, т. е. подсчитать, сколько раз значения меры расстояния, вычисленные с использованием бутстрап выборок, превышают значение, отвечающее исходной выборке, и поделить это число на $K+1$.

Подход $\mathcal{A}_3$

В работе (Genest et al., 2009) предложено применить предыдущий подход $\mathcal{A}_2$ к случайным величинам, подвергнутым преобразованию Розенблатта, т. е. к $\mathbf{V}=(\mathbf{v}_1,\ldots,\mathbf{v}_T)$. Следующий шаг состоит в сравнении $\hat{\mathfrak{C}}(\mathbf{v})$ с копула-функцией $C_\perp(\mathbf{v})$, отвечающей независимости, с использованием статистики Крамера–фон Мизеса, см. (Genest et al., 2009):

$$\hat{T}_3 = T\int_{[0;1]^d}\left(\hat{\mathfrak{C}}(\mathbf{v})-C_\perp(\mathbf{v})\right)^2 d\hat{\mathfrak{C}}(\mathbf{v}) = \sum_{t=1}^{T}\left(\hat{\mathfrak{C}}(\mathbf{v}_t)-C_\perp(\mathbf{v}_t)\right)^2. \qquad (42)$$

Процедура 9.3. ***Реализация теста в подходе*** $\mathcal{A}_3$

Получить псевдо-наблюдения $(\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_T)$ путем преобразования выборочных данных $(\mathbf{x}_1,\ldots,\mathbf{x}_T)$ в нормированные ранги с использованием (32).

Оценить параметры α копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha}=\arg\max_{\alpha} l(\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_T;\alpha)$, т. е. с использованием полупараметрического метода.

В предположении справедливости нулевой гипотезы относительно копула-функции $C_{\hat{\alpha}}$ вычислить интегральное условно-вероятностное преобразование (CPIT) выборочных данных $(\mathbf{v}_1,\ldots,\mathbf{v}_T)$ путем применения CPIT к $(\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_T)$.

Согласно формуле (38) вычислить $\hat{\mathfrak{C}}(\mathbf{v})$.

Согласно формуле (42) вычислить $\hat{T}_3$.

При некотором достаточно большом натуральном K для $k=1,2,\ldots,K$ последовательно повторить следующие шаги:

а) отправляясь от справедливости нулевой гипотезы, сгенерировать случайную выборку $(\mathbf{x}^0_{1,k},\ldots,\mathbf{x}^0_{T,k})$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}^0_{1,k},\ldots,\hat{\mathbf{u}}^0_{T,k})$;

б) по сгенерированной таким образом псевдовыборке оценить параметры α^0 копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha}^0_k=\arg\max_{\alpha^0} l(\hat{\mathbf{u}}^0_{1,k},\ldots,\hat{\mathbf{u}}^0_{T,k};\alpha^0)$, т. е. с использованием полупараметрического метода;

в) в предположении справедливости нулевой гипотезы о параметрическом виде копула-функции $C_{\hat{\alpha}^0_k}$ вычислить интегральное условно-вероятностное преобразование выборочных данных $(\mathbf{v}^0_{1,k},\ldots,\mathbf{v}^0_{T,k})$, применяя CPIT к $(\hat{\mathbf{u}}^0_{1,k},\ldots,\hat{\mathbf{u}}^0_{T,k})$;

г) определить $\hat{\mathfrak{C}}^0_k(\mathbf{u})=\dfrac{1}{T+1}\sum_{t=1}^{T}\mathbf{1}\{\mathbf{v}^0_{t,k}\le\mathbf{u}\}, \quad \mathbf{u}\in[0;1]^n$;

д) вычислить $\hat{T}^0_{3,k}=\sum_{t=1}^{T}\left(\hat{\mathfrak{C}}^0_k(\mathbf{v}^0_{t,k})-C_\perp(\mathbf{v}^0_{t,k})\right)^2$.

Вычислить оценку p-значения следующим образом:

$$\hat{p} = \frac{1}{K+1}\sum_{k=1}^{K}\mathbf{1}\{\hat{T}_{3,k}^{0} > \hat{T}_3\}.$$

Подход $\mathcal{A}_4$

В работах (Genest, Rivest, 1993; Wang, Wells, 2000; Savu, Trede, 2008; Genest et al., 2006a) для построения критериев согласия для копула-функций предлагается использовать так называемую функцию зависимости Кендалла $K(w) = \mathrm{P}\left(C(\hat{\mathbf{U}}) \le w\right)$ (см. (Фантаццини, 2011b)). Тестовый датчик S_4 для подхода $\mathcal{A}_4$ определяется в форме этой функции, т. е.

$$S_4(w) = \mathrm{P}\left(C(\hat{\mathbf{U}}) \le w\right), \quad w \in [0;1] \ ,$$

где $\hat{\mathbf{U}}$ — псевдовектор, каждая компонента которого составлена из соответствующих нормированных рангов. При справедливости нулевой гипотезы $S_4(w) = S_{4,\hat{\alpha}}(w)$, где $S_{4,\hat{\alpha}}(w)$ — функция зависимости Кендалла, рассчитанная в условиях нулевой гипотезы H_0 и зависящая от параметров соответствующей копула-функции. Непараметрическая оценка для тестового датчика S_4 равна:

$$\hat{S}_4(w) = \frac{1}{T+1}\sum_{t=1}^{T}\mathbf{1}\{\hat{\mathfrak{C}}(\hat{\mathbf{u}}_t) \le w\}. \tag{43}$$

Статистика Крамера–фон Мизеса вычисляется следующим образом:

$$\hat{T}_4 = T\int_0^1 \left(\hat{S}_4(w) - S_{4,\hat{\alpha}}(w)\right)^2 d\hat{S}_4(w) = \sum_{t=1}^{T}\left(\hat{S}_4\left(\frac{t}{T+1}\right) - S_{4,\hat{\alpha}}\left(\frac{t}{T+1}\right)\right)^2 . \tag{44}$$

Процедура 9.4. ***Реализация теста в подходе*** $\mathcal{A}_4$

1. Получить псевдо-наблюдения $(\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_T)$ путем преобразования выборочных данных $(\mathbf{x}_1,\ldots,\mathbf{x}_T)$ в нормированные ранги с использованием (32).

2. Оценить параметры α копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha} = \arg\max_{\alpha} l(\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_T;\alpha)$, т. е. полупараметрического метода.

3. В соответствии с формулой (38) вычислить $\hat{\mathfrak{C}}(\hat{\mathbf{u}})$.

4. Если для $S_{4,\alpha}$ можно получить аналитическую выражение, то в соответствии с (43) и (44) вычислить статистику $\hat{T}_4$, а затем перейти к шагу 5. Если аналитическое выражение для $S_{4,\alpha}$ получить невозможно, то выбрать $N_b \ge T$ и реализовать следующие шаги:

а) согласно предположению о справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}_1^*,\ldots,\mathbf{x}_{N_b}^*)$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}_1^*,\ldots,\hat{\mathbf{u}}_{N_b}^*)$;

б) вычислить аппроксимацию $S_{4,\hat{\alpha}}$ при помощи $\hat{S}_4^*(w) = \frac{1}{N_b+1}\sum_{l=1}^{N_b}\mathbf{1}\{\hat{\mathfrak{C}}^*(\hat{\mathbf{u}}_l^*) \le w\}$, где

$$\hat{\mathfrak{C}}^*(\mathbf{u}) = \frac{1}{N_b+1}\sum_{l=1}^{N_b}\mathbf{1}\{\hat{\mathbf{u}}_l^* \le \mathbf{u}\}, \quad \mathbf{u} \in [0;1]^n;$$

в) по формуле (44) вычислить приближение для статистики Крамера–фон Мизеса:

$$\hat{T}_4 = \frac{T}{N_b}\sum_{l=1}^{N_b}\left(\hat{S}_4\left(\hat{\mathfrak{C}}^*(\hat{\mathbf{u}}_l^*)\right) - \hat{S}_4^*\left(\hat{\mathfrak{C}}^*(\hat{\mathbf{u}}_l^*)\right)\right)^2.$$

5. При некотором достаточно большом натуральном K для $k = 1,\ldots,K$ последовательно повторить следующие шаги:

а) согласно предположению нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}_{1,k}^0,\ldots,\mathbf{x}_{T,k}^0)$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}_{1,k}^0,\ldots,\hat{\mathbf{u}}_{T,k}^0)$;

б) оценить параметры α^0 копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha}_k^0 = \arg\max_{\alpha^0} l(\hat{\mathbf{u}}_{1,k}^0,\ldots,\hat{\mathbf{u}}_{T,k}^0;\alpha^0)$, т. е. с использованием полупараметрического метода;

в) определить $\hat{S}_{4,k}^0(w) = \frac{1}{T+1}\sum_{t=1}^{T}\mathbf{1}\{\hat{\mathfrak{C}}_k^0(\hat{\mathbf{u}}_{t,k}^0 \le w\}$, где $\hat{\mathfrak{C}}_k^0(\mathbf{u}) = \frac{1}{T+1}\sum_{t=1}^{T}\mathbf{1}\{\hat{\mathbf{u}}_{t,k}^0 \le \mathbf{u}\}$, $\mathbf{u}\in[0;1]^n$;

г) если для $S_{4,\alpha}$ можно получить аналитическое выражение, то с использованием $\hat{S}_{4,k}^0$ и $S_{4,\hat{\alpha}_k^0}$ в формуле (44) вычислить $\hat{T}_{4,k}^0$ и перейти к шагу 6; если аналитическое выражение для $S_{4,\alpha}$ получить невозможно, тогда выбрать достаточно большое натуральное $N_b \ge T$ и реализовать следующие шаги:

- в условиях справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}_{1,k}^{0*},\ldots,\mathbf{x}_{N_b,k}^{0*})$ и в соответствии с (32) вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}_{1,k}^{0*},\ldots,\hat{\mathbf{u}}_{N_b,k}^{0*})$;
- вычислить аппроксимацию $S_{4,\hat{\alpha}_k^0}$ по формуле: $\hat{S}_{4,k}^{0*}(w) = \frac{1}{N_b+1}\sum_{l=1}^{N_b}\mathbf{1}\{\hat{\mathfrak{C}}_k^{0*}(\hat{\mathbf{u}}_{l,k}^{0*}) \le w\}$, где $\hat{\mathfrak{C}}_k^{0*}(\mathbf{u}) = \frac{1}{N_b+1}\sum_{l=1}^{N_b}\mathbf{1}\{\hat{\mathbf{u}}_{l,k}^{0*} \le \mathbf{u}\}$, $\mathbf{u}\in[0;1]^n$;
- по формуле (44) вычислить приближение для статистики Крамера–фон Мизеса:

$$\hat{T}_{4,k}^0 = \frac{T}{N_b}\sum_{l=1}^{N_b}\left(\hat{S}_{4,k}^0\left(\hat{\mathfrak{C}}_k^{0*}(\hat{\mathbf{u}}_{l,k}^{0*})\right) - \hat{S}_{4,k}^{0*}\left(\hat{\mathfrak{C}}_k^{0*}(\hat{\mathbf{u}}_{l,k}^{0*})\right)\right)^2.$$

6. Вычислить оценку p-значения с использованием формулы:

$$\hat{p} = \frac{1}{K+1}\sum_{k=1}^{K}\mathbf{1}\{\hat{T}_{4,k}^0 > \hat{T}_4\}.$$

Подход $\mathcal{A}_5$

В работе (Quessy et al., 2007) предлагается критерий согласия для двумерных копула-функций, который основан на функции зависимости Спирмена $L_2(w) = \mathrm{P}\left(\hat{U}_1\hat{U}_2 \le w\right)$. При этом $\mathrm{P}\left(\hat{U}_1\hat{U}_2 \le w\right) = \mathrm{P}\left(C_\perp(\hat{U}_1\hat{U}_2) \le w\right)$. Этот подход обобщается на случай произвольной размерности n, при этом $L_n(w) = \mathrm{P}\left(C_\perp(\hat{\mathbf{U}}) \le w\right)$ и тестовый датчик S_5 подхода $\mathcal{A}_5$ определяется соотношением:

$$S_5(w) = \mathrm{P}\left(C_\perp(\hat{\mathbf{U}}) \le w\right), \qquad w\in[0;1],$$

где $\hat{U}$ — псевдовектор, каждая компонента которого составлена из соответствующих нормированных рангов. При справедливости нулевой гипотезы (H_0), $S_5(w) = S_{5,\hat{\alpha}}(w)$, где

$S_{5,\hat{\alpha}}(w)$ — функция зависимости Кендалла, рассчитанная при справедливости H_0 и зависящая от параметров соответствующей копула-функции. Возьмем непараметрическую оценку для S_5:

$$\hat{S}_5(w) = \frac{1}{T+1}\sum_{t=1}^{T}\mathbf{1}\{C_\perp(\hat{\mathbf{u}}_t) \le w\} \ . \tag{45}$$

Статистика Крамера–фон Мизеса вычисляется следующим образом:

$$\hat{T}_5 = T\int_0^1 \left(\hat{S}_5(w) - S_{5,\hat{\alpha}}(w)\right)^2 d\hat{S}_5(w) = \sum_{t=1}^{T}\left(\hat{S}_5\left(\tfrac{t}{T+1}\right) - S_{5,\hat{\alpha}}\left(\tfrac{t}{T+1}\right)\right)^2. \tag{46}$$

Процедура 9.5. ***Реализация теста в подходе*** $\mathcal{A}_5$

1. Получить псевдо-наблюдения $(\hat{\mathbf{u}}_1,...,\hat{\mathbf{u}}_T)$ путем преобразования выборочных данных $(\mathbf{x}_1,...,\mathbf{x}_T)$ в нормированные ранги с использованием (32).

2. Оценить параметры α копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha} = \arg\max_{\alpha} l(\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_T;\alpha)$, т. е. полупараметрического метода.

3. Если для $S_{5,\alpha}$ существует аналитическое выражение, то, используя формулы (45) и (46), вычислить $\hat{T}^0_{5,k}$ и перейти к шагу 4. Если аналитическое выражение для $S_{5,\alpha}$ получить невозможно, тогда выбрать достаточно большое натуральное $N_b \ge T$ и реализовать следующие шаги:

а) согласно предположению о справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}^*_1,...,\mathbf{x}^*_{N_b})$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}^*_1,...,\hat{\mathbf{u}}^*_{N_b})$;

б) вычислить аппроксимацию $S_{5,\hat{\alpha}}$ при помощи: $\hat{S}^*_5(w) = \frac{1}{N_b+1}\sum_{l=1}^{N_b}\mathbf{1}\{C_\perp(\hat{\mathbf{u}}^*_l) \le w\}$;

в) по формуле (46) вычислить приближение для статистики Крамера–фон Мизеса:

$$\hat{T}_5 = \frac{T}{N_b}\sum_{l=1}^{N_b}\left(\hat{S}_5\left(C_\perp(\hat{\mathbf{u}}^*_l)\right) - \hat{S}^*_5\left(C_\perp(\hat{\mathbf{u}}^*_l)\right)\right)^2.$$

4. При некотором достаточно большом натуральном K для $k = 1,...,K$ последовательно повторить следующие шаги:

а) в условиях справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}^0_{1,k},...,\mathbf{x}^0_{T,k})$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}^0_{1,k},\ldots,\hat{\mathbf{u}}^0_{T,k})$;

б) оценить параметры α^0 копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha}^0_k = \arg\max_{\alpha^0} l(\hat{\mathbf{u}}^0_{1,k},\ldots,\hat{\mathbf{u}}^0_{T,k};\alpha^0)$, т. е. полупараметрического метода;

в) определить $\hat{S}^0_{5,k}(w) = \frac{1}{T+1}\sum_{t=1}^{T}\mathbf{1}\{C_\perp(\hat{\mathbf{u}}^0_{t,k}) \le w\}$;

г) если для $S_{5,\alpha}$ существует аналитическое выражение, то с использованием $\hat{S}^0_{5,k}$ и $S_{5,\hat{\alpha}^0_k}$ по формуле (46) вычислить $\hat{T}^0_{5,k}$ и перейти к шагу 5; если аналитическое выражение для $S_{5,\alpha}$

получить невозможно, то выбрать достаточно большое натуральное $N_b \geq T$ и реализовать следующие шаги:

- в условиях справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}_{1,k}^{0*},\ldots,\mathbf{x}_{N_b,k}^{0*})$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}_{1,k}^{0*},\ldots,\hat{\mathbf{u}}_{N_b,k}^{0*})$;
- вычислить аппроксимацию $S_{5,\hat{\alpha}_k^0}$ при помощи формулы

$$\hat{S}_{5,k}^{0*}(w) = \frac{1}{N_b+1}\sum_{l=1}^{N_b}\mathbf{1}\{C_\perp(\hat{\mathbf{u}}_{l,k}^{0*}) \leq w\};$$

- по формуле (46) вычислить приближение для статистики Крамера–фон Мизеса

$$\hat{T}_{5,k}^0 = \frac{T}{N_b}\sum_{l=1}^{N_b}\left(\hat{S}_{5,k}^0\left(C_\perp(\hat{\mathbf{u}}_{l,k}^{0*})\right) - \hat{S}_{5,k}^{0*}\left(C_\perp(\hat{\mathbf{u}}_{l,k}^{0*})\right)\right)^2.$$

5. Вычислить оценку p-значения по формуле

$$\hat{p} = \frac{1}{K+1}\sum_{k=1}^{K}\mathbf{1}\{\hat{T}_{5,k}^0 > \hat{T}_5\}.$$

9.3. Другие подходы к построению критериев согласия

Подход $\mathcal{A}_6$

В работе (Shih, 1998) предложен основанный на моментных тождествах критерий согласия для двумерных моделей, задаваемых копула-функцией Клейтона. В ней рассматриваются невзвешенные и взвешенные оценки параметра зависимости α с использованием τ-Кендалла и взвешенной ранговой оценки, а именно:

$$\hat{\alpha}_\tau = \frac{2\hat{\tau}}{1-\hat{\tau}} \quad \text{и} \quad \hat{\alpha}_W = \frac{\sum_{i<j}\Delta_{ij}/W_{ij}}{\sum_{i<j}(1-\Delta_{ij})/W_{ij}}, \tag{47}$$

где $\hat{\tau} = -1 + 4\sum_{i<j}\Delta_{ij}/T(T-1)$, $\Delta_{ij} = \mathbf{1}\{(\hat{U}_{1i}-\hat{U}_{1j})(\hat{U}_{2i}-\hat{U}_{2j}) > 0\}$,

а $W_{ij} = \sum_{t=1}^{T}\mathbf{1}\{\hat{U}_{1t} \leq \max(\hat{U}_{1i},\hat{U}_{1j}), \hat{U}_{2t} \leq \max(\hat{U}_{2i},\hat{U}_{2j})\}$.

Учитывая, что $\hat{\alpha}_\tau$ и $\hat{\alpha}_W$ являются несмещенными оценками для α, в предположении справедливости нулевой гипотезы ($C = C_\alpha$ для $\alpha \geq 0$) в (Shih, 1998) была предложена следующая критическая статистика:

$$\hat{T}_{\text{Shih}} = \sqrt{T}(\hat{\alpha}_\tau - \hat{\alpha}_W).$$

В (Shih, 1998) показано, что при справедливости нулевой гипотезы приведенная выше статистика имеет асимптотически нормальное распределение. Однако позднее в (Genest et al.,

2006b) было указано, что представленная в (Shih, 1998) формула для дисперсии этого асимптотического нормального распределения неверна, и предложен исправленный вариант этой формулы. В работе (Berg, 2009) предлагается расширить этот подход на случай произвольной размерности n путем сравнения $\hat{\alpha}_\tau$ и $\hat{\alpha}_W$ для каждой пары случайных величин. Как следствие, результирующий вектор, составленный из $n(n-1)/2$ статистик, будет иметь асимптотическое $n(n-1)/2$-мерное нормальное распределение с невырожденной ковариационной матрицей, формула для которой пока не выведена. Если умножить вектор статистик на матрицу, обратную к квадратному корню ковариационной матрицы, то результирующий нормированный вектор имеет асимптотическое стандартное нормальное распределение, а сумма квадратов нормированных статистик будет иметь χ^2-распределение с $n(n-1)/2$ степенями свободы.

Не имея формулы ковариационной матрицы, Berg (2009) вычисляет ненормированные суммы квадратов и использует параметрическую бутстрап процедуру для оценки p-значения. Критическая статистика для подхода $\mathcal{A}_6$ равна:

$$\hat{T}_6 = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\hat{\alpha}_{\tau,ij} - \hat{\alpha}_{W,ij} \right)^2. \tag{48}$$

Отметим, что $\hat{\alpha}_W$ и в целом подход $\mathcal{A}_6$ может быть использован только для тестирования копула-функции Клейтона.

Процедура 9.6. ***Реализация теста в подходе*** $\mathcal{A}_6$

1. Получить псевдо-наблюдения $(\hat{\mathbf{u}}_1,...,\hat{\mathbf{u}}_T)$ путем преобразования выборочных данных $(\mathbf{x}_1,...,\mathbf{x}_T)$ в нормированные ранги с использованием (32).
2. Оценить параметры α копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha} = \arg\max_\alpha l(\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_T;\alpha)$, т. е. полупараметрического метода.
3. В соответствии с (47) вычислить параметры $\hat{\alpha}_\tau$ и $\hat{\alpha}_W$.
4. В соответствии с (48) вычислить $\hat{T}_6$.
5. При некотором достаточно большом натуральном K для $k = 1,...,K$ последовательно повторить следующие шаги:

а) в условиях справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}^0_{1,k},...,\mathbf{x}^0_{T,k})$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}^0_{1,k},\ldots,\hat{\mathbf{u}}^0_{T,k})$;

б) в соответствии с (47) вычислить $\hat{\alpha}^0_{\tau,k}$ и $\hat{\alpha}^0_{W,k}$;

в) в соответствии с (48) и используя полученные на предыдущем шаге $\hat{\alpha}^0_{\tau,k}$ и $\hat{\alpha}^0_{W,k}$, вычислить $\hat{T}^0_{6,k}$.

Вычислить оценку p-значения, используя формулу

$$\hat{p} = \frac{1}{K+1} \sum_{k=1}^{K} \mathbf{1}\{\hat{T}^0_{6,k} > \hat{T}_6\}.$$

Подход $\mathcal{A}_7$

В работе (Panchenko, 2005) предложен тест, основанный на скалярном произведении $\hat{\mathbf{U}}$ и $\hat{\mathbf{U}}_{\hat{\alpha}}$, где $\hat{\mathbf{U}}$ — псевдовектор, каждая компонента которого равна соответствующему нор-

мированному рангу, а $\hat{\mathbf{U}}_{\hat{\alpha}}$ — аналог вектора $\hat{\mathbf{U}}$, отвечающий нулевой гипотезе (при этом $\hat{\alpha}$ — состоятельная оценка параметра копула-функции). Идея этого теста состоит в использовании скалярного произведения в качестве меры расстояния между двумя векторами. Определим квадрат расстояния Q между двумя векторами формулой

$$Q = \langle \hat{\mathbf{U}} - \hat{\mathbf{U}}_{\hat{\alpha}} \,|\, \kappa_n \,|\, \hat{\mathbf{U}} - \hat{\mathbf{U}}_{\hat{\alpha}} \rangle^{5},$$

где κ_n — положительно определенное симметричное ядро; в частности, в качестве κ_n может быть взято гауссовское ядро:

$$\kappa_n(\hat{\mathbf{U}}, \hat{\mathbf{U}}') = \exp\{-\| \hat{\mathbf{U}} - \hat{\mathbf{U}}' \|^2 / (2nh^2)\},$$

где $\|\cdot\|$ обозначает евклидову норму в R^n; $h > 0$ — ширина «окна». Очевидно, что Q будет равно нулю тогда и только тогда, когда $\hat{\mathbf{U}} = \hat{\mathbf{U}}_{\hat{\alpha}}$.

Если имеются случайные выборки $(\hat{\mathbf{u}}_1, ..., \hat{\mathbf{u}}_T)$ из $\hat{\mathbf{U}}$, то следующим шагом является генерация случайных выборок $(\hat{\mathbf{u}}_1^*, ..., \hat{\mathbf{u}}_T^*)$ вектора $\hat{\mathbf{U}}_{\hat{\alpha}}$, отвечающего нулевой гипотезе. Скалярное произведение Q может быть представлено в виде $Q = Q_{11} - 2Q_{12} + Q_{22}$, а каждое слагаемое этого разложения может быть оценено с использованием V-статистики (в качестве введения в теорию U- и V-статистик см. (Denker, Keller, 1983)). Критическая статистика подхода $\mathcal{A}_7$ имеет вид:

$$\hat{T}_7 = \frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} \kappa_n(\hat{\mathbf{u}}_i, \hat{\mathbf{u}}_j) - \frac{2}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} \kappa_n(\hat{\mathbf{u}}_i, \hat{\mathbf{u}}_j^*) + \frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} \kappa_n(\hat{\mathbf{u}}_i^*, \hat{\mathbf{u}}_j^*). \qquad (49)$$

Процедура 9.7. ***Реализация теста в подходе*** $\mathcal{A}_7$

1. Получить псевдо-наблюдения $(\hat{\mathbf{u}}_1, ..., \hat{\mathbf{u}}_T)$ путем преобразования выборочных данных $(\mathbf{x}_1, ..., \mathbf{x}_T)$ в нормированные ранги с использованием (32).

2. Оценить параметры α копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha} = \arg\max_{\alpha} l(\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_T; \alpha)$, т. е. полупараметрического метода.

3. В условиях справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}_1^*, ..., \mathbf{x}_T^*)$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}_1^*, ..., \hat{\mathbf{u}}_T^*)$.

4. В соответствии с (49), используя $(\hat{\mathbf{u}}_1, ..., \hat{\mathbf{u}}_T)$ и $(\hat{\mathbf{u}}_1^*, ..., \hat{\mathbf{u}}_T^*)$, вычислить $\hat{T}_7$.

5. При некотором достаточно большом натуральном K для $k = 1, ..., K$ последовательно повторить следующие шаги:

а) в условиях справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}_{1,k}^0, ..., \mathbf{x}_{T,k}^0)$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}_{1,k}^0, ..., \hat{\mathbf{u}}_{T,k}^0)$;

б) оценить параметры α^0 копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha}_k^0 = \arg\max_{\alpha} l(\hat{\mathbf{u}}_{1,k}^0, \ldots, \hat{\mathbf{u}}_{T,k}^0; \alpha^0)$, т. е. полупараметрического метода;

[5] Определение и подробности см. в (Panchenko, 2005).

в) используя копула-функцию $C_{\hat{\alpha}_k^0}$, отвечающую нулевой гипотезе, сгенерировать случайную выборку $(\mathbf{x}_{1,k}^{0*},...,\mathbf{x}_{T,k}^{0*})$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}_{1,k}^{0*},...,\hat{\mathbf{u}}_{T,k}^{0*})$;

г) в соответствии с (49) и с использованием $(\hat{\mathbf{u}}_{1,k}^{0},...,\hat{\mathbf{u}}_{T,k}^{0})$ и $(\hat{\mathbf{u}}_{1,k}^{0*},...,\hat{\mathbf{u}}_{T,k}^{0*})$ вычислить $\hat{T}_{7,k}^{0}$.

6. Вычислить оценку p-значения по формуле

$$\hat{p} = \frac{1}{K+1}\sum_{k=1}^{K}\mathbf{1}\{\hat{T}_{7,k}^{0} > \hat{T}_7\}.$$

Подход $\mathcal{A}_8$

В работе (Berg, 2009) предложено расширение подхода $\mathcal{A}_7$, которое состоит в применении подхода $\mathcal{A}_7$ к величинам, подвергнутым преобразованию Розенблатта, т.е. к $\mathbf{V} = (\mathbf{v}_1,...,\mathbf{v}_T)$. Если имеются выборки $(\mathbf{v}_1^*,...,\mathbf{v}_T^*)$, извлеченные из копула-функции, отвечающей случаю независимости, статистика подхода $\mathcal{A}_8$ равна:

$$\hat{T}_8 = \frac{1}{T^2}\sum_{i=1}^{T}\sum_{j=1}^{T}\kappa_n(\mathbf{v}_i,\mathbf{v}_j) - \frac{2}{T^2}\sum_{i=1}^{T}\sum_{j=1}^{T}\kappa_n(\mathbf{v}_i,\mathbf{v}_j^*) + \frac{1}{T^2}\sum_{i=1}^{T}\sum_{j=1}^{T}\kappa_n(\mathbf{v}_i^*,\mathbf{v}_j^*). \qquad (50)$$

В работе (Berg, 2009) отмечается, что может показаться странным желание строить выводы относительно отклонения от проверяемой (нулевой) гипотезы на основании единственной выборки, отвечающей нулевой гипотезе. Тем не менее, этот подход изначально рассматривался в работе (Panchenko, 2005), и Берг (2009) предпочел рассматривать этот тест именно в таком виде. Более того, подход $\mathcal{A}_8$ направлен на то, чтобы проверить эффект преобразования Розенблатта в случае использования подхода $\mathcal{A}_7$.

Процедура 9.8. ***Реализация теста в подходе*** $\mathcal{A}_8$

1. Получить псевдо-наблюдения $(\hat{\mathbf{u}}_1,...,\hat{\mathbf{u}}_T)$ путем преобразования выборочных данных $(\mathbf{x}_1,...,\mathbf{x}_T)$ в нормированные ранги с использованием (32).

2. Оценить параметры α копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha} = \arg\max_{\alpha} l(\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_T;\alpha)$, т.е. полупараметрического метода.

3. В предположении справедливости нулевой гипотезы относительно копула-функции $C_{\hat{\alpha}}$ вычислить интегральное условно-вероятностное преобразование (CPIT) выборочных данных $(\mathbf{v}_1,...,\mathbf{v}_T)$ путем применения CPIT к $(\hat{\mathbf{u}}_1,...,\hat{\mathbf{u}}_T)$.

4. Сгенерировать выборку $(\mathbf{v}_1^*,...,\mathbf{v}_T^*)$ из копула-функции, отвечающей независимости.

5. В соответствии с (50) и с использованием $(\mathbf{v}_1,...,\mathbf{v}_T)$ и $(\mathbf{v}_1^*,...,\mathbf{v}_T^*)$ вычислить $\hat{T}_8$.

6. При некотором достаточно большом натуральном K для $k = 1,...,K$ последовательно повторить следующие шаги:

а) в условиях справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}_{1,k}^{0},...,\mathbf{x}_{T,k}^{0})$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}_{1,k}^{0},\ldots,\hat{\mathbf{u}}_{T,k}^{0})$;

б) оценить параметры α^0 копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha}_k^0 = \arg\max_{\alpha} l(\hat{\mathbf{u}}_{1,k}^{0},\ldots,\hat{\mathbf{u}}_{T,k}^{0};\alpha^0)$, т.е. полупараметрического метода;

в) в предположении справедливости нулевой гипотезы относительно копула-функции $C_{\hat{\alpha}_k^0}$ вычислить интегральное условно-вероятностное преобразование (CPIT) выборочных данных $(\mathbf{v}_{1,k}^0,...,\mathbf{v}_{T,k}^0)$ путем применения CPIT к $(\hat{\mathbf{u}}_{1,k}^0,...,\hat{\mathbf{u}}_{T,k}^0)$;

г) используя копула-функцию, отвечающую независимости, сгенерировать случайную выборку $(\mathbf{v}_{1,k}^{0*},...,\mathbf{v}_{T,k}^{0*})$;

д) в соответствии с (50), используя $(\mathbf{v}_{1,k}^0,...,\mathbf{v}_{T,k}^0)$ и $(\mathbf{v}_{1,k}^{0*},...,\mathbf{v}_{T,k}^{0*})$, вычислить $\hat{T}_{8,k}^0$.

7. Вычислить оценку p-значения по формуле

$$\hat{p} = \frac{1}{K+1}\sum_{k=1}^{K}\mathbf{1}\{\hat{T}_{8,k}^0 > \hat{T}_8\}.$$

9.4. Тесты, основанные на усреднении критических статистик

Выше были рассмотрены некоторые критерии согласия для копула-функций. Справедливо задаться вопросом, а может ли усреднение этих тестов (с разными способами измерения отклонений от нулевой гипотезы) дать более устойчивые результаты? Ответ на этот вопрос дается в работе (Berg, 2009), где подход с усреднением обозначен как $\mathcal{A}_9$.

Понятно, что усреднение должно осуществляться по стандартизированным переменным. Более того, следует использовать оптимальные веса для расчета взвешенного среднего. Тем не менее, в силу некоторых вычислительных аспектов, в работе (Berg, 2009) представлены только два варианта усреднения:

первый вариант состоит в усреднении всех девяти указанных выше подходов; ему отвечает статистика:

$$\hat{T}_9^{(a)} = \frac{1}{9}\left\{\hat{T}_1^{(a)} + \hat{T}_1^{(b)} + \sum_{j=2}^{8}\hat{T}_j\right\}; \tag{51}$$

второй вариант состоит в усреднении лишь тех подходов, которые основаны на использовании эмпирической копула-функции, а именно подходов $\mathcal{A}_2$, $\mathcal{A}_3$ и $\mathcal{A}_4$; статистика для такого усреднения равна:

$$\hat{T}_9^{(b)} = \frac{1}{3}\left(\hat{T}_2 + \hat{T}_3 + \hat{T}_4\right). \tag{52}$$

Процедура 9.9. ***Реализация теста в подходе*** $\mathcal{A}_9$

1. Получить псевдо-наблюдения $(\hat{\mathbf{u}}_1,...,\hat{\mathbf{u}}_T)$ путем преобразования выборочных данных $(\mathbf{x}_1,...,\mathbf{x}_T)$ в нормированные ранги с использованием (32).

2. Оценить параметры α копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha} = \arg\max_{\alpha} l(\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_T;\alpha)$, т. е. с использованием полупараметрического метода.

3. Используя $(\hat{\mathbf{u}}_1,...,\hat{\mathbf{u}}_T)$ и $\hat{\alpha}$, реализовать необходимые шаги для вычисления критических статистик процедур 9.1–9.8 и вычислить $\hat{T}_1^{(a)}$, $\hat{T}_1^{(b)}$, $\hat{T}_2 - \hat{T}_8$.

4. В соответствии с (51)–(52) вычислить $\hat{T}_9^{(a)}$ и $\hat{T}_9^{(b)}$ соответственно.

5. При некотором достаточно большом натуральном K для $k = 1,...,K$ последовательно повторить следующие шаги:

а) в условиях справедливости нулевой гипотезы сгенерировать случайную выборку $(\mathbf{x}_{1,k}^0,...,\mathbf{x}_{T,k}^0)$ и, используя (32), вычислить ассоциированную с ней псевдовыборку $(\hat{\mathbf{u}}_{1,k}^0,...,\hat{\mathbf{u}}_{T,k}^0)$;

б) оценить параметры α^0 копула-функции с использованием состоятельной ММП-оценки $\hat{\alpha}_k^0 = \arg\max_{\alpha^0} l(\hat{\mathbf{u}}_{1,k}^0,\ldots,\hat{\mathbf{u}}_{T,k}^0;\alpha^0)$, т. е. с использованием полупараметрического метода;

в) используя $(\hat{\mathbf{u}}_{1,k}^0,...,\hat{\mathbf{u}}_{T,k}^0)$ и $\hat{\alpha}_k^0$, реализовать необходимые шаги для вычисления критических статистик из процедур 9.1 – 9.8 и вычислить $\hat{T}_{1,k}^{0,(a)}$, $\hat{T}_{1,k}^{0,(b)}$, $\hat{T}_{2,k}^0 - \hat{T}_{8,k}^0$;

г) в соответствии с (51)–(52), используя $\hat{T}_{1,k}^{0,(a)}$, $\hat{T}_{1,k}^{0,(b)}$, $\hat{T}_{2,k}^0 - \hat{T}_{8,k}^0$, вычислить $\hat{T}_{9,k}^{0,(a)}$ и $\hat{T}_{9,k}^{0,(b)}$.

Вычислить оценку p-значений по формулам:

$$\hat{p}^{(a)} = \frac{1}{K+1}\sum_{k=1}^{K}\mathbf{1}\{\hat{T}_{9,k}^{0,(a)} > \hat{T}_9^{(a)}\}, \quad \hat{p}^{(b)} = \frac{1}{K+1}\sum_{k=1}^{K}\mathbf{1}\{\hat{T}_{9,k}^{0,(b)} > \hat{T}_9^{(b)}\}.$$

Сравнение подходов с помощью имитационного моделирования. В работе (Berg, 2009) проведено объемное имитационное моделирование с целью сравнения описанных выше подходов к построению критериев согласия для копула-функций. И хотя показано, что не существует критерия, превосходящего по мощности любой другой из рассматриваемых подходов, подходы $\mathcal{A}_2$, $\mathcal{A}_4$ и $\mathcal{A}_9$ дают очень хорошие результаты, причем последний из них можно признать лучшим.

Интересно, что при тестировании гипотезы гауссовости (при альтернативе распределения с «тяжелыми хвостами») слабый в других ситуациях подход $\mathcal{A}_1$ дает очень хорошие результаты для высоких размерностей и больших объемов выборок. Следовательно, его целесообразно использовать в качестве предварительного теста проверки эллипсоидальности распределения с тем, чтобы решить, какой из подходов выбрать для дальнейшего исследования, см. (Huffera, Park, 2007).

9.5. Эмпирические приложения с пакетом *R*: пример реализации подхода $\mathcal{A}_2$

Модули пакета *R* дают возможность численно реализовать критерии согласия, основанные на сравнении *эмпирической* копула-функции с ее *параметрической оценкой* (в предположении справедливости проверяемой гипотезы), т. е. практически осуществить подход $\mathcal{A}_2$. Критическая статистика строится с использованием расстояния Крамера–фон Мизеса. Приближения для p-значений критической статистики вычисляются с использованием параметрических бутстрап-процедур, предложенных в (Genest, Remillard, 2008; Genest et al., 2009), или с помощью подхода, изложенного в (Kojadinovic et al., 2011; Kojadinovic, Yan, 2011). Ниже приведен код процедуры в соответствующем модуле пакета *R*.

```
library(copula)
x <- rcopula(claytonCopula(3), 200)
## Does the Gumbel family seem to be a good choice?
gofCopula(gumbelCopula(l), x)
## What about the Clayton family?
gofCopula(claytonCopula(1) , x)
## The same with a different estimation method
gofCopula(gumbelCopula(l), x, method="itau")
gofCopula(claytonCopula(1), x, method="itau")

## A three-dimensional example
x <- rcopula(tCopula(c(0.5, 0.6, 0.7), dim = 3, dispstr = "un"),200)
## Does the Clayton family seem to be a good choice?
gofCopula(gumbelCopula(l, dim =3), x)
## What about the t copula?
t.copula <- tCopula(rep(0, 3), dim = 3, dispstr = "un", df.fixed=TRUE)
gofCopula(t.copula, x)
## The same with a different estimation method
gofCopula(gumbelCopula(l, dim =3), x, method="itau")
gofCopula(t.copula, x, method="itau")
## The same using the multiplier approach
gofCopula(gumbelCopula(l, dim =3), x, simulation="mult")
gofCopula(t.copula, x, simulation="mult")

For sake of space, we report only the output relative to the first two goodness-of-
fit tests:
>     gofCopula(gumbelCopula(l), x)
Progress will be displayed every 100 iterations.
Iteration 100
…
Iteration 1000
$statistic [1] 0.2648942
$pvalue [1] 0.0004995005
$parameters [1] 1.980118
>     gofCopula(claytonCopula(1), x)
Progress will be displayed every 100 iterations.
Iteration 100
…
Iteration 1000
$statistic [1] 0.01489807
$pvalue [1] 0.54995
$parameters [1] 3.317709
```

Отметим, что в пакете, созданном Даниэлем Бергом, рассмотрен более широкий диапазон критериев согласия для копула-функций. Однако к моменту написания статьи (июль 2011 г.) существовала только предварительная версия этого пакета.

Список литературы

Фантаццини Д. (2011a). Моделирование многомерных распределений с использованием копула-функций. I. *Прикладная эконометрика*, 22 (2), 98–134.

Фантаццини Д. (2011b). Моделирование многомерных распределений с использованием копула-функций. II. *Прикладная эконометрика*, 23 (3), 98–132.

Anderson T. W., Darling D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49, 765–769.

Berg D. (2009). Copula goodness-of-fit testing: An overview and power comparison. *The European Journal of Finance*, 15, 675–701.

Berg D., Bakken H. (2007). A goodness-of-fit test for copulas based on the probability integral transform. *Technical report*, University of Oslo, Statistical research report no. 10.

Breymann W., Dias A., Embrechts P. (2003). Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance*, 3, 1–14.

Chen X., Fan Y. (2006). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of Econometrics*, 135, 125–154.

Chen X., Fan Y., Patton A. (2004). Simple tests for models of dependence between multiple financial time series, with applications to U. S. equity returns and exchange rates. London Economics Financial Markets Group, *Working Paper n. 483*.

D'Agostino R. B., Stephens M. A. (1986). *Goodness-of-fit techniques*. New York: Marcel Dekker Inc.

Danielsson J., de Vries C. G. (2000). Value-at-risk and extreme returns. *Annales d'Economie et de Statistique*, 60, 239–270.

David H. A. (1981). *Order statistics*. 2 ed., Wiley.

Deheuvels P. (1979). La fonction de dépendance empirique et ses proprietés: Un test non paramétrique d'indépendence. *Bulletin de l'Académie Royale de Belgique, Classe des Sciences*, 5e série, 65, 274–292.

Deheuvels P. (1984). The characterization of distributions by order statistics and record values: A unified approach. *Annals of Eugenics*, 21 (2), 326–334.

Denker M., Keller G. (1983). On U-statistics and v. Mises' statistics for weakly dependent processes. *Probability Theory and Related Fields*, 64 (4), 505–522.

Dias A., Embrechts P. (2004). Change-point analysis for dependence structures in finance and insurance. In: *Risk Measures for the 21st Century*, ed. by Giorgio Szegoe, 321–335. Wiley Finance Series.

Dobrić J., Schmid F. (2005). Testing goodness of fit for parametric families of copulas — application to financial data. *Communications in Statistics: Simulation and Computation*, 34 (4), 1053–1068.

Dobric J., Schmid F. (2007). A goodness of fit test for copulas based on Rosenblatt's transformation. *Computational Statistics and Data Analysis*, 51 (9), 4633–4642.

Fang K. T. Kotz S., Ng K. (1990). Symmetric multivariate and related distributions. In: *Monographs on Statistics and Applied Probability*, vol. 36. Chapman and Hall: London.

Fermanian J. (2005). Goodness of fit tests for copulas. *Journal of Multivariate Analysis*, 95, 119–152.

Genest C., Rivest L. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*, 88, 1034–1043.

Genest C., Rémillard B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 44 (6), 1096–1127.

Genest C., Quessy J. F., Rémillard B. (2006a). Goodness-of-fit procedures for copula models based on the probability integral transform. *Scandinavian Journal of Statistics*, 33, 337–366.

Genest C., Quessy J. F., Rémillard B. (2006b). On the joint asymptotic behavior of two rank-based estimators of the association parameter in the gamma frailty model. *Statistics and Probability Letters*, 76, 10–18.

Genest C., Rémillard B., Beaudoin D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44 (2), 199–213.

Glen A. G., Leemis L. M., Barr D. R. (2001). Order statistics in goodness-of-fit testing. *IEEE Transactions on Reliability*, 50 (2), 209–213.

Hong Y., Li H. (2005). Nonparametric specification testing for continuous-time models with application to spot interest rates. *Review of Financial Studies*, 18, 37–84.

Huard D., Ivin G., Favre A. C. (2006). Bayesian copula selection. *Computational Statistics and Data Analysis,* 51 (2), 809–822.

Huffera F. W., Park C. (2007). A test for elliptical symmetry. *Journal of Multivariate Analysis*, 98 (2), 256–281.

Jaynes E. T., Bretthorst G. L. (2003). *Probability theory: The logic of science*. Cambridge University Press.

Kojadinovic I., Yan J. M., Holmes M. (2011). Fast large-sample goodness-of-fit tests for copulas. *Statistica Sinica,* 21 (2), 841–871.

Kojadinovic I., Yan J. (2011). A goodness-of-fit test for multivariate multiparameter copulas based on multiplier central limit theorems. *Statistics and Computing*, 21 (1), 17–30.

Kole E., Koedijk K., Verbeek M. (2007). Selecting copulas for risk management. *Journal of Banking and Finance*, 31, 2405–2423.

Malevergne Y., Sornette D. (2003). Testing the gaussian copula hypothesis for financial assets dependence. *Quantitative Finance*, 3, 231–250.

Marsaglia G., Marsaglia J. (2004). Evaluating the Anderson–Darling distribution. *Journal of Statistical Software*, 9 (2), 1–5.

Panchenko V. (2005). Goodness-of-fit test for copulas. *Physica A*, 355 (1), 176–182.

Quessy J. F., Mesfioui M., Toupin M. H. (2007). A goodness-of-fit test based on Spearmans dependence function. *Working paper*, Université du Québec à Trois-Rivières.

Rosenblatt M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23, 470–472.

Savu C., Trede M. (2008). Goodness-of-fit tests for parametric families of Archimedean copulas. *Quantitative Finance*, 8 (2), 109–116.

Scaillet O. (2005). Kernel based goodness-of-fit tests for copulas with fixed smoothing parameters. *FAME Research Paper Series* n. 145, International Center for Financial Asset Management and Engineering.

Shih J. H. (1998). A goodness-of-fit test for association in a bivariate survival model. *Biometrika*, 85, 189–200.

Takeuchi K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Sciences*, 153, 12–18.

Wang W., Wells M. T. (2000). Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association*, 95, 62–72.

White H. (2000). A reality check for data snooping. *Econometrica*, 68, 1097–1126.

Appendix 9 "Everything you always wanted to know about log-periodic power laws for bubble modeling but were afraid to ask."

Geraskin, P., & Fantazzini, D. (2013). Everything you always wanted to know about log-periodic power laws for bubble modeling but were afraid to ask. *The European Journal of Finance*, 19(5), 366-391. DOI: https://doi.org/10.1080/1351847X.2011.601657

Permisson to post the Publisher's version/PDF: according to https://authorservices.taylorandfrancis.com/copyright-and-you/, it cannot be posted online, even in the case when it is included into a dissertation (an email exchange confirmed this limitation).
⇒ Instead, the authors' accepted manuscript must be used for online dissemination.

Everything You Always Wanted to Know about Log Periodic Power Laws for Bubble Modelling but Were Afraid to Ask

Petr Geraskin* Dean Fantazzini†

Abstract

Sornette et al. (1996), Sornette and Johansen (1997), Johansen et al. (2000) and Sornette (2003a) proposed that, prior to crashes, the mean function of a stock index price time series is characterized by a power law decorated with log-periodic oscillations, leading to a critical point that describes the beginning of the market crash. This paper reviews the original Log-Periodic Power Law (LPPL) model for financial bubble modelling, and discusses early criticism and recent generalizations proposed to answer these remarks. We show how to fit these models with alternative methodologies, together with diagnostic tests and graphical tools to diagnose financial bubbles in the making in real time. An application of this methodology to the Gold bubble which busted in December 2009 is then presented.

Keywords: Log-periodic models, LPPL, Crash, Bubble, Anti-Bubble, GARCH, Forecasting, Gold.

JEL classification: C32, C51, C53, G17.

PRE - PRINT VERSION OF THE FOLLOWING PUBLICATION:

Geraskin, P., Fantazzini, D. (2013). Everything you always wanted to know about log-periodic power laws for bubble modeling but were afraid to ask. *The European Journal of Finance*, 19(5), 366-391

The published version is available at https://doi.org/10.1080/1351847X.2011.601657

*Higher School of Economics, Moscow, Russia. E-mail: petr-geraskin@mail.ru

†Moscow School of Economics, Moscow State University, Russia. E-mail: fantazzini@mse-msu.ru

1 Introduction

Detecting a financial bubble and predicting when it will end has become of crucial importance, given the series of financial bubbles that led to the current "Second Great Contraction", using the definition by Reinhart and Rogoff (2009). As noted by Sornette (2009), Sornette and Woodard (2010), Kaizoji and Sornette (2010), Sornette et al. (2009) and Fantazzini (2010a,b), the global financial crisis that has started in 2007 can be considered an example of how the bursting of a bubble can be dealt with by creating new bubbles. This consideration which is not new in the financial literature, see e.g. Sornette and Woodard (2010) and references therein, was indirectly confirmed by Lou Jiwei, the chairman of the $298 billion sovereign wealth fund named China Investment Corporation (CIC), which was created in 2007 with the goal to manage an important part of the People's Republic of China's foreign exchange reserves. On August the 28th 2009, Lou told reporters on the sidelines of a forum organized by the Washington-based Brookings Institution and the Chinese 'Economists 50 Forum', a Beijing think-tank, that "*both China and America are addressing bubbles by creating more bubbles and we're just taking advantage of that. So we can't lose*". Moreover, Lou also added that "*CIC was building a broad investment portfolio that includes products designed to generate both alpha and beta; to hedge against both inflation and deflation; and to provide guaranteed returns in the event of a new crisis*". See the full Reuters article by Zhou Xin and Alan Wheatley at http://www.reuters.com/article/ousiv/idUSTRE57S0D420090829 for more details. The previous comments clearly point out how important is to have tools able to detect bubbles in the making.

Unfortunately, there is no consensus in the economic literature on what a bubble is: Gürkaynak (2008) surveyed a large set of econometric tests of asset price bubbles and found that for each paper that finds evidence of bubbles, there is another one that fits the data equally well without allowing for a bubble, so that it is not possible to distinguish bubbles from time-varying fundamentals. A similar situation can also be found in the professional literature: for example, Alan Greenspan stated on August the 30th 2002 that " *...We, at the Federal Reserve ... recognized that, despite our suspicions, it was very difficult to definitively identify a bubble until after the fact, that is, when its bursting*

confirmed its existence". So, is this a lost cause? Absolutely not.

A model which has quickly gained a lot of attention among financial practitioners and in the Physics academic literature due to the many successful predictions, is the so called *Log Periodic Power Law* (LPPL) approach proposed by Johansen et al. (2000) and Sornette (2003a,b). The Johansen-Ledoit-Sornette (JLS) model assumes the presence of two types of agents in the market: a group of traders with rational expectations and a second group of so called "noise" traders, that is irrational agents with herding behavior. The idea of the JLS model comes from statistical physics and it shares many elements with a model introduced by Ising for explaining ferromagnetism, see e.g. Goldenfeld (1992). According to this model, traders are organized into networks and can have only two states: buy or sell. In addition, their trading actions depend on the decisions of other traders and on external influences. Due to these interactions, agents can form groups with self-similar behavior which can lead the market to a bubble situation, which can be considered a situation of "order", compared to the "disorder" of normal market conditions. Another important feature introduced in this model is the positive feedbacks which are generated by the increasing risk and the agents' interactions, so that a bubble can be a self-sustained process.

Many examples of calibrations of financial bubbles with LPPLs are reported in Sornette (2003a), who suggests that the LPPL model provides a good starting point to detect bubbles and forecast their most probable end. Johansen and Sornette (2004) identified the most extreme cumulative losses (i.e. drawdowns) in a large set of financial assets and showed that they belong to a probability density distribution, which is distinct from the distribution of the 99% of the smaller drawdowns which represent the normal market regime. Moreover, they showed that, for two-third of these extreme drawdowns, the market prices followed a super-exponential behavior prior to their occurrences, as confirmed by a calibration of a LPPL model. These particular drawdowns (or outliers) are called "dragon kings" in Sornette (2009). Interestingly, this approach allowed to diagnose bubbles ex-ante, as shown in a series of real-life tests, see Sornette and Zhou (2006), Sornette, Woodard and Zhou (2008) and Zhou and Sornette (2003, 2006, 2008, 2009). Furthermore, it is currently being used at the *Financial Crisis Observatory* (FCO), which is a scientific platform set

up at the ETH - Zurich, aimed at "*testing and quantifying rigorously the hypothesis that financial markets exhibit a degree of inefficiency and a potential for predictability, especially during regimes when bubbles develop*", see `http://www.er.ethz.ch/fco` for more details. The goal of this paper is to present an easy-to-use and self-contained guide for bubble modelling and detecting with Log Periodic Power Laws, which contains all the sufficient steps to derive the main models in this growing and interesting field of the literature, and discusses the important aspects for practitioners and researchers.

The rest of the paper is organized as follows. Section 2 reviews the original JLS model with the main steps required for its derivation. Section 3 discusses the early criticism to this approach and recent generalizations proposed to answer these remarks. Section 4 discusses how to fit LPPLs models, by presenting three estimation methodologies: the original 2-step nonlinear optimization by Johansen et al. (2000), the Genetic Algorithm approach proposed by Jacobsson (2009) and the 2-step/3-step ML approach proposed by Fantazzini (2010a). Section 5 is devoted to the diagnosis of bubbles in the making by using a set of different techniques. We describe diagnostic tests based on the LPPL fitting residuals, diagnostic tests based on rational expectation models with stochastic mean-reverting termination times, as well as graphical tools useful for capturing bubble development and for understanding whether a crash is in sight or not. Section 6 presents a detailed empirical application devoted to the burst of the gold bubble in December 2009, while Section 7 briefly concludes.

2 The Original LPPL model

Johansen et al. (JLS, 2000) consider an ideal market with no dividends, and where interest rates, risk aversion and market liquidity constraints are ignored. Therefore, the fundamental value for an asset is $p(t) = 0$, so any positive value of $p(t)$ represents a bubble. In general, $p(t)$ can be viewed as the price in excess of the fundamental value of an asset. In this framework, there are two types of agents: first, a group of rational agents who are identical in their preferences and characteristics, so they can be substituted with a single representative agent. Second, a group of irrational agents whose herding behavior leads to the development of a financial bubble. When this tendency develops till a certain

critical value, a large proportion of agents will then assume the same short position, thus causing a crash. A financial crash is not a certain event in this model, but it is characterized by a probability distribution: as a consequence, it is rational for financial agents to continue investing, because the risk of the crash to happen is compensated by the positive return generated by the financial bubble and there exists a small probability for the bubble to disappear smoothly, without the occurrence of a crash.

The key variable to model the price behavior before a crash is the crash hazard rate $h(t)$, that is the probability per unit of time that the crash will take place, given that it has not yet occurred. The hazard rate $h(t)$ quantifies the probability that a great number of agents will assume the same sell position simultaneously, a position that the market will not be able to satisfy unless the prices decrease substantially. We remark that in this model a strong collective answer (as it is the case for a crash) is not necessarily the consequence of one elaborated internal mechanism of global coordination, but it can appear starting from imitative local micro-interactions which are then transmitted by the market resulting in a macroscopic effect. In this regard, JLS (2000) first discuss a macroscopic "mean field" approach and then turn to a more microscopic approach.

2.1 Macroscopic Modelling

According to the mean field theory from Statistical Mechanics (see e.g. Stanley (1971) and Goldenfeld, (1992)), a simple way for describing an imitative process is by assuming that the hazard rate $h(t)$ can be described by the following equation:

$$\frac{dh}{dt} = Ch^{\delta} \tag{1}$$

where $C > 0$ is a constant, and $\delta > 1$ represents the average number of interactions among traders minus one. Thus, it follows that an amplification of interactions increases the hazard rate. If we integrate (1), we have:

$$h(t) = \left(\frac{h_0}{t_c - t}\right)^{\alpha}, \quad \alpha = \frac{1}{\delta - 1} \tag{2}$$

where t_c is the critical time determined by the initial conditions at some origin of time. It can be shown that the condition $\delta > 1$ (and consequently $\alpha > 0$) is crucial to obtain a growth of $h(t)$ as $t \to t_c$ and therefore a critical point in finite time. Moreover, the condition

that $\alpha < 1$ is required for the price not to diverge at t_c. Rewriting these condition for δ we have that $2 < \delta < \infty$, that is, an agent should be connected at least with two agents. Another important feature of this approach is the possibility of self-fulfilling crisis, which is a concept recently proposed to explain the recession in the '90 in seven countries (Argentina, Indonesia, Hong Kong, Malaysia, Mexico, South Korea and Thailand), see Krugman (1998) and Sornette (2003a). It is suggested that the loss of investor's confidence caused a self-fulfilling process in these countries and thus led to severe recessions. This feedback process can be modelled by using the previous mean field approach:

$$\frac{dh}{dt} = Dp^{\mu}, \quad \mu > 0 \tag{3}$$

where D is a positive constant. The underlying idea is that the lack of confidence quantified by the hazard rate increases when the market price departs from its fundamental value. Therefore, the price has to increase to compensate the increasing risk.

2.2 Microscopic Modelling

JLS (2000) and Sornette (2003a) assume that the group of irrational agents are connected into a network. Each agent is indexed by a integer number $i = 1, \ldots, I$ and $N(i)$ represents the number of agents who are directly connected to agent i in the network. JLS (2000) assume that each agent can have only two possible states s_i: "buy" ($s_i = +1$) or "sell" ($s_i = -1$). JLS (2000) suppose that the state of agent i is determined by the following Markov process:

$$s_i = \text{sign}\left(K \sum_{k \in N(i)} s_j + \sigma \varepsilon_i\right) \tag{4}$$

where the sign function $sign(x)$ is equal to $+1$ if $x > 0$ and to -1 if $x < 0$, K is a positive constant, ε_i is an i.i.d. standard normal random variable. In this model, K governs the tendency of imitation among traders, while σ governs their idiosyncratic behavior. If K increases, the order in the network increases as well, while the reverse is true when σ increases. If order wins, the agents will imitate their close neighbors and their imitation will spread all over the network, thus causing a crash[1]. More specifically and in analogy

[1] In the context of the alignment of atomic spins to create magnetization, this model represented by (4) is identical to the so-called two-dimensional *Ising model* which was solved explicitly by Onsager (1944), and where the disorder parameter is represented by the temperature of the system.

with the Ising model, there exists a critical point K_c, that determines the separation between the different regimes: when $K < K_c$ the disorder reigns and the sensibility to a small global influence is low. When the imitation force K grows approaching K_c, a hierarchy of groups of agents acting collectively and with the same position is formed. As a consequence, the market becomes extremely sensitive to small global disturbances. Finally, for a larger imitation force so that $K > K_c$, the tendency of imitation is so intense that there exists a strong predominance of one state/position among agents.

A physical quantity that represent the degree of a system sensitivity to an external perturbation (or general global influence) is the so-called *susceptibility* of the system. This quantity describes the probability that a large group of agents will have the same state, given the existent external influences in the network. Let us assume the existence of a term G which measures the global influence, and add it to (4):

$$s_i = \text{sign}\left(K \sum_{k \in N(i)} s_j + \sigma\varepsilon_i + G\right) \tag{5}$$

If we define the average state of the market as $M = (1/I)\sum_{i=1}^{I} s_i$, for $G = 0$ we have $E[M] = 0$ by symmetry. For $G > 0$, we have $M > 0$, while for $G < 0$, $M < 0$. Thus, it follows that $E[M] \times G \geq 0$. The susceptibility of the system is then defined as $\chi = \left.\frac{dE[M]}{dG}\right|_{G=0}$. In general, the susceptibility has three possible interpretations: first, it measures the sensitivity of M to a small change in the global influence. Secondly, it is (a constant times) the variance of M around its zero expectation, caused by idiosyncratic shocks ε_i. Finally, if we consider two agents and we force one to be in a certain state, the impact that our intervention will have on the second agent will be proportional to the susceptibility.

2.3 Price Dynamics and Derivation of the JLS Model

As previously anticipated, the rational agent considered by JLS (2000) is risk neutral and has rational expectations. Thus, the asset price $p(t)$ follows a martingale process, i.e. $E_t[p(t')] = p(t)$, $\forall t' > t$, where $E_t[.]$ represents the conditional expectation given all information available up to time t. In the case of market equilibrium, the previous equality is a necessary condition for no arbitrage.

Considering that there exists a not zero probability for the crash to happen, we can define

a jump process j which is equal to zero before crash and one one after the occurrence of the crash at time t_c. Since t_c is unknown, it is described by a stochastic variable with a probability density function $q(t)$, a cumulative distribution function $Q(t)$ and an hazard rate given by $h(t) = q(t)/[1 - Q(t)]$, which is the probability per unit of time of the crash taking place in the next instant, given that it has not yet occurred. Assuming for simplicity that the price falls during a crash by a fixed percentage $k \in (0, 1)$, the asset price dynamics is given by:

$$\begin{aligned} dp &= \mu(t)p(t)dt - kp(t)dj \qquad \Rightarrow \\ E[dp] &= \mu(t)p(t)dt - kp(t)[P(dj = 0) \times (dj = 0) + P(dj = 1) \times (dj = 1)] = \\ &= \mu(t)p(t)dt - kp(t)[0 + h(t)dt] = \mu(t)p(t)dt - kp(t)h(t)dt \end{aligned} \tag{6}$$

The no arbitrage condition and rational expectations together imply that $E[dp] = 0$, so that $\mu(t)p(t)dt - kp(t)h(t)dt = 0$, which yields $\mu(t) = kh(t)$. Substituting this last equality into (6), we obtain the differential equation defining the price dynamics before the occurrence of the crash given by $d(\ln p(t)) = kh(t)$, whose solution is

$$\ln\left[\frac{p(t)}{p(t_0)}\right] = \kappa \int_{t_0}^{t} h(t')dt' \tag{7}$$

The idea is that the higher the probability of the crash is, the faster the price should grow to compensate investors for the increased risk of a crash in the market, see also Blanchard (1979). At this point, JLS (2000) employ the result that a system of variables close to a critical point can be described by a *power law* and the susceptibility of the system diverges as follows:

$$\chi \approx A(K_c - K)^{-\gamma} \tag{8}$$

where A is a positive constant and $\gamma > 0$ is called the critical exponent of the susceptibility (equal to 7/4 for the 2-dimensional Ising model). Unfortunately, the bi-dimensional Ising model considers only investors interconnected in an uniform way, while in real markets some agents can be more connected than others. Modern financial markets are constituted by a collection of interacting investors, that differ substantially in size, going from the individual investors until the large pension funds. Furthermore, all investors in the world are organized inside a network (family, friends, work, etc), within which they locally influence each other. A more appropriate representation for the current structure of financial markets is given by a *hierarchical diamond lattice*, which is used by JLS (2000) to develop a

model of rational imitation. This structure can be described as follows: first, consider two agents linked to each other, so that we have one link and two agents. Secondly, substitute this link with four new links forming a diamond: the two original agents are now situated in the two diametrically opposite vertices, whereas the two other vertices are occupied by two new traders. Thirdly, for each one of these 4 links, substitute them with 4 new links, forming a diamond in the same way. If we repeat this operation an arbitrary number of times, we will get a Hierarchical Diamond Lattice. As a result, after n iterations there will be $N = (2/3) * (2 + 4^n)$ agents and $L = 4^n$ links among them. For example, the last generated agents will have only two links, the initial agents will have 2^n neighbors, while the others will have an intermediate number of neighbors in between. A version of this model was solved by Derrida et al. (1983). The basic properties are similar to those of the rational imitation model using the bi-dimensional network. The only crucial difference is that the critical exponent γ of the susceptibility in (8) can be a complex number. Therefore, the general solution is given by:

$$\begin{aligned} \chi &\approx \mathrm{Re}[A_0(K_c - K)^{-\gamma} + A_1(K_c - K)^{-\gamma + i\omega} + \ldots] \\ &\approx A_0'(K_c - K)^{-\gamma} + A_1'(K_c - K)^{-\gamma} \cos[\omega \ln(K_c - K) + \psi] + \ldots \end{aligned} \tag{9}$$

where A_0, A_1, ω are real numbers and Re[.] represents the real part of a complex number. The power law in (9) is now corrected by oscillations called "log-periodic", because they are periodic in the logarithm of the variable $(K_c - K)$, and $\omega/2$ is their log-frequency. These oscillations are accelerating since their frequency explodes as it reaches the critical time. Considering this mechanism, JLS (2000) assume that the crash hazard rate behave in a similar way to the susceptibility in the neighborhood of a critical point. Therefore, using (9) and considering a hierarchical lattice for the financial market, the hazard rate has the following behavior:

$$h(t) \approx B_0(t_c - t)^{-\alpha} + B_1(t_c - t)^{-\alpha} \cos[\omega \ln(t_c - t) + \psi'] \tag{10}$$

This behavior of the hazard rate shows that the risk of a crash per unit of time, given that it has not yet occurred, increases drastically when the interactions among investors become sufficiently strong. However, this acceleration is interrupted and superimposed with an accelerating sequence of phases where the risk decreases, which is represented by the log-periodic oscillations. Applying (10) to (7), we get the following evolution for the

asset price before a crash:

$$\ln[p(t)] \approx \ln[p(c)] - \frac{\kappa}{\beta}\left\{B_0(t_c - t)^\beta + B_1(t_c - t)^\beta \cos[\omega \ln(t_c - t) + \phi]\right\} \tag{11}$$

which can be rewritten in a more suitable form for fitting a financial time series as follows:

$$\ln[p(t)] \approx A + B(t_c - t)^\beta \left\{1 + C\cos[\omega \ln(t_c - t) + \phi]\right\} \tag{12}$$

where $A > 0$ is the value of $[\ln p(t_c)]$ at the critical time, $B < 0$ the increase in $[\ln p(t)]$ over the time unit before the crash if C were to be close to zero, $C \neq 0$ is the proportional magnitude of the oscillations around the exponential growth, $0 < \beta < 1$ should be positive to ensure a finite price at the critical time t_c of the bubble and quantifies the power law acceleration of prices, ω is the frequency of the oscillations during the bubble, while $0 < \phi < 2\pi$ is a phase parameter. Expression (12), which is known as the Log Periodic Power Law (LPPL), is the fundamental equation that describes the temporal growth of prices before a crash and it has been proposed in different forms in various papers, see e.g. Sornette (2003a) and Lin et al. (2009) and references therein. We remark that A, B, C and ϕ, are just units distributions of betas and omegas, as described in Sornette and Johansen (2001) and Johansen (2003), and do not carry any structural information.

3 Criticism and Recent Generalizations

3.1 Criticism

The most important and detailed criticism against the LPPL approach was put forward by Chang and Feigenbaum (2006), who tested the mechanism underlying the LPPL by using Bayesian methods applied to the time series of returns (see also Laloux et al. (1999) for additional criticism and the reply by Johansen (2002)). By comparing marginal likelihoods, they showed that a null hypothesis model without log-periodical structure outperforms the JLS model. And if the JLS model was true, they found that parameter estimates obtained by curve fitting have small posterior probability. As a consequence, they suggested to abandon the class of models in which the LPPL structure is revealed through the expected return trajectory. These problems are due to the fact that the JLS model considers a deterministic time-varying drift decorated by a non-stationary stochastic random walk component: this latter component has a variance which increases over time, so that the

deterministic trajectory moves away from the observable price path and model estimation with prices is no more consistent. Therefore, Chang and Feigenbaum (2006) considered the time series of returns instead of prices, and resorted to Bayesian methods to simplify the analysis of a complicated time-series model like the JLS model, see Bernardo and Smith (1994) or Koop (2003) for an introduction to Bayesian theory. The benchmark model in Chang and Feigenbaum (2006) is represented by the Black-Scholes model, whose logarithmic returns are given by

$$r_i \sim N(\mu(t_i - t_{i-1}), \sigma^2(t_i - t_{i-1})) \tag{13}$$

where $r_i = q_i - q_{i-1}$ and q_i is the log of the price. The drift μ is drawn from the prior distribution $N(\mu_r, \sigma_r)$, while the variance σ^2 is specified in terms of its inverse $\tau = 1/\sigma^2$, known as the precision, which is higher the more precisely the random variable is known. The precision is drawn from the prior distribution $\tau \sim \Gamma(\alpha_\tau, \beta_\tau)$. The alternative hypothesis model by Chang and Feigenbaum (2006) is the LPPL model with a constant drift μ in the mean function (which was not included in the original JLS model):

$$\begin{aligned} r_i &\sim N(\mu(t_i - t_{i-1}) + \Delta H_{i,i-1}, \sigma^2(t_i - t_{i-1})), \quad \text{where} \\ \Delta H_{i,i-1} &= B(t_c - t_{i-1})^\beta \left[1 + \frac{C}{\sqrt{1 + \left(\frac{\omega}{\beta}\right)^2}} \cos(\omega \ln(t_c - t_{i-1}) + \phi) \right] - \\ &\quad -B(t_c - t_i)^\beta \left[1 + \frac{C}{\sqrt{1 + \left(\frac{\omega}{\beta}\right)^2}} \cos(\omega \ln(t_c - t_i) + \phi) \right] \end{aligned} \tag{14}$$

The LPPL model is characterized by the parameter vector $\xi = (A, B, C, \beta, \omega, \phi, t_c)$, and these parameters are drawn independently from the following prior distributions:

$$A \sim N(\mu_A, \sigma_A), \quad B \sim \Gamma(\alpha_B, \beta_B), \quad C \sim U(0,1), \quad \beta \sim B(\alpha_\beta, \beta_\beta)$$

$$\omega \sim \Gamma(\alpha_\omega, \beta_\omega), \quad \phi \sim U(0, 2\pi), \quad t_c - t_N \sim \Gamma(\alpha_{t_c}, \beta_{t_c})$$

where Γ, B and U denote the Gamma distribution, Beta distribution and uniform distribution, respectively. Given the independence among prior distributions, the prior density for this model is simply given by the product of all marginal priors, while the probability data density for q_i is

$$f(q_i|q_{i-1},\theta_{LPPL};LPPL) = \sqrt{\frac{\tau}{2\pi(t_i - t_{i-1})}} \exp\left[-\frac{\tau(q_i - q_{i-1} - \mu(t_i - t_{i-1}) - \Delta H_{i,i-1})^2}{2(t_i - t_{i-1})}\right]$$

so that the likelihood function for the observed data Q is given by

$$f(Q|\theta_{LPPL};LPPL) = \prod_{i=1}^{N} f(q_i|q_{i-1},\theta_{LPPL};LPPL)$$

Finally, the log marginal likelihood necessary for the computation of the Bayes factor is given by

$$\mathcal{L} = \ln\left(\int_{\Theta} f(Q|\theta_{LPPL};LPPL)\varphi(\theta_{LPPL};LPPL)d\theta_{LPPL}\right)$$

which can be computed with Monte-Carlo methods and a large number of sampling values. By using relatively diffuse priors with large variances in order to encompass the true values of the parameters, Chang and Feigenbaum (2006) found that the marginal likelihood remains basically the same, whether they consider the LPPL specification in the mean function, or only the drift term μ. This result remains robust to a change of prior distributions and they showed that the null hypothesis outperforms the JLS model in terms of marginals likelihood with different sets of priors.

Apart from the problem with weakly informative prior densities, see e.g. Bauwens et al. (2000) for a discussion, Lin et al. (2009) pointed out that the Bayes approach to hypothesis testing assumes that some kind of ergodicity on a single data sample applies, and that this sample has to be of sufficiently large size (which is not always the case). Clearly, this has to be tested and is far from being trivial. Furthermore, it is known that LPPL models can have likelihoods with several local maxima, see Jacobsson (2009) for a recent review, and the Bayes approach aims at solving this problem by integration, that is by smoothing. However, for small to medium sample sizes, the smoothing in the marginal likelihoods can be harmful, particularly in case of poor priors, and can decrease the number of local maxima at the price of a loss of efficiency. This may explain why the null hypothesis model with no log-periodic components showed a better result than the LPPL model.

3.2 The Generalized LPPL Model with Mean-Reversing Residuals

The work by Chang and Feigenbaum (2006) represented the most important challenge to the original JLS model, and this is why it prompted a response by Sornette and his

co-authors in 2009. Lin et al. (2009) proposed a generalization of the original model which wants to make the process consistent with direct price calibration. As we have shown in the previous sections, the original JLS model has a random walk component with increasing variance which makes direct estimation with prices inconsistent, as well as causing the lack of power of Bayesian methods, as shown by Lin et al. (2009). Instead, the "*volatility-confined LPPL model*" proposed by Lin et al. (2009) combines a mean reverting volatility process together with a stochastic conditional return which represents the continuous reassessments of investors' beliefs for future returns. As a consequence, the daily logarithmic returns are no longer described by a deterministic drift decorated by a Gaussian-distributed white noise, and the expected returns become stochastic.

Using the standard framework of rational expectations, Lin et al. (2009) assume that the price dynamics during a bubble is governed by the following process:

$$\begin{aligned} \frac{dI}{I} &= \mu(t)dt + \sigma_Y dY + \sigma_W dW - \kappa dj \\ dY &= -\alpha Y dt + dW \end{aligned}$$

where I is the stock price index or the price of a generic asset, W is the standard Wiener process, $\mu(t)$ is a time-varying drift characteristic of a bubble regime, j is equal to zero before the crash and one afterwards, while κ represents the percentage by which the asset price falls during a crash. When $0 < \alpha < 1$, Y denotes an Ornstein-Uhlenbeck process, so that dY and Y are both stationary, and the volatility remains bounded till the crash. This property guarantees that direct estimation with prices is consistent. We remark that if $\alpha = 0$, we retrieve the original JLS model. The corresponding model in discrete time is given by

$$\ln I_{i+1} - \ln I_i = \mu_i + \sigma_Y(Y_{i+1} - Y_i) + \sigma_W \varepsilon_i - \kappa \Delta j_i \tag{15}$$

$$Y_{i+1} = (1-\alpha)Y_i + \varepsilon_i, \quad \varepsilon_i \sim N(0,1) \tag{16}$$

Using the theory of the Stochastic Discount Factor (SDF), complete markets and no-arbitrage, Lin et al. (2009) show that the asset log returns follow this process,

$$\ln I_{i+1} = \ln I_i + \Delta H_{i,i-1} - \alpha(\ln I_i - H_i) + u_i \tag{17}$$

where $\Delta H_{i,i-1}$ is given by expression (14) and u_i is a Gaussian white noise, while the conditional probability distribution for the logarithmic returns is given by:

$$r_{i+1} = \ln I_{i+1} - \ln I_i \sim N(\Delta H_{i+1,i} - \alpha(\ln I_i - H_i), \sigma_u^2(t_{i+1} - t_i)) \tag{18}$$

Differently from the original JLS model, the additional term $-\alpha(\ln I_i - H_i)$ ensures that the log-price fluctuates around the LPPL trajectory H_t, thus guaranteeing the consistency of direct estimation with prices.

Lin et al. (2009) remarked that the previous model based on rational expectation separates rather artificially the noise traders and the rational investors. Moreover, even though the rational investors cannot make profit on average, rational agents endowed with different preferences may in principle arbitrage the risk-neutral agents. Therefore, assuming that rational investors have homogeneous preferences is rather restrictive. Nevertheless, Lin et al. (2009) show that the previous results can be obtained by using a complete different approach, which considers the theory of the so-called *Behavioral SDF*, where the price movements follows the dynamics of the *market sentiment*, see Shefrin (2005) for a textbook treatment of the behavioral approach to asset pricing. We refer to Lin et al. (2009) for more details about this alternative approach.

3.3 Other Generalizations: The Log-Periodic-AR(1)-GARCH(1,1) Model

While the original LPPL specification can model the long-range dynamics of price movements, nevertheless it is unable to consider the short-term market dynamics, thus showing residual terms which can be strongly autocorrelated and heteroskedastic. As a consequence, Gazola et al. (2008) proposed the following AR(1)-GARCH(1,1) log-periodic model:

$$\begin{aligned} I_i &= A + B(t_c - t_i)^\beta + C(t_c - t_i)^\beta \cos[w \ln(t_c - t_i) + \phi] + u_i \\ u_i &= \rho u_{i-1} + \eta_i \\ \eta_i &= \sigma_i \varepsilon_i, \quad \varepsilon_i \sim N(0,1) \\ \sigma_i^2 &= \alpha_0 + \alpha_1 \eta_{i-1}^2 + \alpha_2 \sigma_{i-1}^2 \end{aligned} \tag{19}$$

where ε_i is a standard white noise term satisfying $\mathrm{E}[\varepsilon_i] = 0$ and $\mathrm{E}[\varepsilon_i^2] = 1$, whereas the conditional variance σ_i^2 follows a GARCH(1,1) process. Under the normality assumption for the error term ε_i, the maximum likelihood estimator for the parameter vector $\mathbf{\Pi} = [A, B, C, t_c, \beta, w, \phi, \rho, \alpha_0, \alpha_1, \alpha_2]$ is obtained through the numerical maximization of the log likelihood:

$$\ln L(\Theta) = -\frac{1}{2}(N-1)\ln(2\pi) - \frac{1}{2}\sum_{i=2}^{N} \ln \sigma_i^2 - \frac{1}{2}\sum_{i=2}^{N} \frac{\eta_i^2}{\sigma_i^2} \tag{20}$$

In order to improve the optimization procedure, each parameter of the log-periodic model (19) denoted by θ and defined in a restricted interval denoted by $[a, b]$, can be re-parameterized according to the following monotonic transformation:

$$\theta = b\frac{\exp(\tilde{\theta})}{1+\exp(\tilde{\theta})} + a\left(1 - \frac{\exp(\tilde{\theta})}{1+\exp(\tilde{\theta})}\right) \tag{21}$$

This monotonic transformation turns the original estimation problem over a restricted space of solutions into an unrestricted problem, which eases estimation particularly when poor starting values are chosen. In this case, the delta method can be used to compute the standard errors of the estimate. We remind that the delta method is used to compute an estimator for the variance of functions of estimators and the corresponding confidence bands. Let $\hat{V}[\tilde{\theta}]$ be the estimated variance-covariance matrix of $\tilde{\theta}$ then, by using the delta method, a variance-covariance matrix for a general nonlinear transformation $g(\tilde{\theta})$ is given by (see Hayashi (2000) for more details):

$$\hat{V}[g(\tilde{\theta})] = \frac{\partial g(\tilde{\theta})}{\partial \tilde{\theta}'}\hat{V}[\tilde{\theta}]\frac{\partial g(\tilde{\theta})}{\partial \tilde{\theta}}'$$

Gazola et al. (2008) use a 2-step procedure to choose the starting values for the numerical maximization of (20):

1. the starting values for the set of parameters $\Phi = [A, B, C, t_c, \beta, w, \phi]$ are retrieved from the estimation of the original LPPL model (12);

2. the starting values for the set of parameters $[\rho, \alpha_0, \alpha_1, \alpha_2]$ of the short-term stochastic component u_i are obtained by estimating an AR(1)-GARCH(1,1) model on the residuals $\hat{u}_i$ from the original LPPL model (12).

4 How to fit LPPL models?

Estimating LPPL models in general has never been easy, due to the frequent presence of many local minima of the cost function where the minimization algorithm can get trapped. However, some recent developments have simplified considerably the estimation process.

4.1 The Original 2-step Nonlinear Optimization

Johansen et al. (2000) noted that noisy data, relatively small samples and a large number of parameters make the estimation of LPPL models rather difficult. Therefore, they proposed to reduce the number of free parameters by slaving the three linear parameters and computing them from the estimated nonlinear parameters.

More specifically, if we rewrite the original LPPL model as follows,

$$y_i = A + B(t_c - t_i)^\beta + C(t_c - t_i)^\beta \cos(\omega \ln(t_c - t_i) + \phi) \tag{22}$$

or more compactly as,

$$y_i = A + Bf_i + Cg_i$$

where

$$\begin{aligned} y_i &= \ln I_i \quad \text{or} \quad I_i \ , \qquad f_i = (t_c - t_i)^\beta \\ g_i &= (t_c - t_i)^\beta \cos(\omega \ln(t_c - t_i) + \phi) \end{aligned}$$

then, it is straightforward to see that the linear parameters A, B, and C can be obtained analytically by using ordinary least squares:

$$\begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} y_i f_i \\ \sum_{i=1}^{N} y_i g_i \end{pmatrix} = \begin{pmatrix} N & \sum_{i=1}^{N} f_i & \sum_{i=1}^{N} g_i \\ \sum_{i=1}^{N} f_i & \sum_{i=1}^{N} f_i^2 & \sum_{i=1}^{N} f_i g_i \\ \sum_{i=1}^{N} g_i & \sum_{i=1}^{N} f_i g_i & \sum_{i=1}^{N} g_i^2 \end{pmatrix} \begin{pmatrix} A \\ B \\ C \end{pmatrix} \tag{23}$$

We can write the previous system compactly by using matrix notation:

$$\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})\mathbf{b}, \quad \text{where} \quad \mathbf{X} = \begin{pmatrix} 1 & f_1 & g_1 \\ \vdots & \vdots & \vdots \\ 1 & f_N & g_N \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} A \\ B \\ C \end{pmatrix} \tag{24}$$

so that

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{y} \tag{25}$$

and we have only four free parameters to estimate (see also Jacobsson (2009) for a similar derivation). We remark that this simplification can also be seen as an example of concentrated maximum likelihood.

The estimation procedure consists of two steps:

1. Use the so-called *Taboo search* (Cvijović and Klinowski, 1995) to find 10 candidate solutions, where only the cases with $B < 0$, $0 < \beta < 1$ and $t_c > t_i$ (if a bubble) are considered, see also Sornette and Johansen (2001). However, alternative grid searches can also be considered. Recently, Lin et al. (2009) have imposed stronger restrictions, by considering $0.1 < \beta < 0.9$, $6 \leq \omega \leq 15$ so that the log-periodic oscillations are neither too fast (to avoid fitting noise) nor too slow (otherwise they would provide a contribution to the trend), and $|C| < 1$ to ensure that the hazard rate $h(t)$ remains always positive[2];

2. Each of these 10 solutions is then used as starting value in a Levenberg-Marquardt nonlinear least squares algorithm. The solution with the minimum sum of squares between the fitted model and the observations is taken as the final solution.

4.2 Genetic Algorithms

The Genetic Algorithm (GA) is an algorithm inspired by Darwin's "survival of the fittest idea", and its theory was developed by John Holland in 1975. The GA is a computer simulation that aims at mimicking the natural selection in biological systems, which is governed by four phases: a *selection mechanism*, a *breeding mechanism*, a *mutation mechanism*, and a *culling mechanism*. The GA does not require the computation of any gradient or curvature and it does not need the cost function to be smooth or continuous.

The use of GA to estimate LPPL models has been proposed by Jacobsson (2009) following the GA methodology by Gulsten et al. (1995). Similarly to Johansen et al. (2000), Jacobsson (2009) reduces the number of free parameters to four, by "slaving" the three linear parameters A, B and C, which are computed by using (25). Her procedure consists of four steps:

1. *Selection Mechanism: Generating the Initial Population.* Each member of the "financial" population is represented by a vector of the four nonlinear coefficients t_c, ϕ, ω, and

[2]The lower bound on ω is rather strong, given that Sornette (2003a) found that $\omega = 6.36 \pm 1.56$ by using a large collection of empirical evidence. Moreover, the recent work about Chinese market bubbles by Jiang et al. (2010) considers only the former set of conditions.

β. The members of the initial population are randomly drawn from a uniform distribution with a pre-specified range, and for each member, the residual sum of squares is calculated. Jacobsson (2009) considers an initial population of 50 members (that is, 50 parameter vectors).

2. *Breeding mechanism.* The 25 members with the best value of the cost function are selected from the population to be included in the breeding program. An offspring is then generated by randomly drawing two parents, without replacement, and taking the arithmetic mean of them. Jacobsson (2009) repeats this procedure 25 times and each pair of parents is drawn randomly with replacement, so that one parent can generate an offspring with another parent (...therefore, betrayals are allowed!).

3. *Mutation mechanism.* Genetic mutations in nature play a key role in the evolution of a species, since they may increase its probability of survival, as well as introduce less favorable characteristics. In our framework, mutations perturb the previous solutions to allow new regions of the search space to be explored, so that premature convergence in local minima can be avoided.

   The mutation process is implemented by computing the statistical range ($\theta_{\max} - \theta_{\min}$) for each parameter in the population. The range for each parameter is then multiplied with a factor $\pm k$, to obtain the perturbation variable ε, which is uniformly distributed over the interval $[-k \times (\text{parameter range}), k \times (\text{parameter range})]$. Jacobsson (2009) considers $k = 2$. 25 members are then drawn randomly, without replacement, from the initial population of 50 computed in the first step. Each selected member is then mutated by adding an exclusive vector of random perturbations for every parameter. Therefore, the mutation mechanism allows to compensate the problem of an inaccurate guess for the initial intervals in the solution space.

4. *Culling mechanism.* Jacobsson (2009) merges the members generated by mutation and breeding into the population, so that a total of 100 solutions is present (50 old, 25 offsprings, and 25 mutations). All of the 100 solutions are ranked according to their cost function in ascending order, and the 50 best solutions are culled, and live on into the next generation. The rest is deleted.

The previous algorithm is then iterated a certain number of times till a desired termination criterion is met. Similarly to the second step of the optimization process used by Johansen et al. (2000) and described in section 4.1, Jacobsson (2009) further refines the parameters estimated with the GA by using them as starting values for the Nelder-Mead Simplex method, also known as the downhill simplex method.

4.3 The 2-step/3-step ML Approach

Fantazzini (2010a) found that estimating LPPL models for "anti-bubbles" was much easier than estimating LPPL models for bubbles: an anti-bubble is symmetric to a bubble, and represents a situation when the market peaks at a critical time t_c and then decreases following a power law with decelerating log-periodic oscillations, see Johansen and Sornette (1999), Johansen and Sornette (2000), Zhou and Sornette (2005) and Fantazzini (2010b) for more details. Furthermore, estimating models with log-prices was much simpler than models with prices in levels, and in the latter case a much more careful choice of the starting values had to be made.

In this regard, we have already seen that the original LPPL model has a stochastic random walk component with increasing variance, so that the deterministic pattern moves away from the observable price path. Therefore, the idea by Fantazzini (2010a) is to reverse the original times series in order to minimize the effect of the non-stationary component during the estimation process. The same idea can be of help also in case of models with stationary error terms, like the one proposed by Lin et al. (2009): a time series with strongly autocorrelated error terms is almost undistinguishable from a non stationary process in small-to-medium sized samples, see e.g. Stock (1994), Ng and Perron (2001) and Stock and Watson (2002) for a discussion of this hot debated issue in the econometric literature. See, for instance, Figure 1 which shows a simulated LPPL model with AR(1) error terms and 1000 observations, as well as its reverse.

The 2-step ML approach used in Fantazzini (2010a) to estimate LPPL models for financial bubbles, also allowing for an AR(1)-GARCH(1,1) model in the error terms as in (19), is given below:

1. Reverse the original time series and estimate the LPPL for the case of an anti-bubble

by using the BFGS (Broyden, Fletcher, Goldfarb, Shanno) algorithm, together with a cubic or quadratic step length method (STEPBT), see e.g. Dennis and Schnabel (1983).

2. Keeping fixed the LPPL parameters $\hat{\Phi} = [\hat{A}, \hat{B}, \hat{C}, \hat{t}_c, \hat{\beta}, \hat{\omega}, \hat{\phi}]$ computed in the first stage, estimate the parameters of the short term stochastic component $[\rho, \alpha_0, \alpha_1, \alpha_2]$.

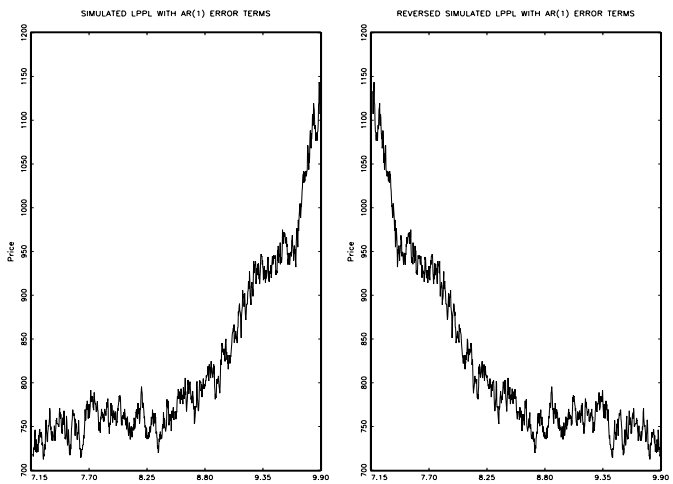


Figure 1: Simulated LPPL with AR(1) error terms and its reverse. Parameters: $A = 7.16$, $B = -0.43$, $C = 0.035$, $\beta = 0.35$, $\omega = 4.15$, $\phi = 2.07$, $T_c = 9.92$, $\rho = 0.88$, $\alpha_0 = 0.00007$.

In case of poor starting values, or when the bubble has just start forming (Jiang et al. (2010) and Sornette (2003a) remarked that a bubble cannot be diagnosed more than 1 year in advance of the crash), the numerical computation can be further eased by considering one additional step. The 3-step ML approach used in Fantazzini (2010a) is described below:

1. Reverse the original time series and then consider the first temporal observation as if it was the date of the crash, that is set $t_c = t_1$. Estimate the remaining LPPL parameters $[A, B, C, \beta, \omega, \phi]$ for the case of an anti-bubble by using the BFGS (Broyden, Fletcher, Goldfarb, Shanno) algorithm, together with a cubic or quadratic step length method (STEPBT).

2. Use the estimated parameters in the previous step as *starting values* for estimating all the LPPL parameters, by using again the reversed times series.

3. Keeping fixed the LPPL parameters $\hat{\Phi} = [\hat{A}, \hat{B}, \hat{C}, \hat{t}_c, \hat{\beta}, \hat{\omega}, \hat{\phi}]$ computed in the second stage, estimate the parameters of the short term stochastic component $[\rho, \alpha_0, \alpha_1, \alpha_2]$.

Being a multi-stage estimation process, the asymptotic efficiency is lower than the 1-step full ML estimation. However, the dramatic improvement in numerical convergence and the improved efficiency in small-to-medium sized samples more than justify the multi-step procedure. An (unreported) simulation study confirms the benefits of this procedure in small-to-medium datasets.

5 Diagnosing Bubbles in the Making

The main method that we have considered so far to detect financial bubbles is by fitting a LPPL model to a price series. However, in order to reduce the possibility of false alarms, it is good practice to implement a battery of tests, so that a prediction must pass all tests to be considered worthy, see e.g. Sornette and Johansen (2001) and Jiang et al. (2010).

5.1 Diagnostic Tests based on the LPPL fitting residuals

We have seen in section 3.2 that Lin et al. (2009) proposed a model for financial bubbles where the LPPL fitting residuals follows a mean-reverting Ornstein-Uhlenbeck process. This implies that the corresponding residuals follow an AR(1) process and we can test this hypothesis by using unit root test.

Lin et al. (2009) use Phillips-Perron (PP) and Augmented Dickey-Fuller (ADF) unit-root tests, where the null hypothesis H_0 is the presence of a unit root. Unfortunately, a well known shortcoming of the previous two unit root tests is their low power when the underlying data-generating process is an AR(1) process with a coefficient close to one. Therefore, we suggest to consider also the test proposed by Kwiatkowski, Phillips, Schmidt and Shin (KPSS, 1992), where the null hypothesis is a stationary process. Considering the null hypothesis of a stationary process and the alternative of a unit root allows us to follow a conservative testing strategy: if we reject the null hypothesis, we can be confident that the series has indeed a unit root; but if the results of the previous two tests indicate a unit root while the result of the KPSS test indicates a stationary process, one should be very cautious and opt for the latter result.

The KPSS test is implemented in the most common statistical and econometric software: see e.g. Griffiths et al. (2008) for applications with Eviews as well as the `Eviews` User's Guide (version 5 or higher), Pfaff (2008) for a description of unit root tests in `R`, together with the many routines written for `Gauss` and `Matlab` which can be found on the web.

5.2 Diagnostic Tests based on Rational Expectation Models with Stochastic Mean-Reverting Termination Times

Lin and Sornette (2009) propose two models of transient bubbles in which their termination dates occur at some potential critical time $\tilde{t}_c$, which follows a stationary process with a unconditional mean T_c. The main advantage of these models is the possibility to compute the potential critical time without the need to estimate the complex stochastic differential equation describing the underlying price dynamics. Interestingly, the rational arbitrageurs discussed in Lin and Sornette (2009) can detect bubbles, but they can not make a deterministic forecast because they have little knowledge about the other arbitrageurs' beliefs about the process governing the stochastic critical time $\tilde{t}_c$. The heterogeneity of the rational agents' expectations determines a synchronization problem among these arbitrageurs, thus allowing the financial bubble to survive till its theoretical end time, see also Abreu and Brunnermeier (2003) for a similar model. Moreover, the two models by Lin and Sornette (2009) can be tested and they allow to diagnose financial bubbles in the making in real time. Besides, both models highlight the importance of positive feedback, that is when a high price pushes even further the demand so that the return and its volatility tend to be a nonlinear accelerating function of the price. This positive feedback mechanism is quantified by a unique exponent m, which is larger than 1 (respectively 2, for the second model) when we are in a bubble regime.

5.2.1 A Test Based on a Finite-Time Singularity in the Price Dynamics with Stochastic Critical Time

The first model proposed by Lin and Sornette (2009) views a bubble as a faster-than-exponential accelerating stochastic price, which leads to a finite-time singularity in the price dynamics at a stochastic critical time. They show in their Proposition 1 that the

price dynamics in a bubble regime follows this process,

$$\begin{aligned} &p(t) = K(\tilde{T}_c - t)^{-\beta}, \\ &\beta = \tfrac{1}{m-1}, \quad K = \left(\tfrac{\beta}{\mu}\right)^{\beta}, \quad T_c = \tfrac{\beta}{\mu} p_0^{-\frac{1}{\beta}}, \quad \tilde{T}_c = T_c + \tilde{t}_c \end{aligned} \tag{26}$$

where μ is the instantaneous return rate, p_0 denotes the price at the start time of the bubble at $t = 0$, while the critical time $\tilde{t}_c$ follows an Ornstein-Uhlenbeck process with zero unconditional mean (see Lin and Sornette (2009) for the full derivation of the model). This last property provides that the end of the bubble cannot be forecasted with certainty but it is a stochastic variable, while the time T_c can be interpreted as the consensus forecast formed by rational arbitrageurs of the stochastic critical time $\tilde{T}_c$. In fact, we have that

$$E[\tilde{T}_c] = E[T_c + \tilde{t}_c] = T_c \tag{27}$$

In order to build a diagnostic test for financial bubbles, Lin and Sornette (2009) invert (26) to obtain an expression for the critical time series $\tilde{T}_{c,i}$:

$$\tilde{T}_{c,i}(t) = \frac{1}{K} \frac{1}{[p(t)]^{1/\beta}} + t, \quad t = t_i - 749, \ldots, t_i \tag{28}$$

where $\tilde{T}_{c,i}$ is defined over the time window i ending at time t_i, and they consider time windows of 750 trading days that slide with a time step of 25 days from the beginning to the end of the available financial time series. We remark that $p(t)$ is known, while the parameters K and β have to be estimated. The previous inversion aims at transforming a non-stationary possibly explosive price process $p(t)$ into what should be a stationary time series $\tilde{T}_{c,i}$, in absence of misspecification. Therefore, we can then estimate T_c according to (27) by using the arithmetic average of $\tilde{T}_{c,i}(t)$,

$$T_{c,i} = \frac{1}{750} \sum_{i=1}^{750} \tilde{T}_{c,i}(t) \tag{29}$$

so that the fluctuations $\tilde{t}_{c,i}(t)$ can be computed as

$$\tilde{t}_{c,i}(t) = \tilde{T}_{c,i}(t) - T_{c,i} \tag{30}$$

The first test by Lin and Sornette (2009) consists of the following two steps:

1. Perform a bivariate grid search over the parameter space of K and β to find the ten best pairs (K, β) such that the resulting time series $\tilde{t}_{c,i}(t)$ given by (30) rejects a standard unit-root test of non-stationarity at the 99.5% significance level. Lin and Sornette (2009) employ the ADF test, but the addition of the KPSS would be

advisable. Needless to say, only a subset of the windows will reject the null hypothesis of a unit root (for the ADF test) or will not reject the null of stationarity (for the KPSS test).

2. If there are time windows for which there are selected pairs (K, β) according to the previous step, select the pair with the smallest variance for its corresponding time series $\tilde{t}_{c,i}(t)$. This gives the optimal pair K_i^* and β_i^* which provides the closest approximation to a stationary time series for $\tilde{t}_{c,i}(t)$ given by (30). For a given window i, an alarm is declared when

   - $\beta^* > 0$, which yields $m > 1$;
   - $T_{c,i} - t_i < 750$, which implies that the termination time of the bubble is not too far. Lin and Sornette (2009) also consider two additional alarm levels: $T_{c,i} - t_i < 500$ and $T_{c,i} - t_i < 250$.

The idea of the last step is that the closer we are to the end of the financial bubble, the stronger should be the evidence for the bubble as a faster-than-exponential growth, and the alarms should be diagnosed repeatedly by several successive windows.

5.2.2 A Test Based on a Finite-Time Singularity in the Momentum Price Dynamics With Stochastic Critical Time

The main disadvantage of the previous model is that the price diverges when approaching the critical time $\tilde{T}_c$ at the end of the bubble. Therefore, Lin and Sornette (2009) consider a second model where the price remains always finite and a bubble is a regime characterized by an accelerating momentum ending at a finite time singularity with a stochastic critical time. They show in their Proposition 3 that the log-price $y(t) = \ln p(t)$ in a bubble regime follows this process,

$$\begin{aligned} &y(t) = A - B(T_c + \tilde{t}_c(t) - t)^{1-\beta}, \\ &\beta = \tfrac{1}{m-1}, \quad T_c = \tfrac{\beta}{\mu} x_0^{\frac{1}{\beta}}, \quad x_0 := x(t=0) \quad B = \tfrac{1}{1-\beta}(\beta/\mu)^{\beta} \quad \tilde{T}_c = T_c + \tilde{t}_c \end{aligned} \tag{31}$$

where μ is the instantaneous return rate, A is a constant, $x(t) = dy/dt$ denotes the effective price momentum, i.e. the instantaneous time derivative of the logarithm of the price, while the critical time $\tilde{t}_c$ follows an Ornstein-Uhlenbeck process with zero unconditional mean

(see Lin and Sornette (2009) for the full derivation of this model). In this second model, it is the high price momentum x which pushes higher the demand, so that the return and its volatility become nonlinear accelerating functions. Instead, in the first model, it is the price that provides a positive feedback on future prices, rather than the price momentum. Using a procedure similar to (28) to transform a non-stationary possibly explosive log-price process $y(t)$ into a stationary time series $\tilde{T}_{c,i}$, Lin and Sornette (2009) invert (31) to obtain an expression for the critical time series $\tilde{T}_{c,i}$:

$$\tilde{T}_{c,i}(t) = \left(\frac{A - \ln p(t)}{B}\right)^{\frac{1}{1-\beta}} + t, \quad t = t_i - 899, \ldots, t_i \tag{32}$$

where $\tilde{T}_{c,i}$ is defined over the time window i ending at time t_i, and they consider time windows of 900 trading days that slide with a time step of 25 days from the beginning to the end of the available financial time series. We can then estimate $T_{c,i}$ according to (29) by computing the arithmetic average of $\tilde{T}_{c,i}(t)$ (with 750 replaced by 900), whereas the fluctuations $\tilde{t}_{c,i}(t)$ around $T_{c,i}$ can be computed by using (30). Similarly to the first model, we remark that $p(t)$ is known while the parameters A, B and β have to be estimated.

The second test by Lin and Sornette (2009) consists of the following two steps:

1. Perform a trivariate grid search over the parameter space of A, B and β to find the ten best triplets (A, B, β) such that the resulting time series $\tilde{t}_{c,i}(t)$ given by (30) rejects a standard unit-root test of non-stationarity at the 99.5% significance level. Lin and Sornette (2009) employ the ADF test, but again the addition of the KPSS would be advisable.

2. If there are time windows for which there are selected triplets (A, B, β) according to the previous step, select the pair with the smallest variance for its corresponding time series $\tilde{t}_{c,i}(t)$. This gives the optimal triplet A_i^*, B_i^* and β_i^* which provides the closest approximation to a stationary time series for $\tilde{t}_{c,i}(t)$ given by (30). For a given window i, an alarm is declared when

   - $0 < \beta_i^* < 0$, which yields $m > 2$ and is called *level 1* filter. Lin and Sornette (2009) also consider two additional alarm levels: $m > 2.5$ (*level 2*) and $m > 3$ (*level 3*);

- $-25 \leq T_{c,i} - t_i \leq 50$.

The stronger upper bound on $T_{c,i} - t_i$ stems from the fact that the finite-time singularity in the price momentum is a weaker singularity which can only be observed only close to the critical time. The lower bound of -25 days is due to the fact that the analysis is performed in sliding windows with a time step of 25 days. Using a dataset covering the last 30 years of the SP500 index, the NASDAQ composite index and the Hong Kong Hang Seng index, Lin and Sornette (2009) found that the second diagnostic method seems to be more reliable and with fewer false alarms than the first method analyzed in Section 5.2.1.

5.3 Graphical tools: The Crash Lock-In Plot (CLIP)

Fantazzini (2010a) proposed a graphical tool that proved to be useful to track the development of a bubble and to understand whether a possible crash is in sight, or at least a bubble deflation. The idea is to plot on the horizontal axis the *date of the last observation in the estimation sample*, while on the vertical axis the *estimated crash date* $\hat{t}_c$ computed by fitting the LPPL to the data: if a change in the stock market regime is approaching, then the recursively estimated $\hat{t}_c$ should stabilize around a constant value close to the critical time. Fantazzini (2010a) called such a plot the Crash Lock-In Plot (CLIP).

This idea can be easily justified theoretically by resorting to the models proposed by Lin and Sornette (2009), in which the critical time $\tilde{T}_c$ follows an Ornstein-Uhlenbeck process. We report in Figure 2 the CLIPs for the Chinese Shanghai Composite Index in July 2009, a case which was analyzed in details in Bastiaensen et al. (2009) and Jiang et al. (2010), and for the SP500 in July 2007, being that the peak of the market in the decade. We used data spanning from the global minima till 1 day before the market peak.

6 An Application: The Burst of the Gold Bubble in December 2009

The gold market peaked on the 02/12/2009, hitting the record high at \$1,216.75 an ounce in Europe, and then started falling on the 04/12/2009, losing more than 10% in two weeks. The main concerns cited to be behind this bubble were the future prospects for

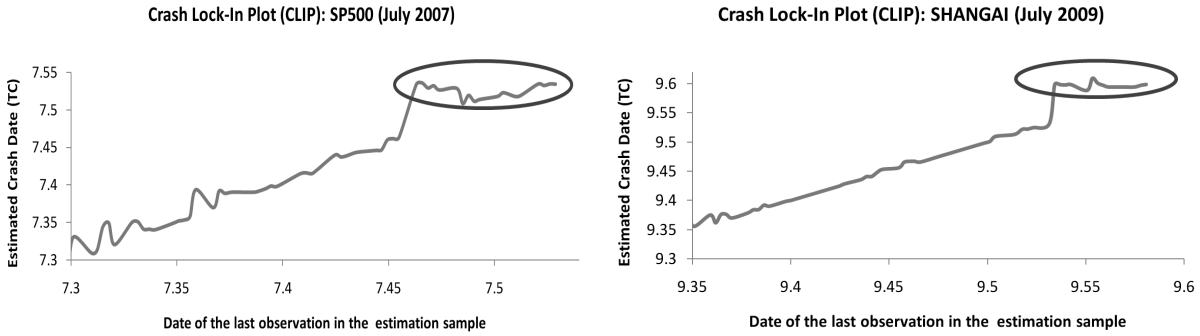


Figure 2: Crash Lock-In Plots for the SP500 and Shanghai Composite Index

a week dollar as well inflationary fears, see e.g. Mogi (Reuters, 2009) and White (The Telegraph, 2009). However, there were also some worried calls about the possibility of a gold bubble: the prestigious magazine *Fortune* wrote on the 12/10/2009 that "*...Signs of gold fever are everywhere ...*" but "*...amid the buying frenzy and after a decade-long run-up that has seen the price quadruple, is gold still a smart investment? The simple answer: Wherever the price of gold is headed in the long term, several market watchers say the fundamentals indicate that gold is poised to fall*" (Cendrowski, "*Beware the gold bubble*", 2009). Interestingly, on the day the gold price peaked, i.e. 02/12/2009, Hu Xiaolian, a vice-governor at the People's Bank of China, told reporters in Taipei that "*... gold prices are currently high and markets should be careful of a potential asset bubble forming...*", see the original report by R. Tung for Reuters (2009). The gold price, starting from November the 12th 2008 (which represents the global minima over a three-year span) till the end of January 2010 is reported in Figure 3. The same Figure reports also the "Search Volume Index" by `Google Trends`, which computes how many searches have been done for the term "Gold Price" on Google over time[3].

The "Search Volume Index" is an interesting tool because it allows us to get some insights as to when the bubble started: looking at Figure 3, we can notice that a massive interest around gold started to build during the year 2008, just before the price minima in November 2008. Therefore, we expect that a possible LPPL model can be fitted using data starting from the year 2008.

[3]See `http://www.google.com/intl/en/trends/about.html` for more details. In this case, the time span starts from 2004, which is the first year available for this analysis.

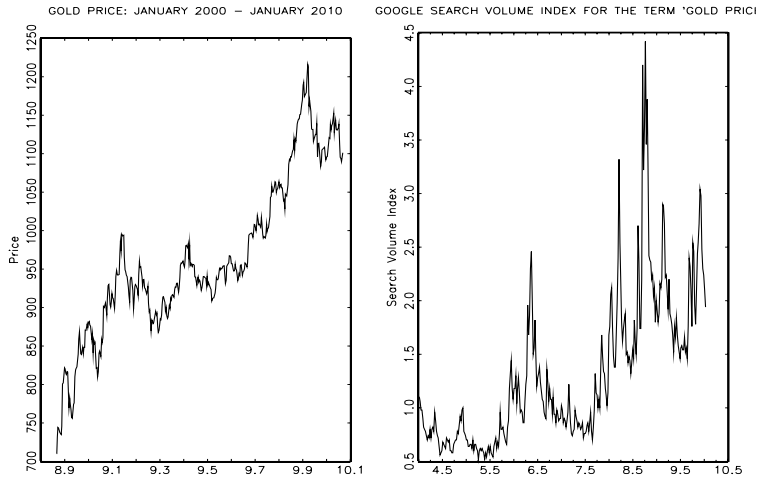


Figure 3: Gold price and `Google` Search Volume Index. Time t converted in units of 1-year

6.1 LPPL Fitting With Varying Window Sizes

Jiang et al. (2010) tested the stability of LPPL estimation parameters by varying the size of the estimation samples and adopting the strategy of fixing one endpoint and varying the other one, see also Sornette and Johansen (2001). By sampling many intervals as well as by using bootstrap techniques, they obtained probabilistic predictions on the time intervals in which a given bubble may end and lead to a new market regime (which may not necessarily be a crash, but also a transition to a plateau or a slower decay). Following their example, we fit the logarithm of the gold price by using the LPPL eq. (22) in *shrinking windows* and in *expanding windows*. The shrinking windows have a fixed end date $t_2 = 01/12/2009$, while the starting date t_1 increases from 12/11/2008 to 17/08/2009 in steps of five (trading) days. The expanding windows have a fixed starting date $t_1 = 12/11/2008$ while the end date t_2 increases from 17/08/2009 to 01/12/2009 in steps of five (trading) days.

Given the stochastic nature of the initial parameter selection and the noisy nature of the underlying generating processes, we employed four estimation algorithms: the original Taboo Search algorithm proposed by Cvijović and Klinowski (1995), the 2-step nonlinear optimization by Johansen et al. (2000), the Pure Random Search (PRS) and the 3-step ML approach by Fantazzini (2010a)[4]. The estimation results are then filtered by the

[4]Similarly to Cvijović and Klinowski (1995), we found that GA has a performance in between Taboo Search and PRS. However, even though PRS is computationally inefficient, it has the benefit to potentially visit regions of the parameter space that sometimes are not visited by the previous algorithms. This is

following LPPL conditions, which were also used in Jiang et al. (2010) for the case of Chinese bubbles: $\hat{t}_c > t_2$, $B < 0$ and $0 < \beta < 1$. The selected $\hat{t}_c$ are then used to compute the 20%/80% and 5%/95% quantile range of values of the crash dates which are reported in Figure 4: the left plot shows the ranges which are obtained by considering the filtered results from all four estimation methods, whereas the right plot shows the ranges obtained by considering only the 2-step nonlinear optimization and the 3-step ML approach.

As expected, the original Taboo Search and the PRS are very inefficient methods compared to the competing 2-step and 3-step approaches and deliver very large quantile ranges. Nevertheless, the two medians $\hat{t}_c$, equal to 11/12/2009 for the left plot and 05/12/2009 for the right plot, are very close to the actual market peak date which occurred on the 02/12/2009 (i.e. 9.9206 when converted in units of one year). Moreover, if we consider only the most efficient methods, the 20%/80% quantile interval is rather close and precise and diagnoses that the critical time t_c for the end of the bubble and the change of market regime lies in the time sample 03/12/2009 - 11/12/2009 (the market started to fall on the 04/12/2009).

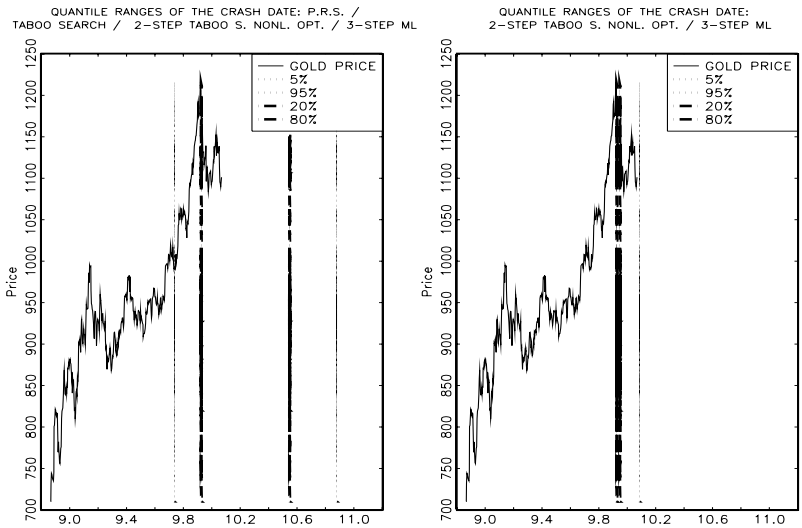


Figure 4: Quantile ranges of the crash date.

why we consider it in our analysis in the place of GA.

6.2 Diagnostic Tests based on the LPPL fitting residuals

We have discussed in section 3.2 that Lin et al. (2009) proposed a model for financial bubbles where the LPPL fitting residuals follows a mean-reverting Ornstein-Uhlenbeck process. Therefore, the corresponding residuals should follow a stationary AR(1) process and this hypothesis can be tested by using unit root tests. We employed ADF and KPSS tests: a rejection of the null hypothesis in the first test together with a failure to reject the null in the second test, indicates that the residuals are stationary and thus compatible with an O-U process.

We used the residuals resulting from the previous estimation windows and numerical algorithms, that is 48 shrinking windows, 19 expanding windows and 4 estimation methods, which gives a total of 268 calibrations. The fraction P_{LPPL} of these different windows that met the LPPL conditions was equal to $P_{LPPL} = 60.1\%$. The conditional probability that, out of the fraction P_{LPPL} of windows that satisfied the LPPL conditions, the null hypothesis of non-stationarity was rejected for the residuals, was equal to $P_{Stat.Res.|LPPL} = 100\%$ when using the ADF test at the probability level $\alpha = 0.001$. As for the KPSS test, the null of stationarity was not rejected at the 10% level or higher in all cases which satisfied the LPPL conditions. Therefore, this empirical evidence is comparable with the results reported by Jiang et al. (2010) for the case of the 2005-2007 and 2008-2009 Chinese stock market bubbles.

6.3 Diagnostic Tests based on Rational Expectation Models with Stochastic Mean-Reverting Termination Times

We employed the two diagnostics proposed by Lin and Sornette (2009) to detect the presence of a bubble (and reviewed in Section 5.2) to the gold price time series, from November the 12th 2008 to December the 1st 2009. As previously discussed, we considered both the ADF and KPSS unit root tests. Moreover, we also used time windows of 500 and 250 trading days to compute the critical time series $\tilde{T}_{c,i}$ in (28)-(29) and (32), together with the original 750 trading days for the first diagnostic and 900 for the second one. The rationale of this choice is that a long time span may include data which are not observed during a bubble regime but during a standard Geometric Brownian Motion regime (or

other regimes). Of course, reducing the time window implies a loss of efficiency.

Interestingly the first procedure did not flag any alarm for the presence of a bubble, whereas the second one flagged three series of alarms close to three important price falls, see Figure 5: the first group of alarms was centered around the local market peak on the 20/02/2009 when gold reached the value of \$995.3 an ounce, very close to the important psychological barrier of \$1000, and after two days it started falling, losing more than 10% in two weeks. The second group of alarms was centered around the local market peak on the 02/06/2009 when gold reached the value of \$982.9 and after two days it started falling, losing more than 5% in a week. Finally, the third group of alarms was centered around the global market peak on the 02/12/2009 when gold reached the value of \$1216 an ounce.

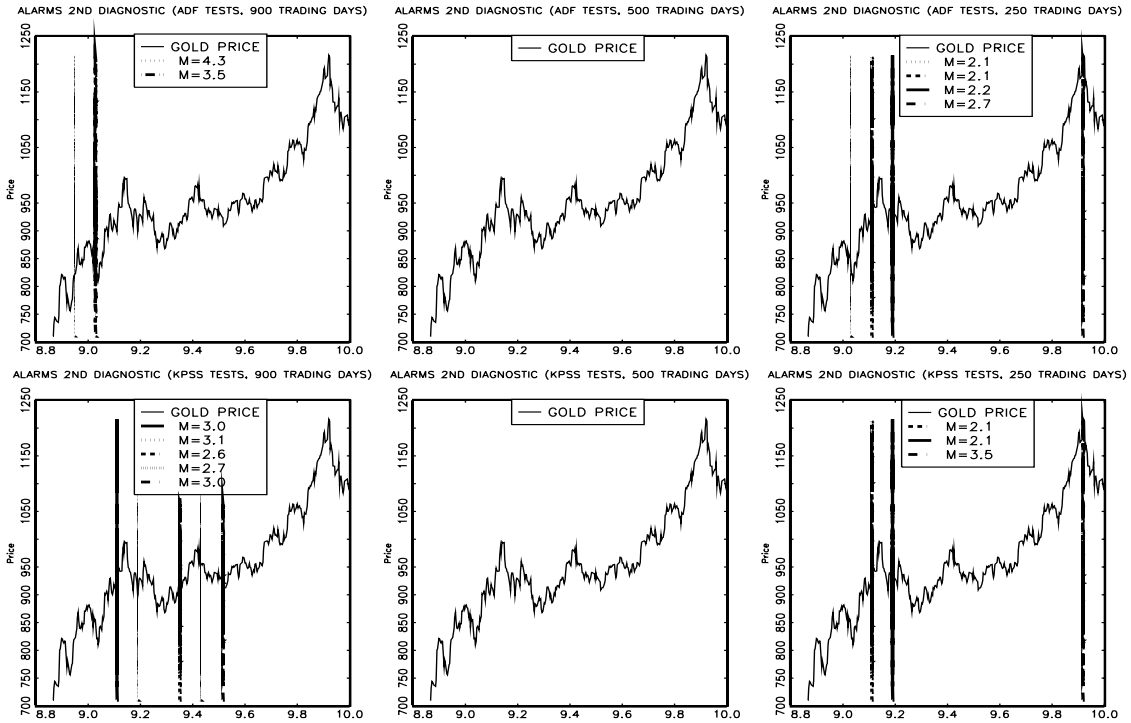


Figure 5: Logarithm of the gold price and corresponding alarms as vertical lines indicating the ends of the windows of T trading days, in which the second diagnostic flags an alarm for the presence of a bubble. The value of the exponent m for each alarm is reported in the legend .

This empirical evidence seems to suggest that the KPSS test provides more precise alarms than the ADF test, which is not a surprise given the well known limitations of the latter test. Moreover, a time window of 250 observations delivers more reliable flags for the presence of a bubble (or imminent price falls) than longer time spans, being more robust

to market regime changes. However, a time window of 900 observations still provides useful information.

6.4 Crash Lock-In Plots (CLIPs) for the Gold Bubble

The CLIP plots on the horizontal axis the date of the last observation in the estimation sample, while on the vertical axis the estimated crash date $\hat{t}_c$ computed by fitting the LPPL to the data. Following the previous empirical evidence as well as the one reported in Lin and Sornette (2009) and Fantazzini (2010a), we computed the CLIP by fitting the data with two rolling estimation windows of 900 and 250 days, and by using the simple average of the estimated $\hat{t}_c$ resulting from the four estimation algorithms discussed in Section 6.1. We used data spanning from the 12/11/2008 till 1 day before the global market peak on the 02/12/2009. The two CLIPs are reported in Figure 6.

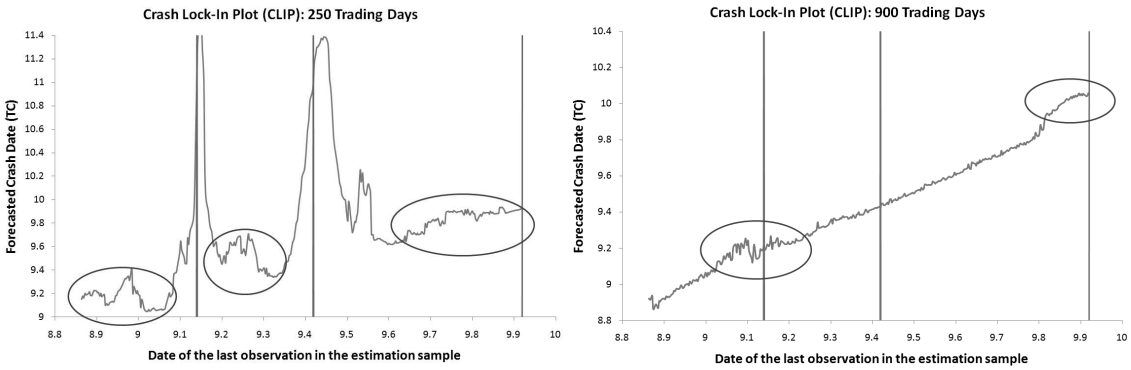


Figure 6: Crash Lock-In Plots for the Gold price series. The three vertical lines correspond to the two local market peaks on the 20/02/2009 and the 02/06/2009, and to the global market peak on the 02/12/2009, respectively.

Not surprisingly, the indications provided by the two CLIPs are rather similar to those provided by the second diagnostic test by Lin and Sornette (2009) discussed in the previous section: the recursive forecasted crash dates computed with time windows of 250 trading days stabilize around three constant values which are very close to the dates corresponding to the two local market peaks on the 20/02/2009 and the 02/06/2009, and to the global market peak on the 02/12/2009. The indications from the second CLIP computed with time windows of 900 trading days are somewhat weaker, but confirm the previous alarms. As expected, the estimates computed with smaller time spans are more noisy than those

with longer time spans.

7 Conclusions

We presented an easy-to-use and self-contained guide for modelling and detecting financial bubbles with Log Periodic Power Laws, which contains the sufficient steps to derive the main models and discusses the important aspects for practitioners and researchers. We reviewed the original JLS model and we discussed early criticism to this approach and recent generalizations proposed to answer these remarks. Moreover, we described three different estimation methodologies which can be employed to estimate LPPLs models. We then examined the issue of diagnosing bubbles in the making by using a set of different techniques, that is by considering diagnostic tests based on the LPPL fitting residuals, diagnostic tests based on rational expectation models with stochastic mean-reverting termination times, as well as graphical tools useful for capturing bubble development and for understanding whether a crash is in sight or not. We finally presented a detailed empirical application devoted to the burst of the gold bubble in December 2009, which highlighted how a series of different diagnostics flagged an alarm for the presence of a bubble before prices started to fall.

References

[1] Abreu, D., and M. K., Brunnermeier. 2003. Bubbles and crashes. *Econometrica*, 71, no. 1: 173204.

[2] Bastiaensen, K., Cauwels, P., Sornette, D., Woodard, R., and W.X. Zhou. 2009. The Chinese equity bubble: Ready to burst. http://arxiv.org/abs/0907.1827.

[3] Bauwens, L., Lubrano, M., and J.F. Richard. 2000. *Bayesian Inference In Dynamic Econometric Models*. New York: OUP.

[4] Bernardo, J.M. and A.F.M. Smith. 1994. *Bayesian Theory*. New York: Wiley.

[5] Blanchard, O. J. 1979. Speculative bubbles, crashes and rational ex- pectations. *Economics Letters*, 3:387-389.

[6] Cendrowski, S. 2009. Beware the gold bubble. http://money.cnn.com/2009/10/06/pf/gold_investing_bubble.fortune/index.htm .

[7] Chang , G. and J. Feigenbaum. 2006. A Bayesian analysis of log-periodic precursors to financial crashes, Quantitative Finance 6 (1), 15-36.

[8] Cvijović, D. and J. Klinowski. (1995). Taboo search An approach to the multiple minima problem. *Science*, 267, no. 5: 664-666.

[9] Dennis, Jr., J.E., and R.B., Schnabel. 1983. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs: Prentice-Hall.

[10] Derrida, B., De Seze, L., and Itzykson, C. 1983. Fractal structure of zeros in hierarchical models. *Journal of Statistical Physics*, 33: 559-569

[11] Fantazzini, D. 2010a. Modelling Bubbles And Anti-Bubbles In Bear Markets: A Medium-Term Trading Analysis. In *Handbook of Trading*, ed. G. Gregoriou, 365-388. New York: McGraw-Hill.

[12] Fantazzini, D. 2010b. Modelling and Forecasting the Global Financial Crisis: Initial Findings using Heterosckedastic Log-Periodic Models. *Economics Bulletin*, 30, no. 3: 1833-1841.

[13] Gazola,L., Fernandes, C., Pizzinga, A., and R. Riera. 2008. The log-periodic-AR(1)-GARCH(1,1) model for financial crashes, *The European Physical Journal B*, 61: 355-362.

[14] Goldenfeld, N. 1992. *Lectures on phase transitions and the renormalization group*. Reading: Addison-Wesley Publishing Company.

[15] Griffiths, W.E., Hill, R., and G.C. Lim. (2008). *Using EViews for Principles of Econometrics*. 3rd Edition, Hoboken, NJ: Wiley.

[16] Gulsten, M., Smith, E. A. and D.M. Tate. 1995. A Genetic Algorithm Approach to Curve Fitting. *International Journal of Production Research*, 33, no. 7: 1911-1923.

[17] Gürkaynak, R. 2008. Econometric tests of asset price bubbles: taking stock. *Journal of Economic Surveys*, 22, 1:166186.

[18] Hayashi, F. 2000. *Econometrics*. Princeton: PUP.

[19] Jacobsson, E. 2009. How to predict crashes in financial markets with the Log-Periodic Power Law. Master diss., Department of Mathematical Statistics, Stockholm University.

[20] Jiang, Z.Q., Zhou, W.H., Sornette, D., Woodard, R., Bastiaensen, K., and P. Cauwels. 2009. Bubble Diagnosis and Prediction of the 2005-2007 and 2008-2009 Chinese stock market bubbles (2009). Journal of Economic Behavior and Organization 74, 149-162 (2010). Available at http://arxiv.org/abs/0909.1007

[21] Johansen, A. 2002. Comment on "Are financial crashes predictable?", *Europhysics Letters*, 60, no. 5: 809-810.

[22] Johansen, A. 2003. Characterization of large price variations in financial markets. *Physica A*, 324: 157-166.

[23] Johansen, A., and D., Sornette. 1999. Financial anti-bubbles: Log-periodicity in Gold and Nikkei collapses. *International Journal of Modern Physics C*, 10, no. 4: 563-575.

[24] Johansen, A., and D., Sornette. 2000. Evaluation of the quantitative prediction of a trend reversal on the Japanese stock market in 1999. *International Journal of Modern Physics C*, 11, no. 2: 359-364.

[25] Johansen, A., and D., Sornette. 2004. Endogenous versus Exogenous Crashes in Financial Markets, in *Contemporary Issues in International Finance*, Nova Science Publishers. Reprinted as 'Shocks, Crashes and Bubbles in Financial Markets', *Brussels Economic Review (Cahiers economiques de Bruxelles)*, 49 (3/4), Special Issue on Nonlinear Analysis, 2006.

[26] Johansen, A., Ledoit, O. and D., Sornette. 2000. Crashes as critical points, *International Journal of Theoretical and Applied Finance*, 3, no. 2: 219-255.

[27] Kaizoji, T., and D. Sornette., Market Bubbles and Crashes, in press in the *Encyclopedia of Quantitative Finance* (Wiley) http://www.wiley.com//legacy/wileychi/eqf/ (long version available at http://arXiv.org/abs/0812.2449)

[28] Koop, G. 2003. *Bayesian Econometrics*. Chichester: Wiley.

[29] Krugman, P. 1998. The Confidence Game, How Washington worsened Asia's crash. *The New Republic*, October 5.

[30] Kwiatkowski, D., Phillips, P., Schmidt, P., and Y., Shin. 1992. Testing the null hypothesis of stationary against the alternative of a unit root. *Journal of Econometrics*, 54: 159-178.

[31] Laloux, L., Potters, M., Cont,R., Aguilar, J.P and J.P. Bouchaud. 1999. Are financial crashes predictable?, *Europhysics Letters*, 45, no. 1: 1-5.

[32] Landau, L.D. and E.M. Lifshitz. (1980). *Statistical Physics, Part 1 and Part 2*. 3rd ediion, Butterworth-Heinemann.

[33] Lin, L. and D. Sornette. 2009. Diagnostics of Rational Expectation Financial Bubbles with Stochastic Mean-Reverting Termination Times. ETH working paper, http://arxiv.org/abs/0911.1921 .

[34] Lin, L., Ren., R.E. and D. Sornette. 2009. A Consistent Model of Explosive Financial Bubbles With Mean-Reversing Residuals, *Quantitative Finance*, in press. Available at http://arxiv.org/abs/0905.0128 and http://papers.ssrn.com/abstract=1407574.

[35] Mogi, C. 2009. Gold hits record high as dollar sets new lows. *Reuters*, November 26 2009, http://www.reuters.com/article/goldMktRpt/idUSSP7486520091126

[36] Ng, S. and P. Perron. 2001. Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power. *Econometrica*, 69, no. 6: 1519-1554.

[37] Onsager, L. 1944. Crystal statistics. A two-dimensional model with an order-disorder transition. *Physics Review*, 65: 117-149.

[38] Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R*. New York: Wiley.

[39] Reinhart, C. and K. Rogoff. 2009. *This Time is Different: Eight Centuries of Financial Folly*. Princeton: PUP.

[40] Shefrin. H. 2005. *A Behavioral Approach to Asset Pricing. Academic Press Advanced Finance Series*. Burlington: Academic Press.

[41] Sornette, D. 2003a. *Why Stock Markets Crash (Critical Events in Complex Financial Systems)*, Princeton: PUP.

[42] Sornette, D. 2003b. Critical market crashes. *Physics Reports*, 378, no. 1: 1-98.

[43] Sornette, D. 2009. Dragon-Kings, Black Swans and the Prediction of Crises, *International Journal of Terraspace Science and Engineering*, 2, no. 1: 1-18.

[44] Sornette, D. and A., Johansen. 1997. Large financial crashes, *Physica A*, 245: 411-422.

[45] Sornette, D. and A., Johansen. 2001. Significance of log-periodic precursors to financial crashes, *Quantitative Finance*, 1, no. 4: 452-471.

[46] Sornette, D., and R. Woodard. 2010. Financial Bubbles, Real Estate bubbles, Derivative Bubbles, and the Financial and Economic Crisis, in "Proceedings of APFA7 (Applications of Physics in Financial Analysis)", *New Approaches to the Analysis of Large-Scale Business and Economic Data*, Misako Takayasu, Tsutomu Watanabe and Hideki Takayasu, eds., Springer. Available at http://www.thic-apfa7.com/en/htm/index.html , http://arxiv.org/abs/0905.0220, http://web.sg.ethz.ch/wps/CCSS-09-00003/, and http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1407608

[47] Sornette, D. and W.X. Zhou. 2006. Predictability of Large Future Changes in major financial indices. *International Journal of Forecasting*, 22: 153-168.

[48] Sornette, D., Johansen, A. and J.P., Bouchaud. 1996. Stock market crashes, Precursors and Replicas, *Journal de Physique I*, 6: 167-175.

[49] Sornette, D., Woodard, R. and W.X. Zhou. 2008. The 2006-2008 Oil Bubble and Beyond. ETH Zurich preprint. http://arXiv.org/abs/0806.1170

[50] Sornette, D., Woodard, R. and W.X. Zhou. 2009. The 2006-2008 Oil Bubble: Evidence of Speculation, and Prediction. *Physica A*, 388: 1571-1576.

[51] Stanley, H. E. 1971. *Introduction to Phase Transitions and Critical Phenomena*. New York: OUP.

[52] Stock J.H. 1994. Unit roots, structural breaks and trends. In *Handbook of econometrics*, ed. R.F. Engle, D.L. McFadden, vol. 4, ch. 47. Elsevier.

[53] Stock J.H., and M.W. Watson. (2002). *Introduction to Econometrics*. Addison Wesley.

[54] Tsay, R. 2005. *Analysis of Financial Time Series*. Chicago: Wiley.

[55] Tung, R. 2009. China c.Bank: Gold Prices High, Warns of Bubble. *Reuters*, December 02 2009, http://www.reuters.com/article/idUSTPU00193020091202 .

[56] White, G. 2009. Golden times for precious metals but pick carefully. *The Telegraph*, September 08 2009, http://www.telegraph.co.uk/finance/markets/questor/6156983/Golden-times-for-precious-metals-but-pick-carefully.html .

[57] Zhou, W.X., and D., Sornette. 2003. 2000-2003 Real Estate Bubble in the UK but not in the USA. *Physica A*, 329: 249-263.

[58] Zhou, W.X., and D., Sornette. 2005. Testing the Stability of the 2000-2003 US Stock Market Antibubble. *Physica A*, 348: 428-452.

[59] Zhou, W.X., and D., Sornette. 2006. Is There a Real-Estate Bubble in the US?. *Physica A*, 361: 297-308.

[60] Zhou, W.X., and D., Sornette. 2008. Analysis of the real estate market in Las Vegas: Bubble, seasonal patterns, and prediction of the CSW indexes. *Physica A*, 387: 243-260.

[61] Zhou, W.X., and D., Sornette. 2009. A case study of speculative financial bubbles in the South African stock market 2003-2006. *Physica A*, 388: 869-880.

Appendix 10 "Nowcasting and Forecasting the Monthly Food Stamps Data in the US Using Online Search Data."

Fantazzini, D. (2014). Nowcasting and Forecasting the Monthly Food Stamps Data in the US Using Online Search Data. *PloS one*, 9(11), e111894. DOI: https://doi.org/10.1371/journal.pone.0111894

Permisson to post the Publisher's version/PDF: according to https://www.plos.org/open-access/, "*PLOS applies the Creative Commons Attribution (CC BY) license to works we publish. This license was developed to facilitate Open Access—namely, free immediate access to, and unrestricted reuse of, original works of all types. Under this license, authors agree to make articles legally available for reuse, without permission or fees, for virtually any purpose. Anyone may copy, distribute or reuse these articles, as long as the author and original source are properly cited. Additionally, the journal platform that PLOS uses to publish research articles is Open Source.*".

OPEN ACCESS Freely available online

PLOS | ONE

Nowcasting and Forecasting the Monthly Food Stamps Data in the US Using Online Search Data

Dean Fantazzini*

Moscow School of Economics, Moscow State University, Moscow, Russia

Abstract

We propose the use of Google online search data for nowcasting and forecasting the number of food stamps recipients. We perform a large out-of-sample forecasting exercise with almost 3000 competing models with forecast horizons up to 2 years ahead, and we show that models including Google search data statistically outperform the competing models at all considered horizons. These results hold also with several robustness checks, considering alternative keywords, a falsification test, different out-of-samples, directional accuracy and forecasts at the state-level.

Citation: Fantazzini D (2014) Nowcasting and Forecasting the Monthly Food Stamps Data in the US Using Online Search Data. PLoS ONE 9(11): e111894. doi:10.1371/journal.pone.0111894

Editor: Tobias Preis, University of Warwick, United Kingdom

Received March 26, 2014; **Accepted** October 8, 2014; **Published** November 4, 2014

Copyright: © 2014 Dean Fantazzini. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. The author confirms that all data underlying the findings are fully available without restriction. All relevant data are within the Supporting Information files.

Funding: This author has no support or funding to report.

Competing Interests: The author has declared that no competing interests exist.

* Email: fantazzini@mse-msu.ru

Introduction

The Supplemental Nutrition Assistance Program (SNAP), which was known as the Food Stamp Program until it was renamed in the 2008 US farm bill, is a federal aid program designed to give low- and no-income people living in the US a means to buy food. Since 2011, more than 40 million Americans have received this kind of aid. The number of monthly food stamps recipients has become increasingly scrutinized worldwide as an important indicator of the US economy: see Figure 1 which reports the monthly (absolute) number of news related to food stamps in Bloomberg since 2000, and the monthly (standardized) number of news in Google since 2006 worldwide.

There are several reasons behind this phenomenon: one is the lack of trust in classical indicators like the GDP, particularly during the last global recession, due to subsequent downward GDP revisions. This has sparked a hot debate about the veracity of official data, forcing even an official declaration by Mark Doms, the Chief Economist of the US Department of Commerce, who said on the 26/11/2011 that "…*as many outside economists and GDP experts will attest to, the integrity of BEA [Bureau of Economic Analysis]'s data and its recent revisions to the latest U.S. recessionary period should not be suspect. But there is always room for improvement, and BEA and the Commerce Department continue to search for ways to improve its data collection and analysis to best serve the American people*"(see the full note by Mark Doms at http://www.esa.doc.gov/Blog/2011/08/26/no-smoke-and-mirrors-gdp-note-bea's-recent-revisions). Another reason is the criticism about the official unemployment rate: the official rate is the so-called U3 (i.e. people without jobs who have actively looked for work within the past four weeks) which can be quite restrictive and underestimate the real rate. Many analysts prefer to consider U6 (=U3 + "discouraged workers" + "marginally attached workers" + Part-time workers who want to work full-time, but cannot due to economic reasons), but even this last measure does not include long-term discouraged workers, which were excluded by the US Bureau of Labor Statistics in 1994. Finally, in 2008, Moody's Analytics found that food stamps were the most effective form of economic stimulus, increasing economic activity by $1.73 for every dollar spent (that is, the one-year fiscal multiplier effect). Unemployment insurance came in second, at $1.62, whereas most tax cuts yielded a dollar or less. The reason for this high effectiveness is the fact that "…*food stamps recipients are so poor that they tend to spend them immediately*", [1]. In 2011, US Secretary of Agriculture Tom Vilsack gave a higher estimate of $1.84, based on a 2002 USDA study.

Given this background, models for nowcasting (i.e. forecasting in real time, since the official release is published with a 2-month lag) can be very important for financial analysts and economists, since they do not have access to the initial estimates by the USDA, which are not released due to the high noise in the data. Moreover, models for forecasting can be very important for policy makers like the USDA when preparing public budgets: for example, it can be of great interest to know when an increase of the number of food stamps recipients will start abating. Similarly, economists and financial professionals worldwide can benefit from good forecasts, since the number of food stamps recipients is an important indicator of the US economy.

Unfortunately, food stamp caseloads are difficult to predict and the academic literature in this regard is very limited: the main paper dealing with food stamps forecasting is in fact the one by [2] for the USDA in 1991. Despite an extensive modelling effort, [2] concluded that their "[…] *model did not yield highly accurate forecasts of the Food Stamp caseload*", and that "*none of the […] models would have captured the increase in participation that began in 1989*". This is probably one of the reason why the (vast) literature since then mainly focused only on the determinants of

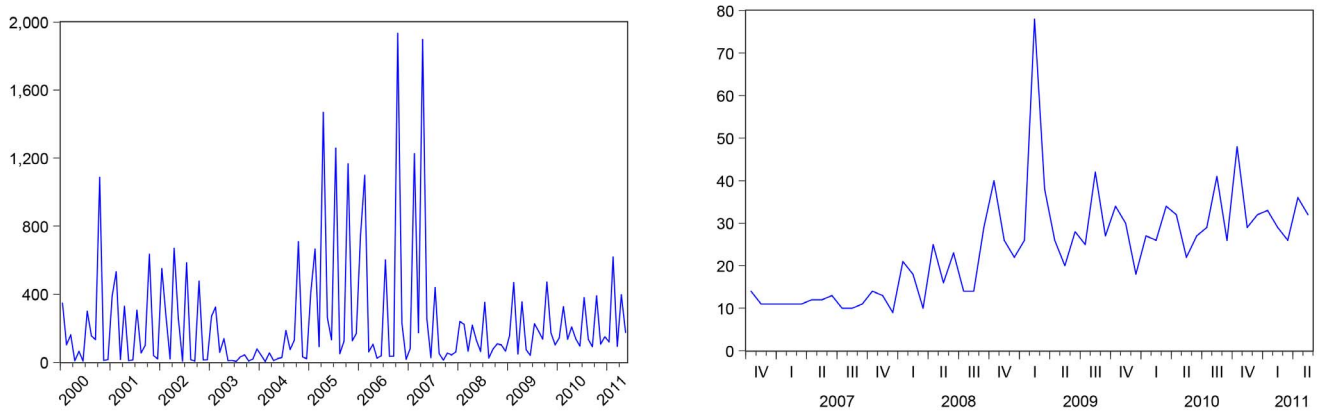


Figure 1. Bloomberg story-count for ''food stamps'' worldwide (left plot); Google standardized volume of news related to ''food stamps'' worldwide (right plot). Google data are registered trademarks of Google Inc., used with permission.
doi:10.1371/journal.pone.0111894.g001

welfare caseloads, analyzing the effects of SNAP policies, welfare policies, and the economy on SNAP participation rates and other characteristics, without dealing with forecasting: see the recent study by [3], the review by [4] and references therein for a discussion and an overview of this literature. A much smaller strand of the literature kept on dealing with welfare caseload forecasting, even though on a more limited scale than [2] –only at the state level– and not always specifically with the food stamps program: [5] discussed the forecasting of child abuse and neglect reports in urban, suburban, and rural counties; [6] dealt with the income assistance caseloads for the state of Washington; [7] developed a forecasting model for the Aid to Families with Dependent Children (AFDC) caseloads; [8] dealt with monthly state-level welfare caseloads in California; [9] provided a review of the literature about welfare caseloads and forecasting methods at the state level, showing an example with Georgia Temporary Assistance for Needy Families (TANF) caseloads, while [10] forecasted the number of participants in the special supplemental nutrition program for Women, Infants, and Children (WIC) using Vector Autoregression and ARIMA models. Differently from the previous literature, [11] is the first work to employ several methods to forecast Japanese welfare caseloads at the national level and to compare their performances.

Twenty years after [2], many interesting models have been developed: cointegration methods, nonlinear methods, periodic models. Even more interesting, now we have free access to Google online search data. Google holds the world leadership among all search engines with 82.8% market share (Net Applications, 2013) and it receives several hundred million queries each day: since January 2004, Google has started to make available the standardized number of the internet search queries for a keyword (or a group of keywords) with a tool called *Google Trends*. It provides information of users' relative interest for a particular search query at a given geographic region and at a given time. The Google Index (GI) for a specific query is standardized between 0 to 100%, where 100% is the peak of the search queries. The academic literature has started using Google search data for both forecasting and nowcasting purposes: [12] proposed Google Trends data for predicting various economic and financial indicators [13], used Google search data for forecasting the German unemployment rate [14], for the Italian unemployment rate [15], for the Israeli unemployment rate, while [16] for the US unemployment rate [17], [18] and [19]. estimated the 'influenza' activity in the US, China and South Korea, respectively, using online influenza-related internet queries [20]. used Google data to measure investors' attention for a sample of Russell 3000 stocks, while [21] used Google data to forecast the real price of oil. See [22] for a survey of this literature. Recently, the Google Trends literature has become much broader: [23] quantified the degree to which Internet users worldwide seek more information about years in the future than years in the past, and found a strong correlation between the country's GDP and the predisposition of its inhabitants to look forward [24]. analyzed changes in Google query volumes for search terms related to finance and found patterns that may be interpreted as ''early warning signs'' of stock market moves [25]. proposed a novel approach to portfolio diversification using Google Trends, which is based on the idea that the popularity of a stock measured by search queries is correlated with the stock riskiness, while [26] analyzed the dynamic relationship between the BitCoin price and the interest in the currency measured by search queries on Google Trends and frequency of visits on the Wikipedia page on BitCoin.

In this perspective, we propose to use online search data for nowcasting and forecasting the monthly number of food stamps recipients: we justify this choice because the administrative burden for enrolling and remaining enrolled in the food stamps program is nontrivial, see e.g. [27], [28] and [3], and searching the web for information is one of the main strategies a potential applicant can do: for example, the most searched query related to the food stamps program for the US in the years 2004-2011 as provided by Google on 16/01/2012 was ''*apply food stamps*''. Therefore, using Google online query statistics can provide real time information about the number of current and future food stamps recipients.

The first contribution of the paper is a detailed analysis of the main determinants of food stamps dynamics using the structural relationship identification methodology discussed by [29] and [30], which is a robust method of model selection in case of small samples. The second contribution of the paper is a large scale forecasting comparison with a set of almost 3000 models. In this regard, we computed nowcasts 1 step and 2 steps ahead, as well as out-of-sample forecasts up to 24 steps ahead, showing that models using Google data statistically outperform the competing models both for short term and long term forecasting. More specifically, we found that linear autoregressive models augmented with Google data definitively improve nowcasting food stamps data 2 months ahead, while simple linear models (eventually augmented with unemployment rates or initial claims data) are sufficient for nowcasting 1 month ahead. However, Google based linear models

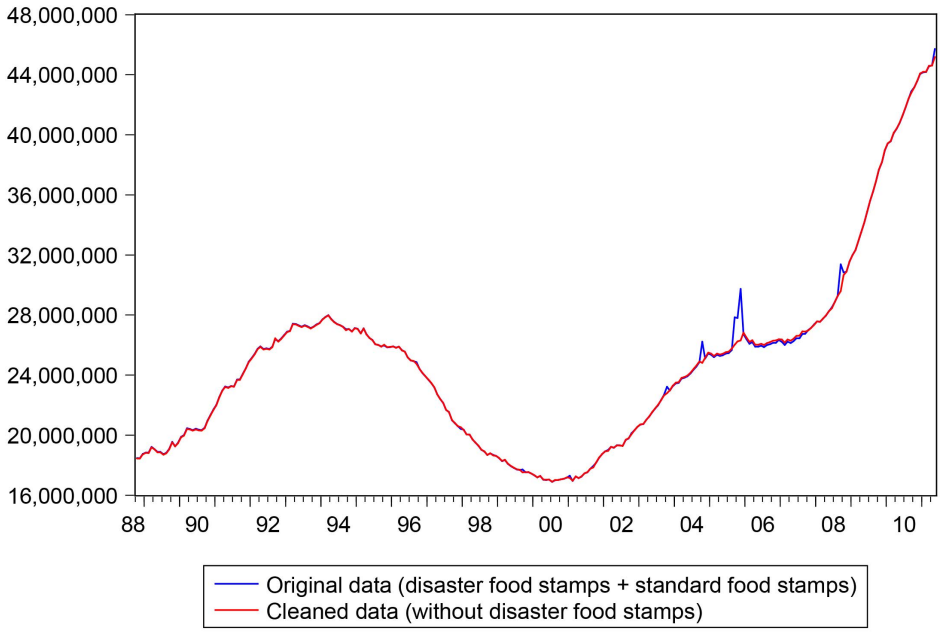


Figure 2. Original an cleaned food stamp data at the US national level. Sample: 1988M10 - 2011M5.
doi:10.1371/journal.pone.0111894.g002

provided superior forecasts in case of 12 steps and 24 steps forecast ahead, whereas most nonlinear models performed very poorly, were computationally intensive, and in several cases did not reach numerical convergence. In this regard, the best models had specifications always close to the ARX(4) model (Auto-Regressive model with eXogenous variables). which was found using the structural relationship identification methodology in the in-sample analysis. Our results hold also with alternative Google keywords and with alternative out-of-sample periods which either include the NBER recession of the years 2007–2009 or start after the end of this recession. Moreover, they passed a falsification test recently proposed by [16]. Similar results were found when considering the directional accuracy of the models' forecasts and when forecasting at the state-level. We remark that the out-of-sample forecasting comparison was structured to replicate the situation that real forecasters face when they compute their forecasts, so that all exogenous variables (for example Google data) have to be

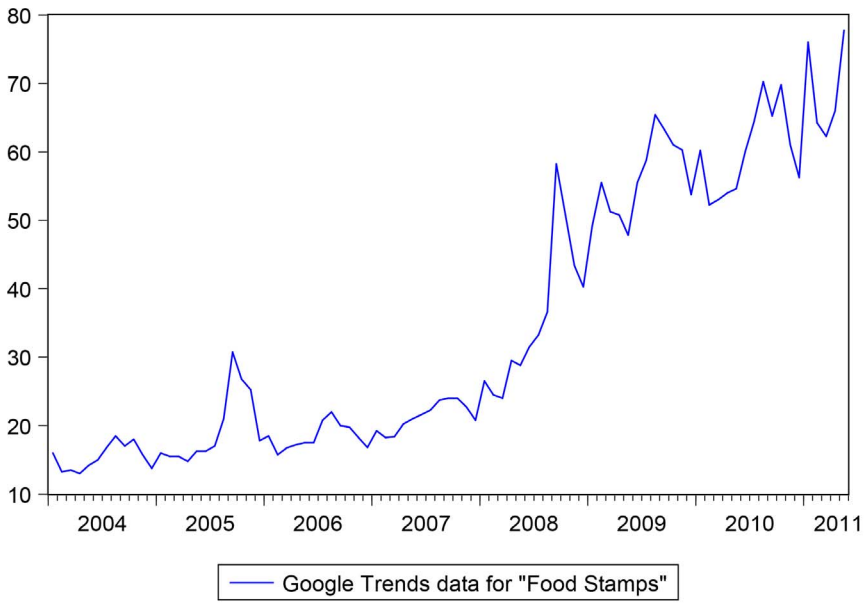


Figure 3. GI for the keywords "*food stamps*". Sample: 2004M1 - 2011M5. Google data are registered trademarks of Google Inc., used with permission.
doi:10.1371/journal.pone.0111894.g003

Table 1. Unit root tests.

	RUR	FB RUR	LS (1 break)	LS (2 breaks)
	(H_0: unit root)	(H_0: unit root)	(H_0: unit root)	(H_0: unit root)
Unemployment	1.34	1.85	−3.26 [1997M4]	−4.06[1996M4,2008M8]
Initial Claims	0.85*	1.07*	−3.94[2007M9]	−5.22[1992M11,200810]
GI Food Stamps	1.60	2.94*	−4.53*[2008M7]	−5.60*[2006M8,2008M8]
GI Jobs	1.17*	1.58	−5.52*[2007M4]	−6.48*[2007M3,2009M9]
Food Stamps	5.16*	7.86*	−3.10[1998M10]	−3.72[1992M5,1999M3]
	Periodic u.r. test - Sample: 1988–2011		**Periodic u.r. test – Sample: 2008–2011**	
	1nd step (LR stat.)	**2nd step (p-value)**	**1nd step (LR stat.)**	**2nd step (p-value)**
	H_0: periodic u.r.	H_0: non periodic u.r.	H_0: periodic u.r.	H_0: non periodic u.r.
Unemployment	0.02	**0.00**	31.55*	/
Initial Claims	15.05*	/	3.94*	/
GI Food Stamps	NA	NA	1.41	**0.01**
GI Jobs	NA	NA	6.46*	/
Food Stamps	20.20*	/	8.06*	/

Unit root tests: RUR = Range Unit Root test by [33]; FB = Forward-Backward RUR test by [33]; LS = unit root test with breaks by [32] - the estimated break dates are reported in brackets. The second step for the periodic unit root tests by [35] and [34] is performed only if the first step did not reject the null hypothesis of a periodic unit root. P-values smaller than 0.05 are in bold font. * Significant at the 5%, level.

doi:10.1371/journal.pone.0111894.t001

predicted to forecast the endogenous variables of interest (in our case the number of food stamps) and avoid any look-ahead bias.

Materials and Methods

Data and In-Sample Analysis

The monthly number of individuals enrolled in the Food Stamps/SNAP program were collected from the USDA, for the period from October 1988 till May 2011, both at the US national level and at the state level. Unfortunately, these data included not only the standard income-based food stamps but also the so called *disaster food stamps*, which "… *provide replacement benefits for regular food stamp recipients who lose food in a disaster and extends benefits to many households which would not ordinarily be eligible but suddenly need food assistance*" (see the full details at http://frac.org/federal-foodnutrition-programs/snapfood-stamps/disaster-snapfood-stamps/). Following an interesting discussion with the people working at USDA who provided us with the data, we proceeded to clean the original data from the disaster food stamps for two main reasons:

- The two food stamps programs have very different economic rationale: the disaster food stamps are usually a very short term phenomenon which follows from natural disasters (floods, tornados, and so on), while food stamps for income reasons are a much more persistent process;
- Disaster food stamps create spikes/jumps in the data which can hinder considerably the estimation of any econometric models.

The cleaning process was very long, since the disaster food stamps were not in standardized format and were reported in different data type, so that the correction was made month by month, state by state, for all states, and with these data we then reconstructed the total number of food stamps recipients at the US national level for all months considered. The original and cleaned datasets are reported in Figure 2.

The spike in disaster food stamps following the havoc caused by hurricane Katrina is clearly visible.

We then collected the GI for the keywords "*food stamps*" at the US national level for the period from January 2004 till May 2011. We remark that the GI is computed as the ratio of the search queries for a specific keyword (or group of keywords) relative to the total number of searches performed in the selected region at a given point of time, and then standardized between 0 and 100 (where the standardization is done over the whole time period). It is usually updated weekly, if not daily. The GI had a weekly frequency but was transformed into a monthly frequency to match food stamps data, see Figure 3.

Among the set of variables that we used to forecast food stamps, we also considered the unemployment rate. Its monthly data are available from January 1976 and can be downloaded from the U.S. Bureau of Labor Statistics, both at the US national and state level. This is one of the most frequently used variables used to model food stamps in the US, and was found able to explain a large part of the variation in food stamps dynamics, see e.g. [31], [3] and references therein. Moreover, in the fewer cases when forecasting was of concern, like [2], [10] and [11], the unemployment rate was one of the variables with the highest forecasting power [16]. found that models augmented with the GI for the keyword "jobs" significantly outperformed a very wide range of competing models when forecasting the US unemployment rate. Given this evidence and considering that the unemployment rate is one of the major predictor of the number of food stamps recipients, we also included this GI in our set of

Table 2. P-values of sequential tests for weak exogeneity.

	Wald test		Toda-Yamamoto	
	1st step	2nd step	1st step	2nd step
Unemployment	**0.01**	**0.00**	0.05	/
Initial Claims	**0.00**	**0.00**	**0.00**	0.07
GI Food Stamps	0.58	/	0.79	/
GI Jobs	**0.01**	**0.00**	0.26	/
Food Stamps	**0.00**	**0.00**	**0.00**	**0.00**

P-values of sequential tests for weak exogeneity: standard Wald test and Wald test using the approach by [37]. P-values smaller than 0.05 are in bold font.
doi:10.1371/journal.pone.0111894.t002

predictors. Finally, the monthly Initial Claims (IC) were also considered: they are available from January 1971 and can be downloaded from the US Department of Labor, both at the US national and state level. We employed this time series because it is a widely accepted leading indicator for the US unemployment rate, see [16] and references therein.

To analyze the food stamps dynamics, we employed the structural relationship identification methodology discussed by [29] and [30], which is specifically designed for small samples. The first step is to identify the order of integration using unit root tests: if all variables are stationary, VAR and VARX (Vector Autoregressive with exogenous variables) models are used. The second step determines the exogeneity of each variable using the sequential reduction method for weak exogeneity by [30], who consider weakly exogenous each variable for which the test is not rejected and re-test the remaining variables until all weakly exogenous variables are identified. For non-stationary variables, cointegration rank tests are employed to determine the presence of a long-run relationship among the endogenous variables: if this is the case, VECM or VECMX (Vector Error Correction model with exogenous variables) models are used, otherwise VAR or VARX models in differences are applied. The last step is to compute out-of-sample forecasts, see [29] for more details. However, our approach differs from the latter in that we employ unit root tests and cointegration tests allowing for structural breaks.

Stationarity. We analyzed the stationarity of food stamps data using a set of unit root tests allowing for structural break(s) because the food stamps legislation underwent a series of reforms during its history: the 1993 Mickey Leland Childhood Hunger Relief Act, the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA), the 1996 Farm Bill, and the 2008 Farm Bill, just to name the most important, see [3] for an overview. Moreover, a global recession hit worldwide in 2007–2009, reaching the apex with the bankruptcy of Lehman Brothers. More specifically, we employed five unit root tests: the [32] unit root tests allowing for one and two breaks, respectively, and the Range Unit Root test (RUR) and the Forward-Backward RUR test suggested by [33], which are non-parametric tests robust against nonlinearities, error distributions, structural breaks and outliers. Furthermore, we also employ a periodic unit root test, given a mild presence of periodicity in the US food stamps data: if we perform a simple regression of the log returns of the monthly food stamps on 12 seasonal dummies over the full time sample, four seasonal dummies are statistically significant at the 1% level (using HAC robust standard errors) and the adjusted R^2 is 12%. This is a first-stage regression that was suggested by [34] to verify the potential presence of periodicity in the data: the mild value of the R^2 highlight the need to take periodicity into account. To test the null hypothesis of a periodic unit root, we follow the two-step strategy suggested by [35] and [34]: in the first step, a likelihood ratio test for testing a single unit root in a Periodic Auto-Regressive (PAR) model of order p is performed (the order p is chosen by using the Schwartz information criterion and checking that the residuals are approximately white noise); if the null of a periodic unit root cannot be rejected [35] and [34], suggest to test in a second step whether the process contain a non periodic unit root equal to 1 for all seasons. Since there is no version of this test with endogenous breaks, we estimated it both with the full sample starting in 1988, and with a smaller sample starting in 2008 to take the global financial crisis into account. As for the GIs, we want to remark that even though they are bounded between 0 and 100, this does not imply that they are stationary: for example, a random walk divided by its maximum value and multiplied by 100 (i.e. the procedure for computing the GI) remains non-stationary. Besides, the statistical effects of dividing the original search data for a specific keyword (which can be non stationary) by the total number of web searches in the same week and same area (which can be non stationary as well) are unknown, see also [21] for a discussion. The results of these tests for the log-transformed data of all variables are reported in Table 1 (the results for data in levels are similar and are not reported for sake of space - the software used to compute these tests is discussed in Software Description S1).

The evidence emerging from the (non-periodic) unit root tests is somewhat mixed but points to stationarity for almost all time series, with structural breaks at the end of the '90s and at the beginning of the global financial crisis in 2007–2008. This evidence is also indirectly confirmed by the periodic unit root tests, whose outcomes changes substantially if the sample used changes, particularly for the unemployment rate. The latter data is probably the one which has the more mixed evidence: in this regard, we are aware of the very hot discussion about the stationarity of unemployment rates and we refer to [36] and [16] for a review of this debate. Given this evidence, we decided to follow a neutral approach and in the forecasting section we compared both models with the data in levels and models with first-differenced data.

Weak exogeneity and Cointegration. The next step in the structural relationship identification methodology suggested by [29] is to determine the exogeneity of each variable using the sequential reduction method for weak exogeneity proposed by [30], which is specifically designed for small samples: once a weakly exogenous variable is found, the remaining variables are re-tested until all weakly exogenous variables are identified. Given the previous mixed evidence of stationarity, we employed both the

Table 3. Single-equation and multivariate cointegration tests with and without structural break(s).

Single-equation cointegration tests				
Engle and Granger (1987)	**Gregory and Hansen (1996)**		**Hatemi (2008)**	
No breaks	**1 (endogenous) break**		**up to 2 (endogenous) breaks**	
Tau statistic	*Z-t statistic*	*Break date*	*Z-t statistic*	*Break dates*
−3.83	−4.82	2009M11	−5.29	2006M1 2009M1
Multivariate cointegration tests				
Johansen (1995)	**Johansen et al.(2000)**		**Johansen et al. (2000)**	
No breaks	**1 (exogenous) break**		**2 (exogenous) breaks**	
N. of CEs	*N. of CEs*	*Break date from*	*N. of CEs*	*Break dates from*
at 5% level	*at 5% level*	*G.H. (1996)*	*at 5% level*	*H. (2008)*
4 CEs	5 CEs −> stationarity	2009M11	5 CEs −> stationarity	2006M1 2009M1

Single-equation and multivariate cointegration tests with and without structural break(s). The null hypothesis for all tests is the absence of cointegration. All the tests considered the case of no deterministic trend in the data and an intercept in the cointegration equation (CE), centered seasonal dummies outside the CE, while the number of lags is chosen using the Schwartz criterion. The tests allowing for break(s) considered the case of a level shift. * Significant at the 5% level.
doi:10.1371/journal.pone.0111894.t003

standard Wald test using a VAR model in levels with centered seasonal dummies, and the Wald test proposed by [37] which is valid in case the variables may be integrated or cointegrated of an arbitrary order (we included centered seasonal dummies because they sum to zero over time and therefore do not affect the asymptotic distributions of the tests, see [38] and [39] for details). This last approach requires, first, to determine the appropriate maximum lag length k for the variables in the VAR in levels using information criteria; then, to estimate a $(k+d_{max})$ th-order VAR where d_{max} is the maximum order of integration that we suspect for our group of time-series. Finally, [37] show that we can test linear or nonlinear restrictions on the first k coefficient matrices using standard asymptotic theory, while the coefficient matrices of the last $k+d_{max}$ lagged vectors have to be ignored. We chose $k=7$ after looking at a battery of information criteria (AIC, BIC, Hannan-Quinn, Forecast Prediction Error) and checking that the residuals behave approximately as a multivariate white noise. Moreover, in our case $d_{max}=1$. The results of the sequential reduction method for weak exogeneity using the standard Wald test with a VAR(7) and the Wald test proposed by [37] with a VAR(8) are reported in Table 2. Variables whose Wald test has a p-value larger than 5% are considered weakly exogenous and are excluded from further testing.

The results of the two approaches differ considerably: for the standard Wald test, only the GI for the keyword "food stamps" is weakly exogenous, while for the Toda and Yamamoto approach all four predictors are weakly exogenous (the unemployment rate and the two GIs in the first step, while the initial claims in the second step). It may well be the case that the global financial crisis in 2008, which was a significant break in the previous unit root tests, could be one of the main reasons of these different results. Unfortunately, our sample is too short to estimate VAR(7) and VAR(8) models starting in 2008.

Given the somewhat mixed evidence about stationarity, we proceeded nonetheless to test for cointegration among our five variables as a potential cross-check: if the variables are all stationary, the multivariate cointegration tests should find a number of cointegration relationships equal to the number of the variables examined. In this regard, the Johansen cointegration tests can be used as panel unit root tests, as discussed by [40] and [41]. More specifically, we used a set of cointegration tests allowing for the presence of structural break(s):

- [42] single-equation cointegration test allowing for one endogenous break;
- [43] single-equation cointegration test allowing for two endogenous breaks;
- [44] multivariate test allowing for the presence of one or two exogenous break(s), where the dates of the breaks are the ones selected by the [42] and [43] tests, respectively.

For sake of generality, we also considered the single-equation test by [45] and multivariate cointegration test by [38], both of them without breaks. The main advantage of single-equation approaches is that they allow for endogenous breaks. However, these tests are not suitable when the right hand variables in the cointegration vector are not weakly exogenous (which is not our case, according to the approach by [37]) and when there are more than one cointegrating vector. The only problem with the multivariate tests by [44] is that they allow only for exogenous breaks. Therefore, we followed a 2-step strategy: we first estimated the single-equation approaches by [42] and [43] to have an indication of the structural breaks dates, and we then used these dates to compute the multivariate tests by [44], see Table 3.

All single-equation tests do not reject the null of no cointegration, while the Johansen tests allowing for break(s) found evidence of five CEs in a system of five variables, which means that all the five variables are stationary. Only the Johansen test with no breaks found evidence of a cointegrated system with 4 CEs, but the presence of a break during the global financial crisis suggests some caution when interpreting this last result. Therefore, this evidence of absence of cointegration and stationary variables is consistent with the previous weak exogeneity tests and unit root tests. We remark that periodic cointegration tests using all variables could not be implemented due to the high number of parameters to be estimated. This "curse of dimensionality" is a well known problem for this kind of tests, see [34] for more details.

Finally, the values of the significant parameters at the 5% level for the equation of the monthly number of food stamps recipients in log-levels are reported in Table 4. A battery of misspecification and stability tests is also reported in the same table.

Table 4. Estimated coefficients in the equation of food stamps recipients (left block) and misspecification and stability tests (right block).

Regressors	*Coeff.*	*T-stat*	*Tests*	*p-value*
log(Food stamps(-1))	0.59	5.40	Ljung-Box(12)	0.52
log(Food stamps(-2))	0.30	2.31	Ljung-Box(24)	0.65
log(Food stamps(-3))	0.29	2.22	Ljung-Box(12) res. sq.	0.79
log(Food stamps(-4))	-0.23	-2.25	Ljung-Box(24) res. sq.	0.79
log(Unemployment rate)	0.02	3.13	ARCH(12)	0.89
log(GI - Food Stamps)	0.01	3.96	ARCH(24)	0.98
log(GI - Jobs)	0.02	2.03	Jarque-Bera	**0.00**
constant	0.87	4.63	RESET	0.56
S1	−0.02	−5.74	BDS (dim = 2)	0.12
S2	−0.02	−8.07	BDS (dim = 6)	**0.00**
S3	−0.01	−4.37	OLS-CUSUM	0.99
S4	−0.02	−4.44	Rec-CUSUM	0.06
S5	−0.01	−3.43	OLS-MOSUM	0.51
S6	−0.02	−4.04	Rec-MOSUM	0.39
S7	−0.01	−3.95	Andrews max-F	**0.03**
S8	−0.01	−3.89	Andrews exp-F	0.22
S9	−0.01	−4.64	Andrews ave-F	0.09
S10	−0.01	−3.69	Optimal n. breakpoints (BIC)	0
S11	−0.01	−4.36	Optimal n. breakpoints (LWZ)	0

Estimated coefficients in the equation of food stamps recipients (left block) and misspecification and stability tests (right block). Sample: 2004M1- 2011M05. P-values smaller than 0.05 are in bold font.

Misspecification tests: the [46] statistics for testing the absence of autocorrelation up to order k in the models' residuals and residuals squared; the Lagrange multiplier test for Auto-Regressive Conditional Heteroskedasticity (ARCH) in the residuals proposed by [47]; the [48] test for checking whether a time series is normally distributed; the REgression Specification Error Test (RESET) proposed by [49], which is a general test for incorrect functional form, omitted variables, and correlation between the regressors and the error term; the BDS test by [50] to test whether the residuals are independent and identically distributed (iid) and which is robust against a variety of possible deviations from independence, including linear dependence, non-linear dependence, or chaos.

Stability tests: the test for parameter instability by [51] which is based on the CUmulative SUM of the recursive residuals (Rec-CUSUM); [52] suggested to modify the previous structural change test and use the cumulative sums of the common OLS residuals (OLS-CUSUM). [53] proposed a structural change test which analyzes moving sums of residuals (MOSUM) instead of cumulative sums. We remark that a unifying view of the previous structural change tests within a generalized M-fluctuation test framework was proposed by [54] and [55]. [56] was the first to suggest an F-test for structural change when the break point is known: [57] and [58] extended the Chow test by computing the F statistics for all potential break points and suggested three different test statistics, the sup-F, the ave-F and the exp-F, which are based on Wald, Lagrange Multiplier or Likelihood Ratio statistics respectively, in a very general class of models fitted by Generalized Method of Moments. See [59] for a review and a step-by-step description of stability tests using R software. Besides, [60], following [61], suggested to find the optimal number of breakpoints by optimizing the Bayesian Information Criterion (BIC) and the modified BIC by [62] (LWZ, 1997).

doi:10.1371/journal.pone.0111894.t004

The GIs for the keywords "food stamps" and "jobs" and the unemployment rate have all a positive effect on the number of food stamps recipients: an increase in these variables, increase the the number of food stamps. Instead, the number of initial claims was found not significant at the 5% level and therefore was not reported in Table 4. The sum of the autoregressive coefficients is 0.94, somewhat close to 1, thus confirming the mixed evidence about stationarity which emerged from unit root tests in Table 1. As previously highlighted by the stationarity and cointegration tests, this may be due to a break in 2008–2009. However, the parameter stability tests do not signal strong evidence of model instability, and similarly the misspecification tests do not show any serious problem in the model's residuals, except for some nonlinearity and the lack of residuals' normality. The latter issue suggests caution when reading the previous t-statistics, considering that our sample consists of 89 observations: one possibility could be to resort to bootstrap methods or to use robust estimation methods. Given that the focus of this work is forecasting, we preferred to deal with this issue by comparing the forecasting performances of a very large set of model specifications: with different number of autoregressive lags, with and without seasonal dummies, with and without Google indexes, with and without weakly exogenous regressors, with data in levels and in first differences. Such an approach allows us to take a neutral stance towards the competing models and avoid any form of look-ahead bias. Moreover, we could analyze the models' behavior during the potential structural break caused by the global financial crisis.

Forecasting Models

Food stamps data are reported with a 2-month lag: the latest data relative to month $t-2$ are issued in the first working days of month $t+1$. For example, in the first days of January 2011, the data about October 2010 were released. As discussed before, the data are released with a 2-month delay due to the high noise in the initial data. Therefore, in order to "*nowcast*" the value of food stamps for November 2010 (i.e. month $t-1$) and December 2010 (i.e. month t), we can use the Google data up to December 2010 (i.e. month t), the initial claims up to December 2010 (i.e. month t), and the unemployment rate up to November 2010 (i.e. month $t-1$), since it is released with 1-month lag. Besides nowcasting, we also consider forecasting monthly food stamps 12 months ahead

Table 5. Models used for nowcasting and forecasting: Linear and Periodic models.

	LAGS	Additional regressors													Time Samples		Data Transformations				Total models
		NO REG.	*GI(FS)*		*GI(JOBS)*		*GIs(JOBS & FS)*		*UR*		*IC*		*GI(FS), UR, IC*	*G(JOBS), UR, IC*	*GI(JOBS & FS), UR, IC*	1988M 10-2011M5	2004M 1-2011M5	Y_t	$\log(Y_t)$	$\Delta(Y_t)$	$\Delta\log(Y_t)$	(sum by row)
Regressors:			*sim.*	*sim. and*	*sim.*	*sim. and*	*sim.*	*sim. and*	*lag1*	*lagged*	*sim.*	*sim. and*	*sim. +*	*sim. +*	*sim. +*							
sim./ lagged				*lagged*		*Lagged*		*lagged*				*lagged*	*lag1*	*lag1*	*lag1*							
AR(p)	up to 12	yes														yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
	up to 12		yes														yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12			yes													yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12				yes												yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12					yes											yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12						yes										yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12							yes									yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12								yes							yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
	up to 12									yes						yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
	up to 12										yes					yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
	up to 12											yes				yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
	up to 12												yes				yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12													yes			yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12														yes		yes	yes	yes	yes	yes	12 × 4 × 1 = 48
ARMA (p, q)	up to 12	yes														yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
	up to 12		yes														yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12				yes												yes	yes	yes	yes	yes	12 × 4 × 1 = 48

Table 5. Cont.

	LAGS	Additional regressors														Time Samples		Data Transformations				Total models
		NO REG.	*GI(FS)*		*GI(JOBS)*		*GIs(JOBS & FS)*		*UR*		*IC*		*GI(FS), UR, IC*	*G(JOBS), UR, IC*	*GI(JOBS & FS), UR, IC*	1988M 10-2011M5	2004M 1-2011M5	Y_t	$\log(Y_t)$	$\Delta(Y_t)$	$\Delta\log(Y_t)$	(sum by row)
Regressors:			*sim.*	*sim. and*	*sim.*	*sim. and*	*sim.*	*sim. and*	*lag1*	*lagged*	*sim.*	*sim. and*	*sim. +*	*sim. +*	*sim. +*							
sim./ lagged				*lagged*		*Lagged*		*lagged*				*lagged*	*lag1*	*lag1*	*lag1*							
	up to 12						yes										yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12								yes							yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
	up to 12										yes					yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
	up to 12												yes				yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12													yes			yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12														yes		yes	yes	yes	yes	yes	12 × 4 × 1 = 48
AR(p)	up to 12	yes														yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
+ seasonal	up to 12		yes														yes	yes	yes	yes	yes	12 × 4 × 1 = 48
constants	up to 12				yes												yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12						yes										yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12								yes							yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
	up to 12										yes					yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
	up to 12												yes				yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12													yes			yes	yes	yes	yes	yes	12 × 4 × 1 = 48
	up to 12														yes		yes	yes	yes	yes	yes	12 × 4 × 1 = 48
PAR(p)	up to 3	yes														yes	yes	yes	yes	yes	yes	3 × 4 × 2 = 24
	up to 3		yes														yes	yes	yes	yes	yes	3 × 4 × 1 = 12

Table 5. Cont.

	LAGS	Additional regressors														Time Samples		Data Transformations				Total models
		NO REG.	*GI(FS)*		*GI(JOBS)*		*GIs(JOBS & FS)*		*UR*		*IC*		*GI(FS), UR, IC*	*G(JOBS), UR, IC*	*GI(JOBS & FS), UR, IC*	1988M 10-2011M5	2004M 1-2011M5	Y_t	$\log(Y_t)$	$\Delta(Y_t)$	$\Delta\log(Y_t)$	(sum by row)
Regressors:			*sim.*	*sim. and*	*sim.*	*sim. and*	*sim.*	*sim. and*	*lag1*	*lagged*	*sim.*	*sim. and*	*sim. +*	*sim. +*	*sim. +*							
sim./ lagged				*lagged*		*Lagged*		*lagged*				*lagged*	*lag1*	*lag1*	*lag1*							
	up to 3			yes													yes	yes	yes	yes	yes	3 × 4 × 1 = 12
	up to 3				yes												yes	yes	yes	yes	yes	3 × 4 × 1 = 12
	up to 3					yes											yes	yes	yes	yes	yes	3 × 4 × 1 = 12
	up to 3						yes										yes	yes	yes	yes	yes	3 × 4 × 1 = 12
	up to 3							yes									yes	yes	yes	yes	yes	3 × 4 × 1 = 12
	up to 1								yes							yes	yes	yes	yes	yes	yes	1 × 4 × 2 = 8
	up to 3									yes						yes	yes	yes	yes	yes	yes	3 × 4 × 2 = 24
	up to 1										yes					yes	yes	yes	yes	yes	yes	1 × 4 × 2 = 8
	up to 3											yes				yes	yes	yes	yes	yes	yes	3 × 4 × 2 = 24
	up to 1												yes				yes	yes	yes	yes	yes	1 × 4 × 1 = 4
	up to 1													yes			yes	yes	yes	yes	yes	1 × 4 × 1 = 4
	up to 1														yes		yes	yes	yes	yes	yes	1 × 4 × 1 = 4
PAR(p)	up to 3	yes														yes	yes	yes	yes	yes	yes	3 × 4 × 2 = 24
+ periodic	up to 3		yes														yes	yes	yes	yes	yes	3 × 4 × 1 = 12
trends	up to 3				yes												yes	yes	yes	yes	yes	3 × 4 × 1 = 12
	up to 3						yes										yes	yes	yes	yes	yes	3 × 4 × 1 = 12
	up to 1								yes							yes	yes	yes	yes	yes	yes	1 × 4 × 2 = 8

Table 5. Cont.

	LAGS	Additional regressors													Time Samples		Data Transformations				Total models
		NO REG.	*GI(FS)*		*GI(JOBS)*		*GIs(JOBS & FS)*		*UR*		*IC*		*GI(FS), UR, IC*	*G(JOBS), UR, IC*	*GI(JOBS & FS), UR, IC*	1988M 10-2011M5	2004M 1-2011M5	Y_t	$\log(Y_t)$	$\Delta(Y_t)$	$\Delta\log(Y_t)$	(sum by row)
Regressors:			*sim.*	*sim. and*	*sim.*	*sim. and*	*sim.*	*sim. and*	*lag1*	*lagged*	*sim.*	*sim. and*	*sim. +*	*sim. +*	*sim. +*							
sim./ lagged				*lagged*		*Lagged*		*lagged*				*lagged*	*lag1*	*lag1*	*lag1*							
	up to 1										yes					yes	yes	yes	yes	yes	yes	1 × 4 × 2 = 8
	up to 1												yes				yes	yes	yes	yes	yes	1 × 4 × 1 = 4
	up to 1													yes			yes	yes	yes	yes	yes	1 × 4 × 1 = 4
	up to 1														yes		yes	yes	yes	yes	yes	1 × 4 × 1 = 4
PAR(p)	up to 3	yes														yes	yes	yes	yes	yes	yes	3 × 4 × 2 = 24
+ ARCH(1)	up to 3			yes													yes	yes	yes	yes	yes	3 × 4 × 1 = 12
	up to 3									yes						yes	yes	yes	yes	yes	yes	3 × 4 × 2 = 24
	up to 3											yes				yes	yes	yes	yes	yes	yes	3 × 4 × 2 = 24
PEC	1,12		yes														yes	yes	yes			1 × 2 × 1 = 2
PEC	1,12								yes							yes	yes	yes	yes			1 × 2 × 2 = 4
PEC	1,12										yes					yes	yes	yes	yes			1 × 2 × 2 = 4
PEC	1,12								yes		yes					yes	yes	yes	yes			1 × 2 × 2 = 4
PEC	1,12						yes										yes	yes	yes			1 × 2 × 1 = 2

doi:10.1371/journal.pone.0111894.t005

Table 6. Models used for nowcasting and forecasting: Multivariate models, Nonlinear models and Random Walk with drift.

		Additional regressors																	Time Samples		Data Transformations				Row Total
MODELS	**LAGS**	**NO REG.**	*GI (JOBS & FS)*	*UR*	*IC*	*UR&IC*	*GI (FS)*	*GI (FS), UR*	*GI (FS), IC*	*GI (FS), UR, IC*	*GI (JOBS)*	*GI (JOBS), UR*	*GI (JOBS), IC*	*GI (JOBS), UR, IC*	*GI (JOBS & FS)*	*GI (JOBS & FS), UR*	*GI (JOBS & FS), IC*	*GI (JOBS & FS), UR, IC*	**1988M10-2011M5**	**2004M1-2011M5**	Y_t	**log** (Y_t)	Δ (Y_t)	$\Delta\log$ (Y_t)	**models**
VAR	1–7		yes																	yes	yes	yes			1 × 2 × 1 = 2
VAR	1–7					yes													yes	yes	yes	yes			2 × 2 × 1 = 4
VAR	1–6		yes																	yes			yes	yes	1 × 2 × 1 = 2
VAR	1–6					yes													yes	yes			yes	yes	2 × 2 × 1 = 4
BVAR	1–7						yes													yes	yes	yes			1 × 2 × 1 = 2
BVAR	1–6						yes													yes			yes	yes	1 × 2 × 1 = 2
BVAR	1,12						yes													yes	yes	yes	yes	yes	1 × 4 × 1 = 4
BVAR	1–7			yes	yes														yes	yes	yes	yes			2 × 2 × 1 = 4
BVAR	1–6			yes	yes														yes	yes			yes	yes	2 × 2 × 1 = 4
BVAR	1,12			yes	yes														yes	yes	yes	yes	yes	yes	2 × 4 × 1 = 8
BVAR	1–7		yes																	yes	yes	yes			1 × 2 × 1 = 2
BVAR	1–6		yes																	yes			yes	yes	1 × 2 × 1 = 2
BVAR	1,12		yes																	yes	yes	yes	yes	yes	1 × 4 × 1 = 4
VEC	1,12			yes															yes	yes	yes	yes			2 × 2 × 1 = 4
VEC	1,12				yes														yes	yes	yes	yes			2 × 2 × 1 = 4
VEC	1,12					yes													yes	yes	yes	yes			2 × 2 × 1 = 4
VEC	1,12						yes													yes	yes	yes			1 × 2 × 1 = 2
VEC	1,12							yes												yes	yes	yes			1 × 2 × 1 = 2
VEC	1,12								yes											yes	yes	yes			1 × 2 × 1 = 2

Table 6. Cont.

		Additional regressors																	Time Samples		Data Transformations				Row Total
MODELS	**LAGS**	**NO REG.**	*GI (JOBS & FS)*	*UR*	*IC*	*UR&IC*	*GI (FS)*	*GI (FS), UR*	*GI (FS), IC*	*GI (FS), UR, IC*	*GI (JOBS)*	*GI (JOBS), UR*	*GI (JOBS), IC*	*GI (JOBS), UR, IC*	*GI (JOBS & FS)*	*GI (JOBS & FS), UR*	*GI (JOBS & FS), IC*	*GI (JOBS & FS), UR, IC*	**1988M10-2011M5**	**2004M1-2011M5**	Y_t	**log** (Y_t)	Δ (Y_t)	$\Delta\log$ (Y_t)	**models**
VEC	1,12									yes										yes	yes	yes			1 × 2 × 1 = 2
VEC	1,12										yes									yes	yes	yes			1 × 2 × 1 = 2
VEC	1,12											yes								yes	yes	yes			1 × 2 × 1 = 2
VEC	1,12												yes							yes	yes	yes			1 × 2 × 1 = 2
VEC	1,12													yes						yes	yes	yes			1 × 2 × 1 = 2
VEC	1,12														yes					yes	yes	yes			1 × 2 × 1 = 2
VEC	1,12															yes				yes	yes	yes			1 × 2 × 1 = 2
VEC	1,12																yes			yes	yes	yes			1 × 2 × 1 = 2
VEC	1,12																	yes		yes	yes	yes			1 × 2 × 1 = 2
SETAR	up to 12	yes																	yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
LSTAR	up to 12	yes																	yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
AAR	up to 12	yes																	yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
NNET	up to 12	yes																	yes	yes	yes	yes	yes	yes	12 × 4 × 2 = 96
RW		yes																	yes	yes				yes	1 × 2 × 1 = 2
GRAND TOTAL: 2890 MODELS																									

doi:10.1371/journal.pone.0111894.t006

Table 7. Number of models with Google data out of the Top 100 models according to the RMSE.

	Nowcasting 1 s.a.	Nowcasting 2 s.a.	Forecasting 12 s.a.	Forecasting 24 s.a.
RMSE	41	90	92	91

doi:10.1371/journal.pone.0111894.t007

and 24 months ahead, given its importance for policy makers and public planners when preparing public budgets.

The regressors used to explain the dynamics of the monthly food stamps are the aforementioned monthly Google indexes for the keyword "food stamps" and "jobs" [in the following tables, GI(J.&F.S.) will represent the case where both the GIs for "food stamps" and "jobs" are present as regressors, GI(F.S.) the case with only the GI for "food stamps", whereas GI(J.) the case with only the GI for "jobs"], the monthly unemployment rates (UR) and the monthly initial claims (IC): these regressors may enter the equation simultaneously as weakly exogenous variables (in case of GIs and IC), with 1 lag (in case of the UR, since it is released with 1-month lag), simultaneously and with lags up to order p (in case of GIs and IC) and with lags up to order p (in case of the UR).

Models without Google data were estimated on two different time samples (1988M10-2011M5, 2004M1-2011M5) to consider the effects of potential structural breaks. In this regard [63], showed that in a regression with a single break, the optimal window for estimation includes all of the observations after the break, plus a limited number of observations before the break, and similar results also hold for multiple breaks (in this case the last break has to be considered): since the evidence so far points to a break at the end of the '90s and at the beginning of the global financial crisis, using a second estimation sample starting from 2004 should be a good compromise between efficiency and bias. Moreover, we also considered four possible data transformation: the original data in levels, the log-transformed data, the first differences and the first differences in logs (i.e. the log-returns). This was done to consider both stationarity and non stationarity, as well as simple nonlinearity captured by the log transformation.

The wide range of models we considered for nowcasting and forecasting can be grouped into five general classes:

Table 8. Ranking of the best models within each class according to the RMSE.

	Type of models	Nowcasting 1 s.a.	Nowcasting 2 s.a.	Forecasting 12 s.a.	Forecasting 24 s.a.
Linear	AR w/GI	17	81	127	177
models	AR w/o GI	5	75	75	128
	ARMA w/GI	138	51	74	153
	ARMA w/o GI	1	113	87	123
	AR + s.d. w/GI	17	1	1	1
	AR + s.d. w/o GI	14	38	180	180
Periodic	PAR w/GI	2530	2364	17	41
models	PAR w/o GI	444	948	690	822
	PAR+p.t. w/GI	2632	2623	959	145
	PAR+p.t. w/o GI	391	613	377	463
	PAR-ARCH w/GI	2635	2514	555	159
	PAR-ARCH w/o GI	1138	1459	610	836
	PEC w/GI	2538	2547	53	44
	PEC w/o GI	1783	2442	72	703
Multivariate	VAR w/GI	236	441	2053	2462
models	VAR w/o GI	293	345	61	229
	VEC w/GI	102	194	856	1518
	VEC w/o GI	209	367	257	627
	BVAR w/GI	7	370	515	411
	BVAR w/o GI	197	907	925	1301
Nonlinear	SETAR	Not converged	Not converged	Not converged	Not converged
models	LSTAR	716	1144	410	137
	NNET	1359	1595	923	797
	AAR	383	704	82	40
Random W.	RW	2562	2585	1847	1183

doi:10.1371/journal.pone.0111894.t008

Table 9. Top 10 models in terms of RMSE - baseline case. Nowcasting: 1 step and 2 steps ahead.

1 STEP ahead (baseline case)		2 STEPS ahead (baseline case)	
Top 10 models	**RMSE**	**Top 10 models**	**RMSE**
ARMA(10,10) dlog 1988	159024	AR(3)+S.D.+UR+IC+GI(J.&F.S.) log 2004	211508
ARMA(10,10) + UR dlog 1988	160819	AR(8)+S.D.+GI(F.S.) lev 2004	211784
ARMA(12,12) dlog 1988	161311	AR(7)+ S.D.+GI(F.S.) lev 2004	211843
ARMA(11,11) + UR diff 1988	162494	AR(2)+S.D.+UR+IC+GI(J.&F.S.) log 2004	212644
AR(12) + IC(sim+lags) dlog 1988	164194	AR(7)+S.D.+GI(J.&F.S.) lev 2004	212878
ARMA(12,12) + IC diff 1988	165172	AR(4)+ S.D.+GI(F.S.) lev 2004	214086
BVAR(1,12) FS+GI(F.S.) lev 2004	165369	AR(4)+S.D.+GI(J.&F.S.) lev 2004	215379
BVAR(1,12) FS+UR+IC+GI(F.S.) lev 2004	165531	AR(8)+S.D.+GI(J.&F.S.) lev 2004	215468
ARMA(12,12) + UR dlog 1988	166215	AR(5)+ S.D.+GI(F.S.) lev 2004	216076
ARMA(11,11) dlog 1988	167503	AR(6)+S.D.+UR+IC+GI(J.&F.S.) log 2004	216667

In each row, the following information is reported: the model, the number of lags, (eventual) exogenous regressors, the data transformation, the first year of the estimation sample.

doi:10.1371/journal.pone.0111894.t009

- ***Linear models***. In this class, we included three types of models:

- AR(p) models, eventually augmented with additional regressors, simultaneous and/or lagged as discussed above (i.e. ARX(p) models):

$$\Delta y_t = \mu + \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + \text{Eventual Regressors} + \varepsilon_t$$

- ARMA(p,q) models, eventually augmented with additional regressors, simultaneous and/or lagged (i.e. ARMAX(p) models):

$$\Delta y_t = \mu + \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + \text{Eventual Regressors} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

- AR(p) models with seasonal dummies, eventually augmented with additional regressors, simultaneous and/or lagged (i.e. AR-SD-X(p) models):

$$\Delta y_t = \sum_{s=1}^{12} \mu_s D_{s,t} + \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + \text{Eventual Regressors} + \varepsilon_t$$

Table 10. Top 10 models in terms of RMSE - baseline case. Forecasting: 12 steps and 24 steps ahead.

12 STEPS ahead (baseline case)		24 STEPS ahead (baseline case)	
Top 10 models	**RMSE**	**Top 10 models**	**RMSE**
AR(2)+S.D.+UR+IC+GI(J.) log 2004	1495400	AR(6)+S.D.+UR+IC+GI(J.&F.S.) log 2004	3775883
AR(5)+S.D.+UR+IC+GI(J.&F.S.) log 2004	1527588	AR(5)+S.D.+UR+IC+GI(J.&F.S.) log 2004	3777359
AR(3)+S.D.+UR+IC+GI(J.) log 2004	1534364	AR(2)+S.D.+UR+IC+GI(J.&F.S.) log 2004	3830094
AR(4)+S.D.+UR+IC+GI(J.&F.S.) log 2004	1544779	AR(2)+S.D.+UR+IC+GI(J.) log 2004	3839694
AR(2)+S.D.+UR+IC+GI(J.&F.S.) log 2004	1565497	AR(7)+S.D.+UR+IC+GI(J.&F.S.) log 2004	3861489
AR(6)+S.D.+UR+IC+GI(J.&F.S.) log 2004	1576811	AR(4)+S.D.+UR+IC+GI(J.&F.S.) log 2004	3887615
AR(6)+S.D.+UR+IC+GI(J.) log 2004	1593775	AR(6)+S.D.+UR+IC+GI(J.) log 2004	3914935
AR(7)+S.D.+UR+IC+GI(J.&F.S.) log 2004	1595086	AR(3)+S.D.+UR+IC+GI(J.&F.S.) log 2004	3939222
AR(3)+S.D.+UR+IC+GI(J.&F.S.) log 2004	1595117	AR(7)+S.D.+UR+IC+GI(J.) log 2004	3973551
AR(8)+S.D.+UR+IC+GI(J.&F.S.) log 2004	1608689	AR(8)+S.D.+UR+IC+GI(J.&F.S.) log 2004	3999943

In each row, the following information is reported: the model, the number of lags, (eventual) exogenous regressors, the data transformation, the first year of the estimation sample.

doi:10.1371/journal.pone.0111894.t010

Table 11. Number of models included in the MCS, at the 90% confidence level, using the T_{max} and T_{SQ} statistics and the MSE loss function.

	1 step		2 step		12 steps		24 step	
	T_{max}	T_{SQ}	T_{max}	T_{SQ}	T_{max}	T_{SQ}	T_{max}	T_{SQ}
Total n. of models selected	683	6	119	2	11	87	37	20
Google models	334	2	102	2	11	79	37	20
Nonlinear models	7	0	0	0	0	3	0	0

doi:10.1371/journal.pone.0111894.t011

- ***Periodic models***. Four types of models were considered:

- PAR(p) models, eventually augmented with additional regressors, simultaneous and/or lagged (i.e. PAR-X(p) models):

$$\Delta y_t = \sum_{s=1}^{12} \mu_s D_{s,t} + \sum_{i=1}^{p} \sum_{s=1}^{12} \phi_{i,s} D_{s,t} \Delta y_{t-i} + \text{Eventual Regressors} + \varepsilon_t,$$

- PAR(p) models with periodic trends, eventually augmented with additional regressors, simultaneous and/or lagged (i.e. PAR-T-X(p) models):

$$\Delta y_t = \sum_{s=1}^{12} \mu_s D_{s,t} + \sum_{s=1}^{12} \tau_s D_{s,t} T_t + \sum_{i=1}^{p} \sum_{s=1}^{12} \phi_{i,s} D_{s,t} \Delta y_{t-i} + \text{Eventual Regressors} + \varepsilon_t,$$

where $T_t = \lfloor (t-1)/12 \rfloor + 1$ with $\lfloor \cdot \rfloor$ the integer function, represents an annual linear deterministic trend.

- PAR(p)-ARCH(1) models, eventually augmented with additional regressors, simultaneous and/or lagged (i.e. PAR-X(p)-ARCH(1) models):

$$\Delta y_t = \sum_{s=1}^{12} \mu_s D_{s,t} + \sum_{i=1}^{p} \sum_{s=1}^{12} \phi_{i,s} D_{s,t} \Delta y_{t-i} + \text{Eventual Regressors} + \varepsilon_t,$$

$$\varepsilon_t = \eta_t \sqrt{\sigma_t^2}, \quad \eta_t \sim N(0,1)$$

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2$$

- Periodic Error Correction (PEC) models: we considered the case of periodic cointegration when the variables have a non-periodic unit root:

$$\Delta_1 y_{1,t} = \sum_{s=1}^{12} \gamma_{1,s} (y_{1,t-1} - \alpha - \kappa y_{2,t-1}) + \sum_{i=1}^{p} \psi_{1,i} \Delta_1 y_{1,t-i} \sum_{i=1}^{p} \psi_{2,i} \Delta_1 y_{2,t-i} + \varepsilon_{1,t} \quad (1)$$

where $y_{1,t}$ is the number of food stamps recipients while $y_{2,t}$ is a set of regressors, see [34] for more details about this single-equation cointegration model. For computational tractability, only the two

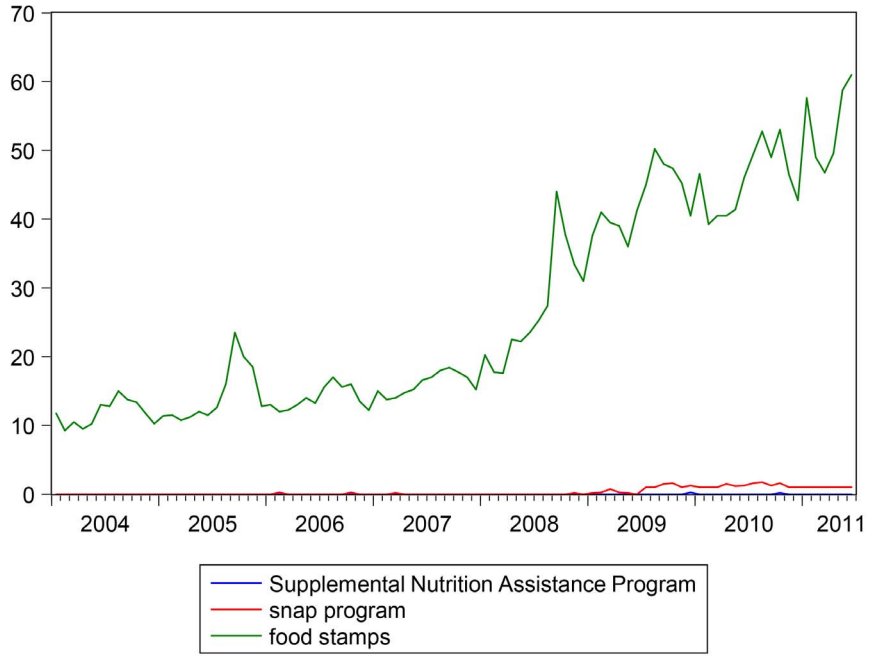


Figure 4. GIs for the keywords *"Supplemental Nutrition Assistance Program"*, *"snap program"*, and *"food stamps"*. Sample: 2004M1 - 2011M5. Search keywords are not case sensitive. Google data are registered trademarks of Google Inc., used with permission.
doi:10.1371/journal.pone.0111894.g004

cases of cointegration between food stamps and GIs, and cointegration between food stamps, UR and IC were considered. Considering the relative small out-of-sample (more below) and the number of variables involved, we considered PEC(1,12) models.

- ***Multivariate models***. Three types of models were considered in this class:

- Vector Auto-Regressive (VAR) models: given the sample dimension and the number of variables, we considered only trivariate VAR models including either food stamps and the two GIs, or food stamps and the UR and the IC.

- Vector Error Correction (VEC) models, where all potential cointegration relationship between food stamps and the four regressors (UR, IC, GIs for "food stamps" and "jobs") were explored. We considered the case of no trend in data and no trend in cointegration relation, but with intercepts in the cointegration relations:

$$\Delta \boldsymbol{Y}_t = \boldsymbol{B}(\boldsymbol{\Gamma}' \boldsymbol{Y}_{t-1} + \boldsymbol{a}_0) + \sum_{j=1}^{p-1} \zeta_j \Delta \boldsymbol{Y}_{t-j} + \varepsilon_t$$

where $\boldsymbol{Y}_t$ is an $n \times 1$ vector process, $\mathbf{B}$ is an $n \times h$ matrix of loadings (or correction factors), $\boldsymbol{\Gamma}$ is an $n \times h$ matrix containing the cointegrating vectors, and h is the number of cointegrating relationships, i.e. the cointegration rank, see [38] for details. Similarly to PEC models, we considered VEC(1,12) models, with only 1 CE for computational tractability.

- Bayesian Vector Auto-Regressive (BVAR) models: when there are a lot of variables and lags, estimating VAR models can be challenging, if not impossible. One way to solve this issue is to shrinkage the parameters by using Bayesian methods. Bayesian VAR models has recently enjoyed a large success in macroeconomic forecasting, see [64] for a recent review and [21] for a recent application with Google data. More specifically, we used the so called Litterman/Minnesota prior, which is a common choice in empirical applications due to its computational speed and forecasting success, see [65], [66] and [64] for more details.

- ***Nonlinear models***. Four types of models were considered (see [21] for a recent application of these nonlinear models to forecast the real price of oil using Google data - the nonlinear model proposed by [8] to forecast food stamps caseloads was not considered because we did not have neither the monthly data relative to the new entries and exits for each state for the

Table 12. Number of nowcasting and forecasting models selected in the MCS at the 90% confidence level, using the T_{max} statistic and the MSE loss function, as well as number of selected models using the "false" Google Index.

	N. 1 step	N. 2 steps	F. 12 steps	F. 24 steps
Models selected	614	122	37	36
Models using the "false" Google Index	29	2	0	0

doi:10.1371/journal.pone.0111894.t012

Table 13. Top 10 models in terms of RMSE - different out-of-sample periods. Nowcasting: 1 step ahead.

1 STEP ahead (baseline case)	**1 STEP ahead (recession 2008)**
Top 10 models	**Top 10 models**
ARMA(10,10) dlog 1988	BVAR(1,12) FS+UR+IC+GI(F.S.) lev 2004
ARMA(10,10) + UR dlog 1988	BVAR(7) FS+UR+IC+GI(F.S.) lev 2004
ARMA(12,12) dlog 1988	AR(4)+S.D.+GI(F.S.) lev 2004
ARMA(11,11)+UR diff 1988	AR(7)+S.D.+GI(F.S.) lev 2004
AR(12) + IC (sim + lags) dlog 1988	AR(5)+S.D.+GI(F.S.) lev 2004
ARMA(12,12) + IC diff 1988	AR(8)+S.D.+GI(F.S.) lev 2004
BVAR(1,12) FS+GI(F.S.) lev 2004	AR(6)+S.D.+GI(F.S.) lev 2004
BVAR(1,12) FS+UR+IC+GI(F.S.) lev 2004	ARMA(11,11)+UR diff 1988
ARMA(12,12) + UR dlog 1988	ARMA(12,12)+GI(J.&F.S.) lev 2004
ARMA(11,11) dlog 1988	AR(6)+S.D.+GI(J.&F.S.) log 2004
1 STEP ahead (baseline case)	**1 STEP ahead (expansion 2009)**
Top 10 models	**Top 10 models**
ARMA(10,10) dlog 1988	AR(8)+S.D.+IC lev 2004
ARMA(10,10) + UR dlog 1988	AR(9)+S.D.+IC lev 2004
ARMA(12,12) dlog 1988	AR(7)+S.D.+IC lev 2004
ARMA(11,11) + UR diff 1988	AR(7)+S.D.+UR+IC+GI(J.) log 2004
AR(12) + IC (sim + lags) dlog 1988	AR(10)+S.D.+IC lev 2004
ARMA(12,12) + IC diff 1988	AR(8)+S.D.+UR+IC+GI(J.) log 2004
BVAR(1,12) FS+GI(F.S.) lev 2004	AR(6)+S.D.+GI(J.) log 2004
BVAR(1,12) FS+UR+IC+GI(F.S.) lev 2004	AR(4)+S.D.+IC lev 2004
ARMA(12,12) + UR dlog 1988	AR(5)+S.D.+IC lev 2004
ARMA(11,11) dlog 1988	AR(10)+S.D.+UR+IC+GI(J.) log 2004

Baseline case (left column) and the two cases including the 2008 recession (top right column) and the expansion starting in 2009 (low right column).
doi:10.1371/journal.pone.0111894.t013

income-based food stamps program, nor the monthly data relative to the new entries and exits for the disaster food stamps):

- Self-Exciting Threshold AutoRegressive (SETAR) models with 2 regimes:

$$Y_t = \begin{cases} \phi_{0,1} + \phi_{1,1} Y_{t-1} + \ldots + \phi_{1,p} Y_{t-p} + \varepsilon_t, & \text{if} \quad Y_{t-1} \le c \\ \phi_{0,2} + \phi_{1,2} Y_{t-1} + \ldots + \phi_{1,p} Y_{t-p} + \varepsilon_t, & \text{if} \quad Y_{t-1} > c \end{cases}$$

where c is a threshold to be estimated and which identifies the two regimes.

- Logistic smooth transition autoregressive (LSTAR) models, which are a generalization of the SETAR model:

$$Y_t = (\phi_{0,1} + \phi_{1,1} Y_{t-1} + \ldots + \phi_{1,p} Y_{t-p})[1 - G(Y_{t-1}, \gamma, c)] +$$

$$+ (\phi_{0,2} + \phi_{1,2} Y_{t-1} + \ldots + \phi_{1,p} Y_{t-p})[G(Y_{t-1}, \gamma, c)] + \varepsilon_t$$

where $G(Y_{t-1}, \gamma, c) = [1 + \exp(-\gamma(Y_{t-1} - c))]^{-1}$ is the first order logistic transition function, bounded between 0 and 1, γ is the slope parameter and c is the location parameter. In contrast with the SETAR models, the LSTAR model assumes that the change between the two regimes is gradual and smooth. This model belongs to the class of Smooth Transition AR models, see [67] for details.

- Neural Network (NNET) models with linear output, defined as follows:

$$Y_t = \beta_0 + \sum_{j=1}^{D} \beta_j g\left(\phi_{0,j} + \phi_{j,1} Y_{t-1} + \ldots + \phi_{j,p} Y_{t-p}\right) + \varepsilon_t$$

where D is the number of hidden units and g is the activation function, given by the logistic function. See [68] (chapter 8) and [69] (chapter 5) for details. In this case, we chose the number of hidden units D to be 3 based on information criteria.

- Additive Autoregressive models (AAR), also known as generalized additive models, since they combines generalized linear models and additive models:

$$Y_t = \phi_0 + s_1(Y_{t-1}) + \ldots + s_p(Y_{t-p}) + \varepsilon_t$$

where s_i are smooth functions represented by penalized cubic regression splines, see [70] for details.

Table 14. Top 10 models in terms of RMSE - different out-of-sample periods. Nowcasting: 2 steps ahead.

2 STEPS ahead (baseline case)	**2 STEPS ahead (recession 2008)**
Top 10 models	**Top 10 models**
AR(3)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(8)+S.D.+GI(F.S.) lev 2004
AR(8)+S.D.+GI(F.S.) lev 2004	AR(7)+ S.D.+GI(F.S.) lev 2004
AR(7)+ S.D.+GI(F.S.) lev 2004	AR(10)+ S.D.+GI(F.S.) lev 2004
AR(2)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(4)+ S.D.+GI(F.S.) lev 2004
AR(7)+S.D.+GI(J.&F.S.) lev 2004	AR(9)+ S.D.+GI(F.S.) lev 2005
AR(4)+ S.D.+GI(F.S.) lev 2004	AR(5)+ S.D.+GI(F.S.) lev 2006
AR(4)+S.D.+GI(J.&F.S.) lev 2004	AR(6)+ S.D.+GI(F.S.) lev 2007
AR(8)+S.D.+GI(J.&F.S.) lev 2004	AR(11)+ S.D.+GI(F.S.) lev 2007
AR(5)+ S.D.+GI(F.S.) lev 2004	ARMA(12,12)+GI(J.&F.S.) log 2004
AR(6)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(6)+S.D.+GI(J.&F.S.) log 2004
2 STEPS ahead (baseline case)	**2 STEPS ahead (expansion 2008)**
Top 10 models	**Top 10 models**
AR(3)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(8)+S.D.+IC lev 2004
AR(8)+S.D.+GI(F.S.) lev 2004	AR(7)+S.D.+IC lev 2004
AR(7)+ S.D.+GI(F.S.) lev 2004	AR(10)+S.D.+IC lev 2004
AR(2)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(7)+S.D.+UR+IC+GI(J.&F.S.) log 2004
AR(7)+S.D.+GI(J.&F.S.) lev 2004	AR(9)+S.D.+IC lev 2004
AR(4)+ S.D.+GI(F.S.) lev 2004	AR(2)+S.D.+UR+IC+GI(F.S.) lev 2004
AR(4)+S.D.+GI(J.&F.S.) lev 2004	AR(2)+S.D.+UR+IC+GI(J.&F.S.) log 2004
AR(8)+S.D.+GI(J.&F.S.) lev 2004	AR(6)+S.D.+UR+IC+GI(J.&F.S.) log 2004
AR(5)+ S.D.+GI(F.S.) lev 2004	AR(3)+S.D.+UR+IC+GI(F.S.) log 2004
AR(6)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(5)+S.D.+IC lev 2004

Baseline case (left column) and the two cases including the 2008 recession (top right column) and the expansion starting in 2009 (low right column).
doi:10.1371/journal.pone.0111894.t014

- Last, but not least, ***the Random Walk with Drift model***, which is the classical benchmark model for economic and financial forecasting: $\Delta Y = c + \varepsilon_t$.

The full list of the 2890 models used in our forecasting exercise is reported in the Tables 5–6. Finally, we remark that our forecasting comparison was structured to replicate the situation that real forecasters face when they compute their forecasts, and they have to use only the information available at each point in time: therefore, predictions of all the exogenous variables (for example Google data) have to be computed to forecast the endogenous variables of interest (in our case, the number of food stamps recipients). To satisfy this criterion and avoid any problem of look-ahead bias, we had to choose a forecasting model for the unemployment rate, the initial claims and the Google indexes: based on information criteria and residuals properties, we selected a PAR(1) model for the monthly unemployment rate and initial claims in logs, and an AR(12) model for the log-returns of Google indexes. The forecasts of these exogenous variables then served as inputs in the forecasting models for monthly food stamps data. Clearly, we could have considered a range of models for the exogenous regressors, but this would have increased exponentially the total number of models, making the forecasting exercise computationally untractable. Therefore, we leave this issue as an avenue for further research.

Results

Out-of-Sample Forecasting Analysis

We used the data between 1988M10 and 2007M2 as the first initialization sample for the models without GIs, while we used the initialization sample 2004M1-2007M2 for the models with GIs and for those models without GIs but estimated on a shorter sample. The evaluation period ranged from 2007M3 till 2011M5 and was used to compare the nowcasts 1 step and 2 steps ahead, as well as the forecasts 12 steps and 24 steps ahead. The total number of models using Google data among the Top 100 models in terms of Root Mean Square Error (RMSE) is reported in Table 7, while Table 8 reports the ranking of the best models within each class according to the RMSE. Finally, the top 10 models in terms of the RMSE for nowcasting and forecasting are reported in Tables 9–10.

In general, Google-based models performed very well both for nowcasting and forecasting. In this regard, Table 7 shows that the number of models with Google data in the Top 100 ranked models in terms of RMSE is very high, particularly for nowcasting 2 steps ahead and forecasting, where more than 90 models include Google data.

In case of *nowcasting*, linear AR and ARMA models augmented with seasonal dummies were sufficient to provide good nowcasts of the food stamps data. Particularly, simple linear models using the log-returns of food stamps and no additional regressors, were sufficient for nowcasting 1 step ahead. Instead, ARX(p) models with seasonal dummies and Google data were the best choice for

Table 15. Top 10 models in terms of RMSE - different out-of-sample periods. Forecasting: 12 step ahead.

12 STEPS ahead (baseline case)	**12 STEPS ahead (recession 2008)**
Top 10 models	**Top 10 models**
AR(2)+S.D.+UR+IC+GI(J.) log 2004	PAR(1)+UR+IC+GI(J.) log 2004
AR(5)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(2)+S.D.+UR+IC+GI(J.) log 2004
AR(3)+S.D.+UR+IC+GI(J.) log 2004	AR(3)+S.D.+UR+IC+GI(J.) log 2004
AR(4)+S.D.+UR+IC+GI(J.&F.S.) log 2004	PAR(1)+UR+IC+GI(J.) lev 2004
AR(2)+S.D.+UR+IC+GI(J.&F.S.) log 2004	PAR(1)+UR+IC+GI(J.&F.S.) lev 2004
AR(6)+S.D.+UR+IC+GI(J.&F.S.) log 2004	PAR(1)+UR+IC+GI(J.&F.S.) log 2004
AR(6)+S.D.+UR+IC+GI(J.) log 2004	AR(5)+S.D.+UR+IC+GI(J.&F.S.) log 2004
AR(7)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(6)+S.D.+UR+IC+GI(J.) log 2004
AR(3)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(4)+S.D.+UR+IC+GI(J.&F.S.) log 2004
AR(8)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(5)+S.D.+UR+IC+GI(J.) log 2004
12 STEPS ahead (baseline case)	**12 STEPS ahead (expansion 2009)**
Top 10 models	**Top 10 models**
AR(2)+S.D.+UR+IC+GI(J.) log 2004	PAR(1)+UR+IC+GI(J.&F.S.) log 2004
AR(5)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(4)+S.D.+GI(J.&F.S.) log 2004
AR(3)+S.D.+UR+IC+GI(J.) log 2004	AR(5)+S.D.+GI(J.&F.S.) log 2004
AR(4)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(6)+S.D.+GI(J.&F.S.) log 2004
AR(2)+S.D.+UR+IC+GI(J.&F.S.) log 2004	PAR(1)+UR+IC+GI(J.&F.S.) lev 2004
AR(6)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(7)+S.D.+GI(J.&F.S.) log 2004
AR(6)+S.D.+UR+IC+GI(J.) log 2004	AR(4)+S.D.+UR+IC+GI(J.&F.S.) log 2004
AR(7)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(5)+S.D.+UR+IC+GI(J.&F.S.) log 2004
AR(3)+S.D.+UR+IC+GI(J.&F.S.) log 2004	PAR(1)+UR+IC+GI(J.) log 2004
AR(8)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(2)+S.D.+UR+IC+GI(J.) log 2004

Baseline case (left column) and the two cases including the 2008 recession (top right column) and the expansion starting in 2009 (low right column).
doi:10.1371/journal.pone.0111894.t015

nowcasting 2 steps ahead, see Table 9. Interestingly, the specification of the best models in this case is quite close to the one selected by the structural relationship identification methodology in Table 4. Moreover, the best models were those with the food stamps data in levels or in log-levels, thus confirming the previous evidence of stationarity.

As for *forecasting*, the evidence is strongly in favor of Google-based models, where all top models in terms of RMSE have a specification very close to the ARX(4) model with seasonal dummies reported in Table 4. Considering that we compared almost 3000 models, this is rather encouraging and confirms that the structural relationship identification methodology by [29] and [30] is a rather robust method of model selection. With regard to

Table 16. Top 10 models in terms of RMSE - different out-of-sample periods. Forecasting: 24 steps ahead.

24 STEPS ahead (baseline case)	**24 STEPS ahead (expansion 2009)**
Top 10 models	**Top 10 models**
AR(6)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(2)+S.D.+UR+IC+GI(J.) log 2004
AR(5)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(6)+S.D.+UR+IC+GI(J.) log 2004
AR(2)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(3)+S.D.+UR+IC+GI(J.) log 2004
AR(2)+S.D.+UR+IC+GI(J.) log 2004	AR(5)+S.D.+UR+IC+GI(J.) log 2004
AR(7)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(7)+S.D.+UR+IC+GI(J.) log 2004
AR(4)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(8)+S.D.+UR+IC+GI(J.) log 2004
AR(6)+S.D.+UR+IC+GI(J.) log 2004	AR(9)+S.D.+UR+IC+GI(J.) log 2004
AR(3)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(11)+S.D.+UR+IC+GI(J.) log 2004
AR(7)+S.D.+UR+IC+GI(J.) log 2004	AR(4)+S.D.+UR+IC+GI(J.) log 2004
AR(8)+S.D.+UR+IC+GI(J.&F.S.) log 2004	AR(6)+S.D.+UR+IC+GI(J.&F.S.) log 2004

Baseline case (left column) and the case including the expansion starting in 2009 (low right column).
doi:10.1371/journal.pone.0111894.t016

Table 17. Number of nowcasting and forecasting models selected in the MCS at the 90% confidence level, using the T_{max} statistic and the MSE loss function, as well as number of selected Google based models.

		Recession 2008	Expansion 2009
Nowcasting 1 step	Models selected	173	82
	Google based models	101	68
Nowcasting 2 steps	Models selected	101	51
	Google based models	89	42
Forecasting 12 steps	Models selected	22	5
	Google based models	22	5
Forecasting 24 steps	Models selected	NA	13
	Google based models	NA	13

doi:10.1371/journal.pone.0111894.t017

nonlinear models, only AAR models showed good performances, while this was not the case for the other three nonlinear models that we considered in our analysis: SETAR models did not reach numerical convergence under all possible configurations; LSTAR were a little bit better, but they were computationally demanding and almost 15% of the considered configurations did not reach convergence. Moreover, most of them had a ranking position above the 1500th place in terms of RMSE. Neural Networks were rather quick to estimate but similarly to LSTAR models they did not fare well in terms of ranking positions: in case of nowcasting, the majority of them ranked higher than the 1500th place, while in case of forecasting most of them ranked above the 1000th position. As for periodic models, simple PAR(p) models and PECM models including only food stamps data and GIs performed rather very well in case of forecasting, whereas more complex configurations with periodic trends, ARCH effects or alternative periodic cointegration models performed rather poorly: most likely, the wealth of parameters that these complex periodic models involves resulted in very imprecise estimates. As for multivariate models, they were generally out of the top 100 models in terms of RMSE and most likely they suffered from efficient loss due to the high number of parameters (the only exception were Bayesian models which performed very good for nowcasting 1 step ahead).

We then tested for statistically significant differences in the forecast performances among the competing models by using the Model Confidence Set (MCS) approach developed by [71]. The MCS is a sequential test of equal predictive ability: given an initial set of forecasting models it tests the null that no forecasting model is distinguishable from any other, so that the starting hypothesis is that all models considered have equal forecasting performances. The MCS procedure yields a model confidence set containing the best forecasting models at some confidence level. An additional advantage of the MCS is that it acknowledges the limits to the informational content of the data: informative dataset will deliver a set that contains only the best model, while less informative data will not be able to distinguish between the competing models and the final set may contain several, if not all, models. We considered the maximum t statistic T_{max}, which is the default statistic in [71], as well as the semi-quadratic statistic T_{SQ}, which is more computationally intensive but more selective (see e.g. [72] and [21] for some recent applications). The loss function used was the Mean Squared Error (MSE), while the p-values for the test statistic were obtained by using the stationary block bootstrap with a block length of 12 months and 1000 re-samples: if the p-value was lower than a defined threshold probability level α, the model was not included in the MCS and vice-versa. We set $\alpha = 0.10$ as in [71].

We report in Table 11 the number of models selected by the MCS procedure according to the MSE loss function, for nowcasting and forecasting. Moreover, we also report the number of selected models including Google data, as well as the number of selected nonlinear models.

In general, we can note that the number of models selected is quite small, with the only exception of the T_{max} statistic for the case of nowcasting, which selected from a minimum of 119 models up to 683 models. In all other cases, the selected models are no more than 40, which indicates that our dataset is rather informative and it can be used to separate poor forecasting models from superior ones. Moreover, the semi-quadratic statistic is much more selective than the maximum t statistic, as expected, and in the case of nowcasting 2 steps ahead it selects only two models: the ARX(3) with seasonal dummies, data in log levels and all exogenous variables and the ARX(8) with seasonal dummies, data in levels and only one exogenous variables included (the GI for the keyword "food stamps"). With the exception of nowcasting 1 step ahead, models with Google data represent the vast majority of the models included in the MCS: this is quite strong evidence that Google online search data provide additional information content not included in standard economic variables.

The fact that simple linear models, augmented with the search volumes for the keywords "food stamps" and "jobs", improve so much the forecasting at long horizons is an indirect confirmation of the recent analysis of food stamps caseload dynamics by [73]

Table 18. Directional accuracy of forecasts: number of models with 100% correct predictions for the direction of change.

	N. 1 step	N. 2 steps	F. 12 steps	F. 24 steps
N. of models	1	179	1096	1252
Google based models	1	101	715	815

doi:10.1371/journal.pone.0111894.t018

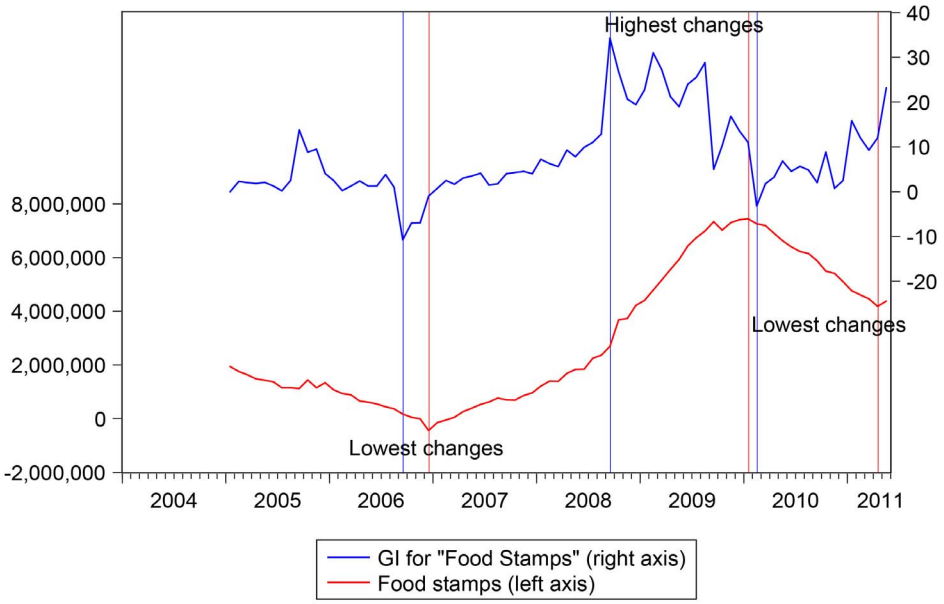


Figure 5. Yearly changes for the food stamps data and the GI for ''food stamps''. Sample: 2004M1 - 2011M5. The turning points for each series is highlighted by a vertical line of the same color. Google data are registered trademarks of Google Inc., used with permission.
doi:10.1371/journal.pone.0111894.g005

and [4], who showed that "*caseloads spell lengths had increased substantially from earlier time periods and… the impact of the current record caseloads may be felt for a long time to come*", [[4], p. 327].

Robustness checks

We wanted to verify that our promising results with Google data hold also with different assumptions, alternative forecasting environments and different comparison methodologies. Therefore, we performed the following robustness checks: a) we verified whether alternative keywords in Google Trends could be used for forecasting the number of food stamps recipients; b) we employed a recent falsification test proposed by [16]; c) we considered alternative out-of-sample intervals with and without the global financial crisis included; d) we compared the models in terms of directional accuracy; e) we repeated the same forecasting exercise for each of the 50 US states plus the Department of Columbia. All checks confirmed the superior forecasting performance of Google based models in general and of ARX models in particular, with specifications always close to that found using the structural relationship identification methodology by [29] and [30].

Alternative Keywords. An important issue is to verify whether alternative keywords for Google searches can be used for forecasting purposes in the place of those used in the main analysis (i.e. "food stamps" and "jobs"). It is well known that in October 2008 the US farm bill renamed the Food Stamp Program as the *Supplemental Nutrition Assistance Program*. However, if we compare the online search volumes for this new name, together with "snap program" and the standard search "food stamps", we can see that the keywords of interest remain only "food stamps" (see Figure 4): the alternative keywords *Supplemental Nutrition Assistance Program* and *snap program* have much lower search volumes and they start having Google indexes different from zero only from the end of 2008, so that they cannot be used in our forecasting exercise. Moreover, the vast majority of searches involving these alternative keywords also includes either "food stamps" or "food program". Therefore, our case seems to be quite different from forecasting the US unemployment rate with Google data as in [16], where there can be alternative important keywords beside the main one given by "*jobs*".

A Falsification Test using Google Correlate. Google has recently developed a new application called Google Correlate which can find out the web searches for keywords that either show the highest correlation with a given keyword search, or show the highest correlation with a given time series, given a specific time horizon. More specifically, its objective is to "*to surface the queries in the [Google] database whose spatial or temporal pattern is most highly correlated (R^2) with a target pattern*" (official Google Correlate white paper, p. 2, available at http://www.google.com/trends/correlate/whitepaper.pdf). Google Correlate is basically Google Trends in reverse.

As a further check, we therefore decided to employ the falsification test proposed by [16], which considers the forecasting performance of an alternative GI chosen by Google Correlate: in our case, the keyword search that had the highest correlation with the US food stamps data during the in-sample period (2004M1-2007M2) was "*pci express slot*." It is clear that such terms have nothing to do with the food stamps program. We thus added 375 forecasting models using the new Google Index to our previous 2890 models considered in the baseline case.

We do not report the top 10 models in terms of the Root Mean Square Error (RMSE) as in the previous section, since no single model using the "false" Google keywords was among the Top 100 models for every forecasting horizon. Instead, we report in Table 12 the number of models selected by the MCS procedure, together with the number of selected models using the "false" Google Index.

Table 12 shows that only a very limited number of models using the "false" GI were selected in case of nowcasting, while not a single model was selected in case of forecasting. These poor results were expected since the "false" Google data were completely disconnected from the food stamps program. Therefore, this evidence suggests that purely automatic methods (like Google

Table 19. Directional accuracy for $\Delta^2 y$: number of first-best and second-best models, together with their percentage of correct predictions for the sign of $\Delta^2 y$.

		N. 1 step		N. 2 steps		F. 12 steps		F. 24 steps	
		N.	*% correct*	*N.*	*% correct*	*N.*	*% correct*	*N.*	*% correct*
(1st best)	Models total	6	82%	3	80%	1	79%	2	81%
	Google based models	1	82%	0	/	0	/	1	81%
(2nd best)	Models total	12	80%	7	78%	2	77%	6	78%
	Google based models	1	80%	4	78%	0	/	5	78%

doi:10.1371/journal.pone.0111894.t019

Correlate) may not always represent the optimal keyword selection choice, see also [21] for similar results in case of oil data.

Different Out-Of-Sample Periods. We wanted to verify the forecasting performance of our competing models using different out-of-sample periods, to check the robustness of our results to different business cycle conditions: for example, our preliminary in-sample analysis highlighted a potential structural break for Google based models in 2008, with a timing close to the Lehman Brothers bankruptcy. In this regard, we followed the approach suggested by [16] and considered the following two alternative out-of-samples:

- 2008M10-2011M5: this sample starts just after the bankruptcy of Lehman Brothers;
- 2009M7- 2011M5: this second sample starts with the end of the (official) NBER recession for the US in 2009.

Due to the new out-of-samples dimensionality, forecasts 24 steps ahead are considered only for the second sample starting in 2009. The top 10 models in terms of the Root Mean Square Error (RMSE) for nowcasting and forecasting are reported in Tables 13–16, together with the top 10 models for the baseline case, which are reported in the left column for ease of reference; Table 17 reports the number of models selected by the MCS procedure according to the MSE loss function and the T_{max} statistic at the 90% confidence level, together with the number of selected models using Google data.

Tables 13–14 show that, in case of nowcasting, Google based models tend to do particularly well during the recession period, while models using the unemployment rate and initial claims tend to perform better in terms of RMSE during the economic expansion, even though in the latter case the difference is rather small. This evidence is confirmed by the Model Confident Set approach, where more than 60% of the selected models are Google based models, for both out-of-sample periods. As for forecasting 12 and 24 steps ahead, Tables 15–16 show that the results are quite similar to the baseline case instead, with Google based models in the top spots. Moreover, all selected models by the MCS are Google based models. Interestingly, these two alternative out-of-sample periods are much more informative for the MCS approach, because the number of selected models is much lower compared to the baseline case reported in Table 11 (particularly for nowcasting). Therefore, this evidence highlights that Google models are much more stable than competing models, and their forecasting performances are robust across different business cycles, as recently found also by [16] and [21].

Directional Accuracy and Turning Points: Can Google Help. The analysis has considered so far only the accuracy of forecasts in terms of magnitude, but also directional accuracy is important: even if forecast errors are large, forecasts with the correct direction of change may still provide useful information about food stamps dynamics. A special case of directional accuracy is the ability to predict a turning point, which is a change in the direction of movement of the variable under investigation, and it exists if $y_t < y_{t-1} > y_{t-2}$ (peak turning point) or $y_t > y_{t-1} < y_{t-2}$ (trough turning point), see [74] and [75] for details.

Unfortunately, our forecasting evaluation period (2007M3-2011M5) spans a limited time sample, where food stamps caseloads mainly increased (see Figure 2). Nevertheless, if we evaluate the directional accuracy of the competing forecasting models, we are still able to identify a limited group of best models, at least for nowcasting (see Table 18).

Table 18 shows that there was only 1 model able to correctly predict all the 51 directions of change in case of nowcasting 1 step

Table 20. US state level forecasts.

US state	1 step	2 steps	12 steps	24 steps	Census Population	Density (inhabitants per square mile)
Alabama	38	100	89	87	4822023	92
Alaska	45	61	62	38	731449	1
Arizona	44	36	4	20	6392017	56
Arkansas	0	9	0	26	2915918	55
California	34	68	84	72	37253956	228
Colorado	1	3	4	53	5029196	48
Connecticut	57	48	17	94	3574097	645
Delaware	24	52	42	40	897934	361
District of Columbia	2	1	41	14	601723	8805
Florida	0	13	13	42	18801310	286
Georgia	35	35	58	75	9687653	163
Hawaii	20	30	35	35	1360301	124
Idaho	35	36	11	10	1567582	19
Illinois	38	68	87	67	12830632	222
Indiana	30	48	22	8	6483802	178
Iowa	34	60	38	23	3046355	54
Kansas	100	98	18	51	2853118	35
Kentucky	66	64	41	44	4339367	107
Louisiana	28	14	98	94	4533372	87
Maine	21	24	51	57	1328361	38
Maryland	64	70	43	60	5296486	427
Massachusetts	73	65	63	50	6349097	602
Michigan	64	65	62	66	9938444	103
Minnesota	16	8	45	65	4919479	57
Mississippi	7	3	35	39	2844658	59
Missouri	8	0	2	1	5595211	80
Montana	53	48	35	36	902195	6
Nebraska	1	0	9	95	1711263	22
Nevada	39	26	13	21	1998257	18
New Hampshire	60	14	41	86	1235786	132
New Jersey	57	72	84	87	8414350	965
New Mexico	43	43	46	58	1819046	15
New York	1	14	76	72	18976457	348
North Carolina	65	80	83	74	8049313	150
North Dakota	4	17	41	38	642200	9
Ohio	16	18	60	73	11353140	253
Oklahoma	57	65	6	14	3450654	49
Oregon	74	56	24	2	3421399	35
Pennsylvania	78	76	49	67	12281054	267
Rhode Island	28	46	76	93	1048319	679
South Carolina	65	70	41	42	4012012	125
South Dakota	43	55	53	95	754844	10
Tennessee	6	53	46	33	5689283	135
Texas	16	75	70	60	20851820	78
Utah	39	56	31	39	2233169	26
Vermont	18	33	29	20	608827	63
Virginia	73	74	69	40	7078515	165
Washington	47	44	29	49	5894121	83

Table 20. Cont.

US state	1 step	2 steps	12 steps	24 steps	Census Population	Density (inhabitants per square mile)
West Virginia	1	12	23	24	1808344	75
Wisconsin	12	4	4	15	5363675	82
Wyoming	5	8	52	33	493782	5

Number of models using Google data out of the Top 100 models (according to the RMSE), 2010 census population data for each US state and population density (inhabitants per square mile, 2010).
doi:10.1371/journal.pone.0111894.t020

ahead, and this is a Google based model (the AR(3)+UR+IC+GI("food stamps") model using log-transformed data). Instead, it is not a surprise that the number of models with a 100% directional accuracy increases with the forecast horizon, since the number of food stamps was continuously increasing in the considered forecasting sample: therefore, directional accuracy cannot be used to discriminate competing models in this case.

As for turning points, we could not evaluate the models' ability to predict them because there were none in the forecasting sample. Nevertheless, a simple indirect way to check whether the forecasting models would have been able to predict them is to evaluate their ability to correctly forecast the sign of second order differenced data (that is $\Delta^2 y = \Delta y - \Delta y_{t-1}$). To get an intuitive idea of this point, we plot in Figure 5 the yearly changes Δy of the number of food stamps recipients and the Google Index for "food stamps": even though the yearly changes of food stamps data were almost always positive, they had a declining rate between 2004 and 2006 (that is $\Delta^2 y < 0$), an increasing rate between 2006 and 2009 ($\Delta^2 y > 0$) and again a declining rate between 2010 and 2011 ($\Delta^2 y < 0$). Interestingly, the yearly changes of the GI for the keywords "food stamps" showed a similar pattern, which always anticipated the turning points of the yearly changes of food stamps data: from a minimum of 3 months in advance in 2006, up to 16 months in 2008 and 14 months in 2010. Therefore, Figure 5 gives some clues for understanding why Google based models forecasted so well food stamps 12 steps and 24 steps ahead.

The directional accuracy of the competing models for the second order differenced data $\Delta^2 y$ is reported in Table 19. Given that the selection is now much higher than for the first order changes discussed in Table 18, we report both the first-best models and the second-best models.

The results in Table 19 are somewhat mixed but partially confirm what we previously found out when examining the accuracy of forecasts in terms of magnitude: simple linear models augmented with initial claims and the unemployment rate are sufficient for nowcasting food stamps, while Google based models perform better for nowcasting 2 steps ahead and for long term forecasting (24 steps ahead).

US State Level Forecasts. The last check was to estimate the same set of forecasting models for each of the 50 US states, together with the District of Columbia. A similar check was implemented by [16] when forecasting the US unemployment rate with Google data. As in the baseline case, the evaluation period ranged from 2007M3 till 2011M5 and was used to compare the nowcasts 1 step and 2 steps ahead, as well as the forecasts 12 steps and 24 steps ahead. For sake of interest and space, we report in Table 20 the number of models using Google data among the Top 100 models according to the RMSE for each US state. Moreover, the same table also reports the US census state population and the population density per square mile as of April 1, 2010.

The results are quite similar to what we saw for the whole US (not surprisingly). However, two outcomes are worth noticing: Google data seems to be more useful for forecasting highly densely populated US states, while its importance is minor for several states with small population and low density. Probably, this may be due to a higher internet penetration in highly densely populated states. However, this issue goes beyond the scope of this paper and we leave it as an interesting avenue for further research. Secondly, the number of models with Google data in the Top 100 increases with the forecasting horizon, thus confirming similar evidence in [16].

Conclusion

We proposed the use of Google data based on internet searches about food stamps as a potential indicator to nowcast and forecast the US monthly number of individuals participating in the Supplemental Nutrition Assistance Program, formerly known as the Food Stamp program. We compared almost 3000 forecasting models and we found that Google based models definitively improved nowcasting food stamps 2 months ahead, while simple linear models (eventually augmented with unemployment rates or initial claims data) are sufficient for nowcasting 1 month ahead. Moreover, Google based models provided statistically significant superior forecasts in case of forecasting 12 steps and 24 steps ahead. More specifically, linear autoregressive models augmented with Google search data for the terms "food stamps" and "jobs", the unemployment rate and initial claims were the best models for forecasting purposes. In this regard, the best models had specifications always close to the ARX(4) model found using the structural relationship identification methodology by [29] and [30] in the in-sample analysis, thus showing that this approach is a rather robust method of model selection in case of small samples.

Nonlinear models performed poorly, were computationally intensive, and in several cases did not reach numerical convergence, with the exception of additive autoregressive models which provided competitive forecasts in case of long term forecasting. Simple periodic autoregressive models performed quite well for 12-step and 24-step ahead forecasts, while more complex periodic models performed poorly, probably due to the high number of estimated parameters which hindered their forecasting performances. Our results hold also with alternative out-of-sample periods which either include the global financial crisis or start after the (official) end of this recession. Besides, our Google based models passed a falsification test which considered the forecasting performance of an alternative Google index chosen by Google Correlate. Similar results were also found when considering the

directional accuracy of the models' forecasts. Finally, the estimates for single US states gave similar results to the case of the whole US, even though we found that Google data are very important in case of highly densely populated US states, while their importance is minor for several states with small population.

We remark that although we considered a very large set of models, we had to restrict the potential range of models to keep the forecasting exercise computationally tractable. An avenue of future research would be to consider additional models like fractional cointegration, exponential smoothing methods in state space form and many others.

Supporting Information

Software Description S1 The Software description.
(PDF)

Data S1 The (cleaned) food stamps data.
(ZIP)

Acknowledgments

We are grateful to the U.S. Department of Agriculture for providing us with the data used in the analysis: particularly, I want to thank John Kirlin (Economic Research Service USDA), Kathryn Law and Jenny Laster Genser (Office of Research and Analysis - USDA) for their great help and assistance. We also thank David Powers (Social, Economic, and Housing Statistics Division - U.S. Census Bureau) for precious comments, Scott C. Linn (R.W. "Dick" Moore Chair in Finance and Economic Development), Mario Maggi (Faculty of Economics - University of Pavia, Italy), and Hugo Benitez Silva (Economics Department, State University of New York at Stony Brook) for useful indications and remarks.

Author Contributions

Conceived and designed the experiments: DF. Performed the experiments: DF. Analyzed the data: DF. Contributed reagents/materials/analysis tools: DF. Wrote the paper: DF.

References

1. The Economist (2011) Food stamps - the struggle to eat. Available: http://www.economist.com/node/18958475.
2. Dynaski M, Rangarajan A, Decker P (1991) Forecasting food stamp program participation and benefits. Technical report, Prepared by Mathematica Policy Research, Inc. for U.S. Department of Agriculture, Food and Nutrition Service, August.
3. Klerman J, Danielson C (2011) Transformation of the supplemental nutrition assistance program. J Policy Anal Manag 30(4): 863–888.
4. Wilde P (2013) The new normal: The supplemental nutrition assistance program (snap). Am J Agr Econ 95(2): 325–331.
5. Albert V, Barth R (1996) Predicting growth in child abuse and neglect reports in urban, suburban, and rural counties. Soc Scie Rev 70(1): 58–82.
6. Opitz W, Nelson H (1996) Short-term, population-based forecasting in the public sector: A dynamic caseload simulation model. Popul Res Policy Rev 15(5/6): 549–563.
7. Conte M, Levy D, Shahrokh F, Staveley J, Thompson S (1998) Economic determinants of income maintenance programs: The maryland forecasting model. J Policy Model 20(4): 461–481.
8. Grogger J (2007) Markov forecasting methods for welfare caseloads. Child Youth Serv Rev 29(7): 900–911.
9. Gurmu S, Smith W (2008) Estimating and forecasting welfare caseloads. In: Sun J, Lynch T, editors, Government Budget Forecasting: Theory and Practice, Auerbach Publications, Boca Raton, Florida. pp. 188–222.
10. Lazariu V, Chengxuan Y, Gundersen C (2011) Forecasting women, infants, and children caseloads: A comparison of vector autoregression and autoregressive integrated moving average approaches. Contemp Econ Policy 29(1): 46–55.
11. Hayashi M (2012) Forecasting welfare caseloads: The case of the japanese public assistance program. Technical report, IPSS Discussion Paper Series No.2011-E03.
12. Choi H, Varian H (2012) Predicting the present with google trends. Econ Rec 88: 2-9.
13. Askitas N, Zimmermann K (2009) Google econometrics and unemployment forecasting. Appl Econ Quart 55: 107–120.
14. DAmuri F (2009) Predicting unemployment in short samples with internet job search query data. Technical report, MPRA working paper n. 1840.
15. Suhoy T (2009) Query indices and a 2008 downturn. Discussion Paper 6, Bank of Israel.
16. D'Amuri F, Marcucci J (2013) The predictive power of google searches in forecasting unemployment. Temi di discussione (Economic working papers) 891, Bank of Italy.
17. Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, et al. (2009) Detecting influenza epidemics using search engine query data. Nature 457: 1012–1014.
18. Kang M, Zhong H, He J, Rutherford S, Yang F (2013) Using google trends for influenza surveillance in south china. PLoS ONE 8: e55205.
19. Cho S, Sohn CH, Jo MW, Shin SY, Lee JH, et al. (2013) Correlation between national influenza surveillance data and google trends in south korea. PLoS ONE 8: e81422.
20. Da Z, Engelberg J, Pengjie G (2011) In search of attention. J Finance 5: 1461–1499.
21. Fantazzini D, Fomichev N (2014) Forecasting the real price of oil using online search data. International Journal of Computational Economics and Econometrics forthcoming.
22. Edelman B (2012) Using internet data for economic research. J Econ Perspect 26(2): 189–206.
23. Preis T, Moat H, Stanley H, Bishop S (2012) Quantifying the advantage of looking forward. Scientific Reports 2(350): DOI:10.1038/srep00350.
24. Preis T, Moat H, Stanley H (2013) Quantifying trading behavior in financial markets using google trends. Scientific Reports 3(1684): DOI:10.1038/srep01684.
25. Kristoufek L (2013) Can google trends search queries contribute to risk diversification? Scientific Reports 3(2713): DOI:10.1038/srep02713.
26. Kristoufek L (2013) Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. Scientific Reports 3(3415): DOI:10.1038/srep03415 (2013).
27. Bartlett S, Burstein N, Hamilton W (2004) Food stamp program access study final report. Technical report, Washington, DC: U.S. Department of Agriculture, Economic Research Service., Available: http://www.myfoodstamps.org/pdf files/ProgAccess.pdf. Accessed 2013 Jan 15.
28. United States General Accounting Office (1999) Food stamp program: Various factors have led to declining participation. Available: http://www.gao.gov/archive/1999/rc99185.pdf. Accessed 2013 Jan 15, GAO/RCED-99-185. Washington, DC.
29. Sa-ngasoongsong A, Bukkapatnam STS, Kim J, Iyer PS, Suresh RP (2012) Multi-step sales forecasting in automotive industry based on structural relationship identification. International Journal of Production Economics 140: 875–887.
30. Hall S, Henry S, Greenslade J (2002) On the identification of cointegrated systems in small samples: A modelling strategy with an application to uk. J Econ Dyn Control 26: 1517–1537.
31. Ziliak J, Gundersen C, Figlio D (2003) Food stamp caseloads over the business cycle. South Econ J 69: 903–919.
32. Lee L, Strazicich M (2003) Minimum lagrange multiplier unit root test with two structural breaks. Rev Econ and Stat 85: 1082–1089.
33. Aparicio F, Escribano A, Garcia A (2006) Range unit-root (rur) tests: Robust against nonlinearities, error distributions, structural breaks and outliers. Journal of Time Series Analysis 27: 545–576.
34. Franses P, Paap R (2004) Periodic time series models. Oxford: Oxford University Press.
35. Boswijk H, Franses P (1996) Unit roots in periodic autoregressions. Journal of Time Series Analysis 17: 221–245.
36. Proietti T (2003) Forecasting the us unemployment rate. Computational Statistics and Data Analysis 42: 451–476.
37. Toda H, Yamamoto T (1995) Statistical inference in vector autoregressions with possibly integrated processes. J Econometrics 66(12): 225–250.
38. Johansen S (1995) Likelihood-based inference in cointegrated Vector autoregressive models. Oxford: Oxford University Press.
39. Johansen S (2006) Cointegration: a survey. In: Mills T, Patterson K, editors, Palgrave Handbook of Econometrics: Volume 1, Econometric Theory, Basingstoke, UK and New York, USA: Palgrave MacMillan. pp. 540–577.
40. Taylor M, Sarno L (1998) The behaviour of real exchange rates during the post-bretton woods period. Journal of International Economics 46(2): 281–312.
41. Österholm P (2004) Killing four unit root birds in the us economy with three panel unit root test stones. Applied Economics Letters 11(4): 213–216.
42. Gregory A, Hansen B (1996) Residual-based tests for cointegration in models with regime shifts. J Econometrics 70: 99–126.
43. Hatemi J (2008) Tests for cointegration with two unknown regime shifts with an application to financial market integration. Empir Econ 35: 497–505.
44. Johansen S, Moscow R, Nielsen B (2000) Cointegration analysis in the presence of structural breaks in the deterministic trend. Economet J 3: 216–249.
45. Engle R, Granger C (1987) Co-integration and error correction: Representation. Econometrica 55: 251–276.
46. Ljung G, Box G (1979) On a measure of lack of fit in time series models. Biometrika 66: 265–270.

47. Engle R (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of u.k. inflation. Econometrica 50: 987–1008.
48. Jarque C, Bera A (1987) A test for normality of observations and regression residuals. Int Stat Rev 55: 163–172.
49. Ramsey J (1969) Tests for specification errors in classical linear least-squares regression analysis. J Roy Stat Soc B 31: 350–371.
50. Brock W, Dechert D, Sheinkman J, LeBaron B (1996) A test for independence based on the correlation dimension. Economet Rev 15(3): 197–235.
51. Brown R, Durbin J, Evans J (1975) Techniques for testing the constancy of regression relationships over time. J Roy Stat Soc B 37: 149–163.
52. Ploberger W, Kramer W (1992) The cusum test with ols residuals. Econometrica 60(2): 271–285.
53. Chu C, Hornik K, Kuan C (1995) Mosum tests for parameter constancy. Biometrika 82: 603–617.
54. Zeileis A, Kleiber C, Kramer W, Hornik K (2003) Testing and dating of structural changes in practice. Comput Stat Data An 44(12): 109–123.
55. Zeileis A (2005) A unified approach to structural change tests based on ml scores, f statistics, and ols residuals. Economet Rev 24(4): 445–466.
56. Chow G (1960) Tests of equality between sets of coefficients in two linear regressions. Econometrica 28: 591–605.
57. Andrews D (1993) Tests for parameter instability and structural change with unknown change point. Econometrica 61: 821–856.
58. Andrews D, Ploberger K (1994) Optimal tests when a nuisance parameter is present only under the alternative.. Econometrica 62: 1383–1414.
59. Zeileis A, Leisch F, Hornik K, Kleiber C (2002) strucchange: An r package for testing for structural change in linear regression models. J Stat Softw 7(2): 1–38.
60. Zeileis A, Shah A, Patnaik I (2010) Testing, monitoring, and dating structural changes in exchange rate regimes. Comput Stat Data An, 54(6): 1696–1706.
61. Bai J, Perron P (2003) Computation and analysis of multiple structural change models. J Appl Econom 18: 1–22.
62. Liu J, Wu S, Zidek J (1997) On segmented multivariate regression. Stat Sinica 7: 497–525.
63. Pesaran M, Timmermann A (2007) Selection of estimation window in the presence of breaks. J Econometrics 137(1): 134–161.
64. Koop G, Korobilis D (2010) Bayesian multivariate time series methods for empirical macroeconomics. Foundations and Trends in Econometrics 3: 267–358.
65. Doan T, Litterman R, Sims C (1984) Forecasting and conditional projection using realistic prior distributions. Econometric Reviews 3: 1–100.
66. Litterman R (1986) Forecasting with bayesian vector autoregressions: Five years of experience. Journal of Business and Economic Statistics 4: 25–38.
67. Tong H (1990) Non-linear time series: a dynamical system approach. Oxford: Oxford University Press.
68. Venables W, Ripley B (2003) Modern Applied Statistics with S. Springer, 4th edition.
69. Franses P, Dijk V (2000) Nonlinear Time Series Models in Empirical Finance. Cambridge University Press.
70. Wood S (2006) Generalized additive models: an introduction with R. Boca Raton: Chapman and Hall/CRC.
71. Hansen P, Lunde A, Nason J (2011) The model confidence set. Econometrica 79: 453–497.
72. Rossi E, Fantazzini D (2014) Long memory and periodicity in intraday volatility. Journal of Financial Econometrics forthcoming.
73. Mabli J, Castner S, Godfrey T, Foran P (2011) Dynamics of supplemental nutrition assistance program participation in the mid-2000s. Technical report, Princeton, NJ: Mathematica Policy Research, Inc.
74. Theil H (1961) Economic Forecasts and Policy. Amsterdam: North Holland Publishing Company.
75. Naik G, Leuthold R (1986) A note on qualitative forecast evaluation. Am J Agr Econ 68: 721–726.

Appendix 11 "Proposed coal power plants and coal-to-liquids plants in the US: Which ones survive and why?"

Fantazzini, D., & Maggi, M. (2015). Proposed coal power plants and coal-to-liquids plants in the US: Which ones survive and why?. *Energy Strategy Reviews*, 7, 9-17. DOI: https://doi.org/10.1016/j.esr.2014.11.005

Permisson to post the Publisher's version/PDF: according to https://www.elsevier.com/about/policies/sharing "*Theses and dissertations which contain embedded Published Journal Article (PJA) as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.*"

Energy Strategy Reviews 7 (2014) 9–17

Contents lists available at ScienceDirect

Energy Strategy Reviews

journal homepage: www.ees.elsevier.com/esr

ANALYSIS

Proposed coal power plants and coal-to-liquids plants in the US: Which ones survive and why?

Dean Fantazzini [a,*], Mario Maggi [b]

[a] Moscow School of Economics, Moscow State University, Leninskie Gory, Building 61/1 119992, Russia
[b] Department of Economics and Management, University of Pavia, via S. Felice, 5, Pavia 27040, Italy

ARTICLE INFO

Article history:
Received 1 July 2014
Received in revised form
18 November 2014
Accepted 26 November 2014
Available online

Keywords:
Coal
Coal plants
Coal-to-liquids
Logit
Probit
Training
Validation
Forecasting
Model confidence set
Google
Google trends
Second great contraction
Global financial crisis

ABSTRACT

The increase of oil and natural gas prices since the year 2000 stimulated the planning and construction of new coal-fired electricity generating plants and coal-to-liquids (CTL) plants in the US. However, many of these projects have been canceled or abandoned since 2007. Using a set of 145 proposed coal power plants and 25 CTL plants, the determinants that influence the decision to abandon a project or to proceed with it are examined using binary data models and 20 regressors. In the case of coal power plants, the number of searches performed on Google relating to coal power plants, the project duration and the prices of alternative fuels for electricity generation are found to be statistically significant at the 5% level. As for CTL plants, the political affiliation of the state governor is the only variable significant at the 5% level across several model specifications. An out-of-sample exercise confirms these findings. These results also hold with robustness checks considering alternative Google search keywords, the potential effects of the recession between 2008 and 2009 and the inclusion of the two dimensions of the Dynamic-Weighted Nominate (DWN) database.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

The first decade of the 21st century witnessed a large increase in oil prices mainly due to the growing demand by China and India, as well as to a growing difficulty to increase oil production worldwide with the notable exception of North America (see Ref. [1], for a recent review). Similarly, US natural gas prices followed a rising trend reaching the level of 13 $/MMBtu in June 2008. The rise in oil and gas prices coincided with increasing power demand in the US. To counter rising fuel cost, coal was a logical choice for power generation, stimulating the planning and/or construction of almost 150 coal-fired electricity generating plants by 2007 [2]. Several coal-to-liquids (CTL hereafter) plants were also proposed (see Ref. [3] for a recent review of hydrocarbon liquefaction as a peak oil mitigation strategy). Since 2007–2008 the energy landscape has changed substantially: the advent of shale gas has reduced considerably the price of natural gas in the US reaching a low of 1.9 $/MMBtu in April 2012. Meanwhile, the construction cost for coal plants has increased considerably but US coal prices remain relatively low (see Refs. [4,5] for recent reviews). Since 2011 the US Environmental Protection Agency (EPA) has began regulating greenhouse gases from mobile and stationary sources of air pollution under the Clean Air Act. There has been an increased awareness of the health risks posed by power plant pollution (as showed by Google data, more below). All this has led to more than 100 coal plants being canceled or abandoned by 2013. This estimate is based on the Sierra Club database [6] and the Coal-Swarm database [7]). The Energy Information Administration (EIA) expects that very few new coal plants will be built through 2040 [8].

Although coal is still the main source for US electricity power production, coal plants are aging. In 2011, the capacity weighted average age of coal-fired plants was 36 years, whereas it was only 18 for natural

* Corresponding author.
E-mail addresses: fantazzini@mse-msu.ru (D. Fantazzini), maggma@eco.unipv.it (M. Maggi).

http://dx.doi.org/10.1016/j.esr.2014.11.005
2211-467X/© 2014 Elsevier Ltd. All rights reserved.

Table 1
Capacity weighted distribution of electricity power production plants by fuel. 2011 data from http://www.eia.gov.

Fuel type	Coal	Natural gas	Petroleum
Average size (MW)	245.54	85.65	15.39
Average age (years)	36.34	17.88	35.16
25% built before	1967	1981	1970
50% built before	1974	2001	1972
75% built before	1981	2003	1978
CO_2/capacity (Million Metric Tons/MWh)	0.9931	0.3972	0.8689

gas-fired plants (and 35 for oil-fired plants[1]), see Table 1. Refitting these coal plants to comply with the recent stricter emission standards is very expensive so many of them face retirement in the coming years [5].

Given this background, we analyze the main determinants that influenced the decision to abandon or to proceed with a coal project using a dataset of 145 coal power plants projects and 25 CTL plants, between 2004 and 2013, from the Coal-Swarm database [7], and binary data models.

Prior knowledge of the variables influencing the viability of a coal plant project is fundamental for successful strategy and policy making. To our knowledge, this is the first study that analyzes these variables after the advent of US shale gas and the global economic crisis in 2008–2009. Our findings are not limited to coal plants but also include CTL plants.

A vast body of the literature has found that the public attitude toward the location of environmentally hazardous facilities is a major determinant of siting costs, which can increase quickly when the local community agreement is missing (see Refs. [9,10] for extensive reviews). We use Google search data to measure public attitudes towards coal plants and the associated environmental issues: a tool called Google Trends provides information about users' relative interest for a particular search query in a given geographic region and at given time (the data are available on a weekly or even a daily basis). In recent years, researchers worldwide have started to use online search data for forecasting purposes (see Refs. [11–17] for some recent applications).

The predictive power of our binary data models is then tested by means of an out-of-sample comparison. The models differ along three dimensions: (i) the variables adopted; (ii) the econometric specification; and (iii) the data transformation (either in logs or in levels). A series of robustness checks is also performed to verify that our previous results hold also with alternative data setups: (i) a dataset with alternative keywords for Google search; (ii) time dummies to evaluate the effect on model estimates of the global financial crisis in 2008 and 2009; (iii) a dataset that includes as additional regressors the two dimensions of the Dynamic-Weighted Nominate (DWN) database developed by the political scientists Poole and Rosenthal in the early 1980s to analyze legislative roll-call voting behavior in the US congress, see Refs. [18,19].

The paper is organized as follows. Section 2 describes the data and methods used in our work while the empirical analysis is performed in Section 3. Robustness checks are discussed in Section 4, while Section 5 includes a brief conclusion.

2. Data and methods

2.1. Data

The National Energy Technology Laboratory (NETL), a division of the Department of Energy, maintained a database of all new projects of coal-fired electricity generating plants until May 2007. Since then, the *Coal Issues Portal* on SourceWatch (a project of CoalSwarm and the Center for Media and Democracy) has maintained a dataset of the proposed coal plants in the US and their latest status [7].[2] We separated the variable "status" into two groups: one collecting all plants that are active/upcoming/operating and another group with all plants that were canceled/abandoned or have an uncertain status.[3]

The Coal Issues portal contains some information about the coal projects, like the US state location and, in some cases, also the total capacity (in MW for power plants and bbl/day for CTL), but this information was not sufficient for the scope of our analysis and was augmented by an extensive online search for each coal project. This search was not successful for several plants, for which budget costs, capacity, carbon dioxide (CO_2) emissions, project beginning year and project duration were not available. The initial dataset was filtered and the final dataset consisted of 145 coal power plant projects and 25 CTL plant projects, observed between 2004 and2013.[4] The dataset of coal power plants projects consists of 97 plants that were canceled/abandoned and 48 plants that are active/operating/upcoming for a total of 574 yearly data samples. The dataset of CTL projects consists of 17 plants that were cancelled/abandoned and 8 plants that are active/upcoming for a total of 94 yearly data samples. The (few) projects that were either operative or cancelled before 2004 were omitted since those early projects had very different economics from subsequent ones (see Refs. [1,3,20]).

The literature has identified four main groups of variables that influence the plant location choice. First Coase [21], suggested that site-specific environmental externalities should be the main determinants of location choices: a profit-maximizing firm will try to find an agreement with the community that causes the least damage, all other things being equal. Hamilton (1993) [22], Hamilton (1995) [23] and Jenkins et al. [24] questioned this hypothesis and advanced the idea that local community's public opinion can influence externality costs: communities that show strong opposition are less likely to host a plant or any environmentally hazardous facility. Therefore, a model trying to explain the location of a (coal) plant should consider a group of "voice" indicators. A third group of variables includes traditional industrial location factors like infrastructure, construction and labor costs, see Refs. [10,25–28]. More recently, given the falling prices of renewables and natural gas, several authors have started comparing the economics of these alternative sources of electricity generation with the economics of coal plants to determine the best choice and location, see Refs. [4,5,29–31]. Table 2 illustrates the regressors that we used to explain the status of a coal plant project.

We used the state population in millions and the CO_2 output in tons to measure the external costs a state can suffer given that the larger the population and the CO_2 produced the larger the perception of the expected environmental damage (see Refs. [10,22,32]).[5]

Four indicators were used to represent the awareness of local residents and their ability to pay for environmental quality: the median household income, the labor force participation, the unemployment rate and the Google Index (GI) for the keyword "jobs" (remark that D'Amuri and Marcucci [16] found this GI to be the best predictor for the US unemployment rate). The GI is computed as the ratio of the search queries for a specific keyword (or group of keywords) relative to the total number

[1] The old age of oil-fired plants is also due to the fact that in US oil produces a small and decreasing portion of electricity production.

[2] The NETL database is no longer available but it is included in the CoalSwarm database.

[3] An online search allowed us to find that all plants with an uncertain status were either cancelled or abandoned. They had no related news for years.

[4] The names of these plants are reported in Tables 1–2 in the Technical Appendix accompanying this paper and is posted on the authors' websites.

[5] We tried population density in place of the population data, as done by Garrone and Groppi [10], but this resulted in worse in-sample results, models' residuals and out-of-sample results; we used the population data instead.

of searches performed in the selected region at a given point of time. The result was then standardized between 0 and 100 (where the standardization is done over the whole time period and all searches).

To measure awareness and the ability to organize protests against coal plant projects as well as to ask for compensation (the so-called "voice" factors), we used the GI for the keyword "coal," the GI for the keywords "coal plant + coal power," the GI for the keywords "coal-to-liquids + ctl coal" and the GI for the keyword "pollution." The analysis of Google data showed that several searches for the previous keywords included and/or were related also to "legal action," "protest," "stop," etc., which shows that it is not easy to separate awareness from voice factors.[6] Following Ansolabehere and Konisky [9], we also used the political affiliation of the state governor as a voice factor.

Five indicators were used to consider traditional industrial location factors: the plant cost estimate, the plant capacity, the coal price, the available rail miles (which is important for coal transportation) and the average electricity price. The latter can also be interpreted as a measure of (past) profitability. The plant cost estimates were updated each year using the Chemical Engineering Plant Cost Index (CEPCI), which is a dimensionless number used to update the capital cost required to build a chemical plant from a past date to a later time. This index is widely accepted and consists of subcomponents dealing with equipment, labor costs, buildings, engineering, supervision and other parameters affecting costs. Kreutz et al. [33] provide a comparison of the CEPCI with the Marshall and Swift index, the US GDP deflator and the Handy-Whitman Total Plant-All Steam Generation Index, while Höök et al. [3] used this index to compute the economics of coal-to-liquids and gas-to-liquids plants.

The competition with alternative energy sources was measured by using the average levelized long-term wind price, the average price of residential and commercial solar photovoltaics, and the Henry Hub natural gas price.

We also considered the number of years since the coal plant project had started: we noted that the more time the project spends in its planning phase the less probable will be its full development. We found two reasons for this phenomenon: strong cost escalations and a prolonged legal battle between the local communities and the plant developers. These two reasons were interconnected: the legal battle delayed the coal project to such an extent that the new price environment was no longer profitable due to cost escalations and falling prices of energy alternatives (see the Coal Issues Portal and the history of each coal plant reported there). This phenomenon also confirms again that separating the different indicators in clear-cut categories is not always possible.

All data were collected for each US state for the period January 2004 through December 2013. The data had yearly frequency or were converted to a yearly frequency to match the coal plants data. All data were transformed into logs, except for the *duration* indicator and the binary variable *governor*. In Section 3.2, devoted to the out-of-sample forecasting analysis, we considered a wide set of models also including models with data in levels that is without any transformation.

2.2. Methods: collinearity, stationarity and econometric analysis

We computed the correlation among regressors (see Figs. 1–2 in the Technical Appendix) as well as the Variance Inflation Factors[7] for each regressor (see Tables 3–4 in the Technical Appendix), where we differentiated between coal power plants and CTL plants.[8] Classical "rules of thumbs" to get rid of collinearity are to eliminate those variables with a VIF higher than 10 or to eliminate one of the two variables with a correlation higher than 0.7–0.8 (in absolute value). Given that simply eliminating variables may not be a good solution, as shown by O'Brien [34] and Dormann et al. [35], we followed a less aggressive approach: when two variables had a correlation coefficient (in absolute value) higher than 0.8 we took the first one and the ratio between the first and the second one. The following ratios were considered: CO_2/capacity in place of CO_2 output (for coal power plants only); cost/capacity in the place of cost; rail/population in place of rail miles; solar price/natural gas price in place of solar price (for coal power plants only); solar price/wind price in place of solar price (for CTL plants only); the GI for the keywords "coal-to-liquids + ctl" divided by the natural gas price, in place of the initial GI (for CTL plants only). The last one is the only ratio without an immediate economic meaning. However, considering that hydrocarbon liquefaction can be implemented either using coal or natural gas, this ratio can be roughly interpreted as a ratio between the interest for coal-to-liquids plants and gas-to-liquids plants.

The next step was to check whether our data are stationary. Given the moderate size of our dataset in case of coal power plants and the small size for CTL plants, we employed a battery of panel unit root tests: the test by Levin et al. [36], the test by Im et al. [37], the Fisher-type tests using ADF and PP tests by Maddala and Wu [38] and Choi [39], and the Hadri test [40]. The results in Table 5 in the Technical Appendix show that our data are stationary.

Our dataset consists of the binary dependent variable "status" Y_{it}, which indicates whether a coal project is active/upcoming ($Y_{it} = 0$) or abandoned/canceled ($Y_{it} = 1$), for observation i, ($i = 1, \ldots, n$) and time t, ($t = 1, \ldots, T$), and of the $p \times 1$ vector of regressors X_{it}. We are interested in predicting the expectation of the response variable as a function of the regressors. The expectation of a simple binary response is just the probability that the response is 1:

$$E(Y_{it}|X_{it}) = \pi(Y_{it} = 1|X_{it}). \tag{1}$$

To model this expectation we use logit and probit models:

$$g\{\pi(Y_{it} = 1|X_{it})\} = \beta X_{it} = \nu_i, \tag{2}$$

where ν_i is referred to as the *linear predictor*, $g(\cdot)$ is the logit or probit *link function*, while the distribution of the response given the regressors is always specified as a Bernoulli distribution (see Ref. [41] for more details)[9].

2.3. Model evaluation

To compare the alternative models, we will report the standard Akaike and Schwartz information criteria (AIC and SIC, respectively). We will also compute the Ljung–Box [42] test statistic for testing the absence of autocorrelation up to order k in the models' standardized residuals and squared residuals, as well as the BDS test by Brock et al. [43], to test whether the standardized residuals are independent and identically distributed. This test is robust against a variety of possible deviations from independence, including linear dependence, nonlinear dependence, or chaos.

We will also report the Area Under the Receiver Operating Characteristic (ROC) curve (AUC) by Metz and Kronman [44], Goin [45] and

[6] Alternative keywords for Google search with smaller search volumes will be analyzed in Section 4, Robustness Checks.

[7] Variance Inflation Factors (VIF) are used to measure the degree of collinearity among the regressors in a linear equation. They can be computed by dividing the variance of a coefficient estimate with all the other regressors included by the variance of the same coefficient estimated from an equation with only that regressor and a constant.

[8] The CO_2 output was not considered for CTL plants because this data was available for very few plants.

[9] Random intercepts and random coefficients binary models were not employed because either they did not converge numerically or their random component showed variances that were not statistically different from zero. In the following discussion, we will only consider simple (pooled) logit and probit models.

Hanley and McNeil [46] for all competing models. The ROC curve is obtained by plotting, for any probability threshold, the proportion of correct predictions that a project is abandoned/canceled (*y*-axis), with respect to the proportion of incorrect predictions that a project is active/upcoming (*x*-axis). In terms of model comparison, the best curve is the one that is leftmost, the ideal one coinciding with the *y*-axis (see Refs. [47,48] for a recent application). The AUC lies between zero and one and the closer it is to one the more accurate the classifier is.

Although the AUC is one of the most common tools to measure classifier predictive performance, it has some drawbacks, as recently reviewed by Figini and Maggi [49] and references therein. We also computed the *Model Confidence Set* (MCS), proposed by Hansen et al. [50] and extended by Figini and Maggi [49] to binary models, to assess the prediction power of the competing models (see also Refs. [17,51] for recent applications in financial forecasting). The MCS is a set containing the best forecasting models at a given confidence level. Following Hansen et al. [50], the MCS procedure selects the best model and computes the probability that other models are undistinguishable from the best one using an evaluation rule based on a *loss function*. The more the data are informative the smaller the MCS will be. We computed the MCS following the procedure set up by Hansen et al. [50], adopting the χ^2 test, at 10% confidence level, for the model elimination rule and using different loss functions.[10] The full technical details and the loss functions used in our analysis are reported in Technical Appendix.

3. Results

3.1. In-sample analysis

Tables 3–4 report the results for coal power plants and CTL plants respectively, where the left columns show the results with all the regressors, while the right columns the restricted models with only the regressors that were significant at the 5% level (for coal power plants) and at the 10% level (for CTL plants).[11]

	Logit		Probit		Logit restricted		Probit restricted	
	Coef.	*p-value*	*Coef.*	*p-value*	*Coef.*	*p-value*	*Coef.*	*p-value*
CO_2/Capacity	0.95	0.21	0.48	0.26				
GI (coal)	0.09	0.62	0.06	0.57				
Coal price	**−1.32**	**0.04**	**−0.71**	**0.04**	**−1.59**	**0.00**	**−0.86**	**0.00**
GI (coal plant + coal power)	**13.17**	**0.00**	**7.35**	**0.00**	**14.18**	**0.00**	**7.85**	**0.00**
Cost/Capacity	−0.53	0.55	−0.28	0.59				
Duration	**0.22**	**0.03**	**0.13**	**0.03**	**0.22**	**0.02**	**0.12**	**0.02**
Electricity	−0.01	0.99	−0.01	0.98				
Governor	−0.29	0.26	−0.17	0.23				
Income	1.36	0.40	0.84	0.37				
GI (jobs)	1.06	0.38	0.63	0.36				
LFP	−3.22	0.37	−1.96	0.34				
NG price	**−10.32**	**0.00**	**−5.72**	**0.00**	**−10.22**	**0.00**	**−5.72**	**0.00**
GI (pollution)	0.02	0.94	0.02	0.88				
Population	−0.25	0.15	−0.14	0.16				
Rail/Population	−4.19	0.16	−2.25	0.16				
Solar price/NG price	**−14.78**	**0.00**	**−8.13**	**0.00**	**−15.34**	**0.00**	**−8.53**	**0.00**
Capacity (MW)	0.51	0.08	0.26	0.11				
UR	−1.26	0.08	−0.74	0.06				
Wind price	0.24	0.80	0.03	0.95				
Constant	−16.63	0.22	−9.51	0.21	**−14.77**	**0.00**	**−8.20**	**0.00**
Information criteria and AUC								
AIC		516.52		515.44		501.22		**500.12**
SIC		603.58		602.50		527.33		**526.23**
AUC		71.02%		71.08%		67.08%		67.11%
Residual tests								
Ljung–Box (50) res. [*p*-val]		0.21		**0.03**		0.36		0.51
Ljung–Box (50) res.sq. [*p*-val]		0.29		0.34		**0.00**		0.50
BDS (dim = 2) [*p*-val]		0.09		0.33		0.14		0.09
BDS (dim = 6) [*p*-val]		0.56		0.91		0.56		0.57

In case of coal power plants, as expected, the longer the planning period the higher the probability that the project will be abandoned/cancelled. The lower the price for natural gas and the lower the price for solar photovoltaics with respect to natural gas the higher the probability that the project will be abandoned. The higher the Google search volumes about coal plants and/or coal power the higher the probability the coal project will be abandoned: an increasing number of people looking for information about coal plants on the web highlights a growing opposition to coal projects. An unexpected result is that a higher coal price will increase the probability that a coal plant will be fully developed. A possible explanation of this result could be that the strong commercial relationships between coal mining companies and coal power companies allow high coal prices to be economically viable. Given the very sketchy information about the business structure of the companies involved in coal projects (particularly for abandoned projects), we leave this issue as an interesting avenue for further research.

As for CTL plants only two regressors were found to be significant at the 10% level: the political affiliation of the state governor and the ratio between solar and wind prices. A republican governor will increase the likelihood that the coal plant will be built, whereas a lower price for solar photovoltaics with respect to wind price will increase the probability that the project will be abandoned. The importance of the governor political affiliation is not a surprise given the greater technical complexity, strong environmental impacts and the higher costs of CTL plants. These plants have very high CO_2 emissions, more than double the amount produced by the oil industry (see Refs. [52,53]), they are extremely water-intensive (see Ref. [3] and references therein), and the discharged water must be treated to avoid environmental harm (see Ref. [54]). The recent analysis in 3 highlights very poor economics and "*a strong risk for CTL plants to become financial black holes*" and helps explain why China has strongly slowed down the development of its CTL program (see Ref. [55]). A CTL project will likely succeed only with strong political support at the level of the local state government otherwise it will be better not to proceed further.

Probit models fared better than logit models showing lower information criteria and better residuals properties. Restricted models

[10] Other tests can be applied: for instance, the *F* statistic or other statistics built on the *t*-statistic that do not require the computation of the model covariance matrix. In our applications, the *F* statistic and other *t*-statistics delivered similar results to the χ^2. However, the *t*-statistics are much more demanding in terms of computing time and are convenient when the number of models is large relative to the sample, which is not our case.

[11] We used a higher probability level for CTL plants due to the small size of the dataset.

Table 2
Regressors: description and source.

Variables	Description	Sources
Externalities costs		
CO_2 (tons)	Carbon dioxide output in tons	Carbon Monitoring for Action (CARMA) database
Population	Population by US state in millions	U.S. Department of Commerce: Census Bureau
Awareness and ability to pay for environmental quality		
Income	Median Household Income by US state	U.S. Department of Commerce: Census Bureau
LFP	Labor Force Participation by US state	U.S. Department of Labor: Bureau of Labor Statistics
UR	Unemployment Rate by US state	U.S. Department of Labor: Bureau of Labor Statistics
GI (JOBS)	Google index for the keyword "jobs"	Google trends
Awareness and voice factors		
GI (coal)	Google index for the keyword "coal"	Google trends
GI (coal power + coal plant)	Google index for the keywords "coal power + coal plant"	Google trends
GI (coal-to-liquids + ctl coal)	Google index for the keywords "coal-to-liquids + ctl coal"	Google trends
GI (pollution)	Google index for the keyword "pollution"	Google trends
Governor	Binary variable that is 1 if Republican and 0 otherwise	www.rulers.org
Traditional industrial location factors		
cost	Plant cost estimate (billion $)	CMD/Google search
Coal price	US Central Appalachian coal spot price ($/ton)	BP Statistical Review of World Energy 2013/US EIA
Rail	Rail miles by US state	Association of American Railroads
Capacity (MW)	Plant capacity expressed in MW for coal power	NETL-US DOE/CMD/Google search
Capacity (BBL/day)	Plant capacity expressed in bbl/day for CTL plants	
Electricity	Average electricity price by US state ($/Kwh)	US Energy Information Administration (EIA)
Economics of alternative energy sources		
Wind price	Average levelized long-term wind power purchase agreement prices ($/Mwh)	US Department of Energy/Energy Analysis and Environmental Impacts Department – Lawrence Berkeley National Laboratory
Solar price	Installed price of residential and commercial solar photovoltaics ($/W)	US Department of Energy (DOE)/Lawrence Berkeley National Laboratory
NG price	US Henry Hub natural gas price ($/MmBtu)	BP Statistical Review of World Energy 2013/US EIA
Additional indicators		
Duration	The number of years that has passed at time *t* since the project started	The National Energy Technology Laboratory (NETL) The Center for Media and Democracy (CMD) Google search

Table 3
Coal power plants: model estimation results. Smallest information criteria and *p*-values smaller than 5% are reported in bold font.

	Logit		Probit		Logit restricted		Probit restricted	
	Coef.	*p-value*	*Coef.*	*p-value*	*Coef.*	*p-value*	*Coef.*	*p-value*
CO_2/Capacity	0.95	0.21	0.48	0.26				
GI (coal)	0.09	0.62	0.06	0.57				
Coal price	**−1.32**	**0.04**	**−0.71**	**0.04**	**−1.59**	**0.00**	**−0.86**	**0.00**
GI (coal plant + coal power)	**13.17**	**0.00**	**7.35**	**0.00**	**14.18**	**0.00**	**7.85**	**0.00**
Cost/Capacity	−0.53	0.55	−0.28	0.59				
Duration	**0.22**	**0.03**	**0.13**	**0.03**	**0.22**	**0.02**	**0.12**	**0.02**
Electricity	−0.01	0.99	−0.01	0.98				
Governor	−0.29	0.26	−0.17	0.23				
Income	1.36	0.40	0.84	0.37				
GI (jobs)	1.06	0.38	0.63	0.36				
LFP	−3.22	0.37	−1.96	0.34				
NG price	**−10.32**	**0.00**	**−5.72**	**0.00**	**−10.22**	**0.00**	**−5.72**	**0.00**
GI (pollution)	0.02	0.94	0.02	0.88				
Population	−0.25	0.15	−0.14	0.16				
Rail/Population	−4.19	0.16	−2.25	0.16				
Solar price/NG price	**−14.78**	**0.00**	**−8.13**	**0.00**	**−15.34**	**0.00**	**−8.53**	**0.00**
Capacity (MW)	0.51	0.08	0.26	0.11				
UR	−1.26	0.08	−0.74	0.06				
Wind price	0.24	0.80	0.03	0.95				
Constant	−16.63	0.22	−9.51	0.21	**−14.77**	**0.00**	**−8.20**	**0.00**
Information criteria and AUC								
AIC		516.52		515.44		501.22		**500.12**
SIC		603.58		602.50		527.33		**526.23**
AUC		71.02%		71.08%		67.08%		67.11%
Residual tests								
Ljung–Box (50) res. [*p*-val]		0.21		**0.03**		0.36		0.51
Ljung–Box (50) res.sq. [*p*-val]		0.29		0.34		**0.00**		0.50
BDS (dim = 2) [*p*-val]		0.09		0.33		0.14		0.09
BDS (dim = 6) [*p*-val]		0.56		0.91		0.56		0.57

Table 4
Coal-to-liquids plants: model estimation results. Smallest information criteria and *p*-values smaller than 5% are reported in bold font.

	Logit		Probit		Logit restricted		Probit restricted	
	Coef.	*p-value*	*Coef.*	*p-value*	*Coef.*	*p-value*	*Coef.*	*p-value*
GI (coal)	1.59	0.31	0.92	0.30				
Coal price	0.31	0.88	0.18	0.87				
GI (coal-to-liquids + ctl)/ng price)	0.57	0.28	0.32	0.29				
Cost/Capacity	−1.83	0.40	−1.08	0.37				
Duration	0.45	0.25	0.26	0.18				
Electricity	3.20	0.28	1.86	0.18				
Governor	**−2.28**	**0.02**	**−1.32**	**0.01**	**−1.27**	**0.03**	**−0.76**	**0.02**
Income	0.81	0.87	0.43	0.86				
GI (jobs)	−4.70	0.38	−2.74	0.31				
LFP	−8.11	0.41	−4.69	0.36				
NG price	−0.22	0.92	0.00	1.00				
GI (pollution)	−0.54	0.37	−0.34	0.32				
Population	1.32	0.14	0.74	0.10				
Rail/Population	2.20	0.86	1.12	0.85				
Solar price/Wind price	*−54.47*	*0.08*	*−31.10*	*0.07*	**−20.32**	**0.03**	**−11.80**	**0.02**
Capacity (BBL/day)	0.06	0.88	0.04	0.85				
UR	−3.91	0.12	−2.21	0.12				
Wind price	1.83	0.49	0.86	0.54				
Constant	42.34	0.28	25.55	0.21	*8.60*	*0.06*	**4.98**	**0.04**
Information criteria and AUC								
AIC		109.51		108.99		87.44		**87.07**
SIC		157.84		157.31		95.07		**94.70**
AUC		79.83%		80.21%		70.51%		70.51%
Residual tests								
Ljung–Box (50) res. [*p*-val]		0.42		0.62		0.20		0.47
Ljung–Box (50) res.sq. [*p*-val]		0.21		0.91		0.06		0.44
BDS (dim = 2) [*p*-val]		0.41		0.53		**0.00**		0.85
BDS (dim = 6) [*p*-val]		0.54		0.95		0.07		0.93

p-values between 5% and 10% are reported in italics.

showed lower information criteria but full models had higher AUC values.

3.2. Out-of-sample forecasting analysis

To better evaluate the predictive performance for each model we also implemented a cross-validation procedure. We divided our dataset into two parts of equal size: the first one was used as the training set while the second one as the validation set. Similarly to Ref. [56], we compared a set of alternative models whose characteristics are reported in Table 5. We considered both logit and probit models, models with all the regressors, as well as restricted models with only significant parameters at the 5% level; models with data in logs and models with data in levels; models without GIs and models with only GIs. In case of CTL plants, due to the very small sample size of the training and validation sets, we only considered restricted models with GIs only and without GIs.

The estimated AUC for all previous models are reported in Table 6. The restricted probit model with data in logs was the best for coal power plants projects, while the probit model with data in logs and no GIs was the best for CTL projects, thus confirming previous in-sample results discussed in Section 3.1.

Table 5
List of forecasting models.

Coal power plants				Coal-to-liquids plants			
Model	*Data transformation*	*All regressors/restricted model*	*Google data*	*Model*	*Data transformation*	*All regressors/restricted model*	*Google data*
Logit	log	All	Yes	Logit	log	Restricted	No
Probit	log	All	Yes	Probit	log	Restricted	No
Logit	log	Restricted	No	Logit	log	Restricted	Only
Probit	log	Restricted	No	Probit	log	Restricted	Only
Logit	log	Restricted	No	Logit	Levels	Restricted	No
Probit	log	Restricted	No	Probit	Levels	Restricted	No
Logit	log	Restricted	Yes	Logit	Levels	Restricted	Only
Probit	log	Restricted	Yes	Probit	Levels	Restricted	Only
Logit	log	Restricted	Only				
Probit	log	Restricted	Only				
Logit	Levels	All	Yes				
Probit	Levels	All	Yes				
Logit	Levels	Restricted	Yes				
Probit	Levels	Restricted	Yes				
Logit	Levels	Restricted	No				
Probit	Levels	Restricted	No				
Logit	Levels	Restricted	No				
Probit	Levels	Restricted	No				
Logit	Levels	Restricted	Only				
Probit	Levels	Restricted	Only				

Table 6
A.U.C. for each forecasting model. The best model is reported in bold font.

Models: coal power plants	AUC	Models: coal-to-liquids	AUC
Logit log	59.48%	Logit log (no Google)	60.74%
Probit log	60.13%	**Probit log (no Google)**	**61.85%**
Logit log (no Google)	57.14%	Logit log (only Google)	52.04%
Probit log (no Google)	57.74%	Probit log (only Google)	52.78%
Logit log (no Google) restricted	58.23%	Logit levels (no Google)	60.37%
Probit log (no Google) restricted	58.25%	Probit levels (no Google)	59.07%
Logit log restricted	63.91%	Logit levels (only Google)	53.89%
Probit log restricted	**64.13%**	Probit levels (only Google)	55.74%
Logit log (only Google)	48.18%		
Probit log (only Google)	48.08%		
Logit levels	60.13%		
Probit levels	60.00%		
Logit levels restricted	62.35%		
Probit levels restricted	63.87%		
Logit levels (no Google)	57.46%		
Probit levels (no Google)	58.01%		
Logit levels (no Google) restricted	60.17%		
Probit levels (no Google) restricted	60.17%		
Logit levels (only Google)	48.11%		
Probit levels (only Google)	48.33%		

We then employed the MCS approach developed by Hansen et al. [50] and discussed in Section 2.2 to test for statistically significant differences in the forecast performances among the competing models. We recall from Section 2.2 that the MCS procedure will yield a set containing the best forecasting models at a given confidence level. The full results of the MCS procedure are reported in Table 6 in the Technical Appendix.

In case of coal power plants, the restricted probit model with data in logs is the model with the lowest loss for almost all loss functions considered, thus confirming the previous results. Moreover, models with Google data represent the majority of models included in the MCS, while models without Google data are seldom included, thus confirming the important information that this type of data can provide. As for CTL plants, the logit models with data in levels without Google data is the one that has the lowest loss across a spectrum of loss functions. Almost all models are now included in the MCS, which highlights that the validation set is not very informative (which was expected given its small size).

4. Robustness checks

To verify that our previous results hold also with alternative data setups, we performed a series of robustness checks: we considered alternative keywords for Google search and we evaluated the effect on our estimates of the global financial crisis in 2008 and 2009. Finally, we included as additional regressors the two dimensions of the Dynamic-Weighted Nominate (DWN) database developed by the political scientists Poole and Rosenthal in the early 1980s to analyze legislative roll-call voting behavior in the US congress, see Refs. [18,19].

4.1. Alternative keywords

One of the regressors in our analysis was the GI for the search term "pollution." While this keyword is very general and should include all possible searches related to environmental hazards, it may be too way general and not related to coal plants: for example, two of the top rising searches for this term in the US were "*pollution in china*" and "*china pollution*". Google Trends provides also the search trends for specific categories, which include all searches related to the chosen category according to some internal selection algorithms. The closest category related to pollution and environmental hazards is *Business and Industrial/Energy and utilities/Waste Management*. Similarly, we also downloaded the GI related to the keywords "*coal power + coal plant*" and "*coal-to-liquids + ctl coal*," but restricted to the category *Business and Industrial/Energy and utilities*. The estimated coefficients for the models including these two alternative GIs in the place of the initial ones are reported in Tables 7–8 in the Technical Appendix for coal power plants and CTL plants, respectively.

In the case of coal power plants the results do not change much in terms of signs and significance with respect to the baseline case in Table 3. The only difference is that now the *coal price* is not statistically significant. The information criteria (AIC and SIC) are higher than in the baseline case in Table 3 and the AUCs are lower. The residuals tests highlights some small misspecification in the squared residuals. In general, this robustness check confirms the results discussed in Section 3.1 but with a worse fit than the baseline case.

In the case of CTL plants, the main findings of the baseline case are also confirmed with some interesting differences. The indicator *governor* continues to be a strong significant variable (now even at the 1% level) and with the same sign as in the baseline case. Similarly, lower solar prices (with respect to wind prices) increase the probability that the project will be abandoned. The GI for "waste management" is not statistically significant as was the case for the GI for the keyword "pollution." Differently from the baseline case in Table 4, the ratio of the GI for the keywords "*coal-to-liquids + ctl coal*" and the natural gas price is now significant at the 5% level with a positive coefficient: the more people look for information about CTL plants (with respect to natural gas prices) the higher the probability the project will be abandoned. The lower the unemployment rate and the lower the number people looking for "jobs," the higher is the probability that the coal project will be abandoned/canceled. Furthermore, a higher population will decrease the odds that the coal plant will be built. The information criteria are now slightly lower than those in Table 4 for the baseline case, the AUCs are slightly higher, while the residuals tests do not highlight any particular misspecification. In general, restricting the selection criteria for Google data seems to be beneficial for the analysis of CTL plants by eliminating unrelated searches and highlighting additional significant factors beyond the political affiliation of the state governor and renewable prices, which still remain the most important factors.[12]

4.2. The recession in the years 2008–2009

The second robustness check was to evaluate the effect on our estimates of the global financial crisis in 2008 and 2009. Given the small temporal dimension of our dataset, we used a dummy variable for the years 2008 and 2009 in correspondence to the official NBER recession for the US. The estimated models including this dummy variable are reported in Tables 9–10 in the Technical Appendix for coal power plants and CTL plants[13], respectively.

In the case of coal power plants, the dummy variable is not statistically significant across all model specifications and the coefficients of all other parameters are very close to the baseline case reported in Table 3. Similar results can be observed also for CTL plants: the coefficient for the dummy variable is not significant and the results remain basically the same as those reported in Table 8 in the Technical Appendix and discussed in Section 4.1.

[12] Eliminating these variables results in steep increases of information criteria and a worse AUC, much more than the other regressors.

[13] We used restricted GIs for CTL plants, given the better results found in Section 4.1.

The global financial crisis seems not to have influenced the fate of coal projects significantly.

4.3. The DW-Nominate database

Poole and Rosenthal [18] showed that roll-call voting in both the US House and the Senate can be organized and explained by no more than two dimensions throughout the whole of American history: the first dimension is the familiar left-right (or liberal-conservative) spectrum on economic matters. The second dimension explains attitudes on cross-cutting, salient issues of the day (which include or have included slavery, bimetallism, civil rights, regional, and social/lifestyle issues), see Ref. [57] for an informal discussion. The DW-Nominate database can allow to more carefully disentangle ideology (which somehow reflects voters-at-large perception) and political affiliation of the politician in charge (possibly more affected by lobbying). We included as additional regressors the two dimensions for the US senators whose jurisdiction covered the location of the coal plants. The estimated models are reported in Tables 11—12 in the Technical Appendix.

The estimates for coal power plants show that the two dimensions are not significant while the other coefficients are very close to the baseline case in Table 3. As for CTL plants, the second dimension turns out to be significant at the 10% level with a positive coefficient: since this dimension is also interpreted as the North/South divide, this may indicate that the lobbies for coal plants are much stronger in socially conservative southern states than in northern socially liberal states. This result and the governor affiliation confirm again that CTL plants can be built only in socially conservative states with a strong lobby. Otherwise, the complex engineering and high environmental impacts make the building of these plants extremely difficult, if not impossible.

5. Conclusions

The construction of new coal plants has become an issue of great relevance in the US given the large fleet of old coal power plants that should be replaced. The analysis of the determinants that influence the success or failure of coal plants projects may be relevant for both energy policy making and project planning. We analyzed 145 coal plants and 25 coal-to-liquid plants that have been proposed in US in the period 2004—2013 and we investigated the decision to complete the plant or abandon the project using several variables and binary data models. Beside common industrial explanatory variables (size, input, output and labor costs, substitute costs, infrastructure), we also considered measures of social and environmental awareness using Google search data. After controlling for collinearity, stationarity and robustness, we performed an extensive model specification, comparison and selection.

We found that the project duration, the prices of energy substitutes for electricity generation and the awareness about the coal projects and its hazards are the main factors for coal power plants. The longer the planning period the less likely the project will be implemented: expensive legal disputes and costly project modifications to meet new requirements can make plant profitability vanish. The lower the price for natural gas and the lower the price for solar photovoltaics with respect to natural gas price the higher is the probability that the project will be abandoned. Awareness by local communities as measured by the Google search volumes about coal plants and/or coal power plants increases the probability that the project will be abandoned.

As for CTL plants, we found that the state governor's political affiliation, the ratio between solar and wind prices, the population size, the unemployment rate and the job searches as measured by Google data are the main drivers (however, the latter three are only weakly significant). CTL plants are more likely to be completed in conservative states where we presume that there is stronger political support for heavy industry projects. The lower price of solar photovoltaics with respect to wind price the higher the probability that the project will be abandoned. Larger state populations make these projects less likely, as expected, while higher unemployment rates and job searches increase the probability of successful implementation.

Appendix A. Supplementary data

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.esr.2014.11.005.

References

[1] D. Fantazzini, M. Höök, A. Angelantoni, Global oil risks in the early 21st century, Energy Policy 39 (12) (2011) 7865—7873.
[2] The National Energy Technology Laboratory Department of Energy. www.netl.-doe.gov, 2007.
[3] M. Höök, D. Fantazzini, A. Angelantoni, S. Snowden, Hydrocarbon liquefaction: viability as a peak oil mitigation strategy, Philos. Trans. R. Soc. A (2014) (forthcoming).
[4] B. Freese, S. Clemmer, C. Martinez, A. Nogee, A Risky Proposition — The Financial Hazards of New Investments in Coal Plants, Working Paper, Union Of Concerned Scientists, March 2011.
[5] L. Fleischman, R. Cleetus, S. Clemmer, J. Deyette, S. Frenkel, Ripe for retirement: an economic analysis of the U.S. coal fleet, Electr. J. 26 (10) (2013) 51—63.
[6] Sierra Club, Proposed Coal Plant Tracker, 2014. http://content.sierraclub.org/coal/environmentallaw/plant-tracker.
[7] CoalSwarm Center for Media and Democracy, Proposed Coal Plants in the United States, 2014. http://www.sourcewatch.org/index.php/Category:Proposed_coal_plants_in_the_United_States.
[8] EIA, EIA Annual Energy Outlook 2014, 2014. http://www.eia.gov/forecasts/aeo/.
[9] S. Ansolabehere, D. Konisky, Public attitudes toward construction of new power plants, Public Opin. Q. (2009). http://dx.doi.org/10.1093/poq/nfp041.
[10] P. Garrone, A. Groppi, Siting locally-unwanted facilities: what can be learnt from the location of Italian power plants, Energy Policy 45 (2012) 176—186.
[11] H. Choi, H. Varian, Predicting the Present with Google Trends, Working Paper, Google Inc., 06 2009.
[12] N. Askitas, K.F. Zimmermann, Google econometrics and unemployment forecasting, Appl. Econ. Q. 55 (2009) 107—120.
[13] T. Suhoy, Query Indices and a 2008 Downturn, Discussion Paper 6, Bank of Israel, 2009.
[14] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (2009) 1012—1014.
[15] Z. Da, J. Engelberg, G. Pengjie, In search of attention, J. Finance 5 (2011) 1461—1499.
[16] F. D'Amuri, J. Marcucci, The Predictive Power of Google Searches in Forecasting Unemployment. Temi di discussione (Economic Working Papers) 891, vol. 11, Bank of Italy, 2013.
[17] D. Fantazzini, N. Fomichev, Forecasting the real price of oil using online search data, Int. J. Comput. Econ. Econ. 4 (1—2) (2014) 4—31.
[18] K.T. Poole, H. Rosenthal, A spatial model for legislative roll call analysis, Am. J. Political Sci. 29 (2) (1985) 357—384.
[19] K.T. Poole, Spatial Models of Parliamentary Voting, Cambridge University Press, 2005.
[20] M. Höök, K. Aleklett, A review on coal to liquid fuels and its coal consumption, Int. J. Energy Res. 34 (10) (2014) 848—864.
[21] R. Coase, The problem of social cost, J. Law Econ. 3 (1960) 1—23.
[22] J.T. Hamilton, Politics and social costs: estimating the impact of collective action on hazardous waste facilities, RAND J. Econ. 24 (1) (1993) 101—125.
[23] J.T. Hamilton, Testing for environmental racism: prejudice, profits, political power? J. Policy Anal. Manag. 14 (1) (1995) 107—132.
[24] R.R. Jenkins, K.B. Maguire, C.L. Morgan, Host community compensation and municipal solid waste landfills, Land Econ. 80 (4) (2004) 513—528.
[25] D.L. Anderton, A.B. Anderson, J.M. Oakes, M.R. Fraser, Environmental equity: the demographics of dumping, Demography 31 (2) (1994) 171—192.
[26] V. Been, F. Gupta, Coming to the nuisance or going to the barrios? A longitudinal analysis of environmental justice claims, Ecol. Law Q. 24 (1) (1997) 1—56.
[27] S. Arora, T.N. Cason, Do community characteristics influence environmental outcomes?: evidence from the toxics release inventory, J. Appl. Econ. 1 (2) (1998) 413—453.
[28] A. Wolverton, Effects of socio-economic and input-related factors on polluting plants' location decisions, B.E. J. Econ. Anal. Policy 9 (1) (2009). Article 14, Available at: http://www.bepress.com/bejeap/vol9/iss1/.
[29] R. Cleetus, S. Clemmer, E. Davis, J. Deyette, J. Downing, S. Frenkel, Ripe for Retirement: the Case for Closing America's Costliest Coal Plants, Working Paper, Union of Concerned Scientists, Cambridge, MA, 2012.
[30] S. Tierney, Why Coal Plants Retire: Power Market Fundamentals as of 2012, 2012. http://www.analysisgroup.com/uploadedFiles/News_and_Events/News/2012_Tierney_WhyCoalPlantsRetire.pdf.

[31] L.F. Pratson, D. Haerer, D. Patino-Echeverri, Fuel prices, emission standards, and generation costs for coal vs natural gas power plants, Environ. Sci. Technol. 47 (9) (2013) 4926—4933.
[32] J.T. Boer, M. Pastor, J.S. Sadd, L.D. Snyder, Is there environmental racism? The demographics of hazardous waste in los angeles county, Soc. Sci. Q. 78 (4) (1997) 793—809.
[33] T. Kreutz, E. Larson, G. Liu, R. Williams, Fischer-Tropsch fuels from coal and biomass, Pittsburgh, USA, in: Proceedings of the 25th Annual International Pittsburgh Coal Conference. 29 September to 2 October, 2008. Available from: http://web.mit.edu/mitei/docs/reports/kreutz-fischer-tropsch.pdf.
[34] R. O'Brien, A caution regarding rules of thumb for variance inflation factors, Qual. Quantity 41 (5) (2007) 673—690.
[35] C. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. García Marquéz, B. Gruber, B. Lafourcade, P. Leitao, T. Münkemüller, C. McClean, P. Osborne, B. Reineking, B. Schröder, A. Skidmore, D. Zurell, S. Lautenbach, Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography 36 (2013) 27—46.
[36] A. Levin, C. Lin, C. Chu, Unit root tests in panel data: asymptotic and finite-sample properties, J. Econ. 108 (2002) 1—24.
[37] K. Im, M. Pesaran, Y. Shin, Testing for unit roots in heterogeneous panels, J. Econ. 115 (2003) 53—74.
[38] G. Maddala, S. Wu, A comparative study of unit root tests with panel data and a new simple test, Oxf. Bull. Econ. Statistics 61 (1999) 631—652.
[39] I. Choi, Unit root tests for panel data, J. Int. Money Finance 20 (2001) 249—272.
[40] K. Hadri, Testing for stationarity in heterogeneous panel data, Econ. J. 3 (2000) 148—161.
[41] P. Giudici, S. Figini, Applied Data Mining for Business and Industry, second ed., Wiley, 2009.
[42] G. Ljung, G. Box, On a measure of lack of fit in time series models, Biometrika 66 (1979) 265—270.
[43] W. Brock, D. Dechert, J. Sheinkman, B. LeBaron, A test for independence based on the correlation dimension, Econ. Rev. 175 (3) (1996) 197—235.
[44] C. Metz, H. Kronman, Statistical significance tests for binormal ROC curves, J. Math. Psychol. 22 (1980) 218—243.
[45] J.E. Goin, ROC curve estimation and hypothesis testing: applications to breast cancer detection, J. Pattern Recognit. Soc. 15 (1982) 263—269.
[46] A. Hanley, B. McNeil, The meaning and use of the area under a receiver operating characteristics (ROC) curve, Diagn. Radiol. 143 (1982) 29—36.
[47] W.J. Krzanowski, D.J. Hand, ROC Curves for Continuous Data, CRC/Chapman Hall, 2009.
[48] D. Fantazzini, M. De Giuli, S. Figini, P. Giudici, Enhanced credit default models for heterogeneous SME segments, J. Financial Transformation 25 (2009) 31—39.
[49] S. Figini, M. Maggi, Performance of Credit Risk Prediction Models via Proper Loss Functions, Working Paper 64, Department of Economics and Management, University of Pavia, 2014.
[50] P. Hansen, A. Lunde, J.M. Nason, The model confidence set, Econometrica 79 (2) (2011) 453—497.
[51] E. Rossi, D. Fantazzini, Long memory and periodicity in intraday volatility, J. Financial Econ. (2014) (forthcoming).
[52] H. Mulder, Coal-To-Liquids: Can CO_2 Emissions be Significantly Reduced?, 2009. http://www.engineering.pitt.edu/Coal_Conference/Mulder_Presentation.aspx.
[53] S. Mui, L. Tonachel, B. McEnaney, E. Shope, Ghg Emission Factors for High Carbon Intensity Crude Oils. Technical Report 2, Natural Resources Defense Council, September 2010.
[54] S. Lei, J. Zhang, An analysis of coal-to-oil industrys impact on environment, Technology 7 (2009) 84—88.
[55] F. Rong, D.G. Victor, Coal liquefaction policy in china: explaining the policy reversal since 2006, Energy Policy 39 (2011) 8175—8184. http://dx.doi.org/10.1016/j.enpol.2011.10.017.
[56] T. Young, R. Zaretski, J. Perdue, F. Guess, X. Liu, Logistic regression models of factors influencing the location of bioenergy and biofuels plants, BioResources 6 (1) (2011) 329—343.
[57] Wikipedia, Nominate (Scaling Method), 2014. http://en.wikipedia.org/wiki/NOMINATE_(scaling_methodhttp://en.wikipedia.org/wiki/NOMINATE_(scaling_method).

Appendix 12 "Forecasting German car sales using Google data and multivariate models."

Fantazzini, D., & Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics*, 170, 97-135. DOI: https://doi.org/10.1016/j.ijpe.2015.09.010

Permisson to post the Publisher's version/PDF: according to https://www.elsevier.com/about/policies/sharing "*Theses and dissertations which contain embedded Published Journal Article (PJA) as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.*"

Int. J. Production Economics 170 (2015) 97–135

Contents lists available at ScienceDirect

Int. J. Production Economics

journal homepage: www.elsevier.com/locate/ijpe

Forecasting German car sales using Google data and multivariate models

Dean Fantazzini [a,*], Zhamal Toktamysova [b]

[a] Moscow School of Economics, Moscow State University, Leninskie Gory, 1, Building 61, 119992, Moscow, Russia
[b] Faculty of Economics, Higher School of Economics, Moscow, Russia

ARTICLE INFO

Article history:
Received 15 April 2014
Accepted 4 September 2015
Available online 25 September 2015

Keywords:
Car sales
Forecasting
Google
Google trends
Global financial crisis
Great recession

ABSTRACT

Long-term forecasts are of key importance for the car industry due to the lengthy period of time required for the development and production processes. With this in mind, this paper proposes new multivariate models to forecast monthly car sales data using economic variables and Google online search data. An out-of-sample forecasting comparison with forecast horizons up to 2 years ahead was implemented using the monthly sales of ten car brands in Germany for the period from 2001M1 to 2014M6. Models including Google search data statistically outperformed the competing models for most of the car brands and forecast horizons. These results also hold after several robustness checks which consider nonlinear models, different out-of-sample forecasts, directional accuracy, the variability of Google data and additional car brands.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Long-term forecasting of car sales plays an important role in the automobile industry. Accurate predictions allow firms to improve market performance, minimize profit losses, and plan manufacturing processes and marketing policies more efficiently.

Tough competition, significant investments, and the need for quick model updates are the specifics of the automotive industry which make forecasting an element of key importance for the sales and production processes. Like other complex industries, it can be characterized by long product development cycles varying from 12 up to 60 months. An effective planning of the production therefore requires accurate long-term sales forecasts. Inaccurate forecasts may result in several negative consequences, such as overstocking or shortage of production supplies, high costs for different workforce activities, loss of reputation for the manufacturer and even bankruptcy.

There are several economic factors affecting the automobile industry, and they can be broadly divided into three groups. The first group incorporates the technological aspects of the products: quality, innovation and technology, performance and economy of the engine, functionality, safety, space management, design and aesthetics (Lin and Zhang, 2004; Sa-ngasoongsong and Bukkapatnam, 2011). The second group comprises promotion and sales factors, including wholesale and retail prices, customer service, advertising campaigns, and brand image (Landwehr et al., 2011). These factors are significant, but usually do not have a long-term effect and automobile producers in most cases can manage and control them (Dekimpe et al., 1998; Nijs et al., 2001; Pauwels et al., 2002, 2004). The third group includes various political, economic and social environmental factors which are generally beyond the control of manufacturers, such as organizational issues, political issues, global economic growth, ecological and physical forces, socio-cultural effects and consumer behavior. The use of these factors for car sales forecasting has been rather limited, see Brühl et al. (2009), Shahabuddin (2009), Wang et al. (2011) and Sa-ngasoongsong et al. (2012). Moreover, most previous studies have focused on the dynamics of car sales in the short-term, with forecast horizons usually less than 4 months, whereas car sales forecasting requires time scales with duration up to one year or more.

Following the growing number of Internet users (International Telecommunications Union, 2014) and the increasing popularity of Google as a search engine for obtaining information about cars, we propose the use of Google search data as a leading indicator for the long-term forecasting of car sales. In this regard, Google Search holds the world leadership among all search engines with a 54% market share (Net Applications, 2014). Since 2004, it has offered a tool called Google Trends, which provides information on the relative interest of users in a particular search query, at a given geographic region and at a given time (the data are available on a weekly or even daily basis). Moreover, Google Trends can attribute queries to different search categories (Autos, Computers, Finance, Health and others). In recent years, researchers worldwide have begun to use online search data to produce real-time forecasts

* Corresponding author. Tel.: +7 4955105267; fax: +7 4955105256.
E-mail address: fantazzini@mse-msu.ru (D. Fantazzini).

http://dx.doi.org/10.1016/j.ijpe.2015.09.010
0925-5273/© 2015 Elsevier B.V. All rights reserved.

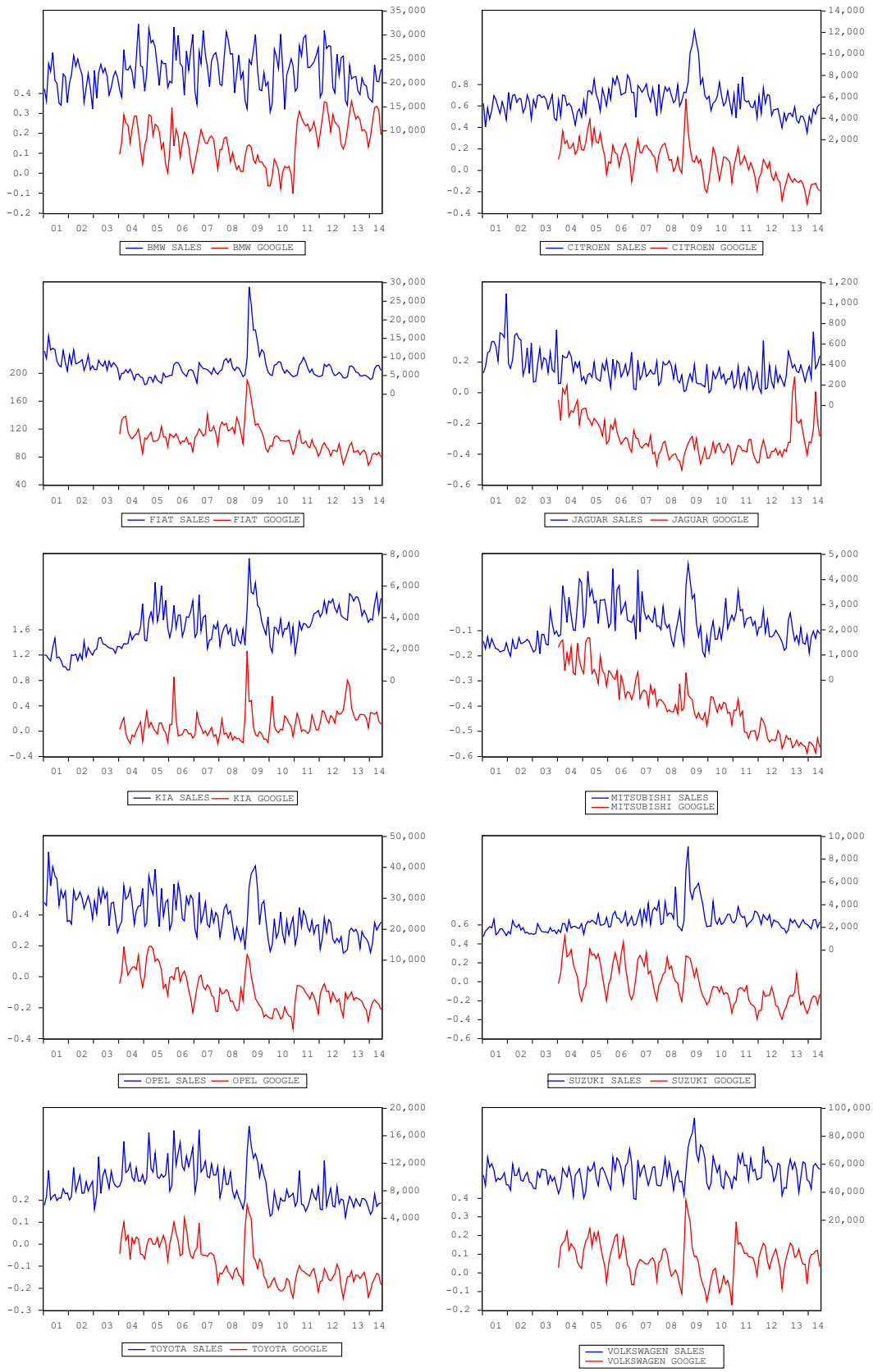


Fig. 1. Car sales (right vertical axis) and relative GIs (left vertical axis) – not seasonally adjusted. Sample: 2001M1–2014M6.

where information from official sources is released with a lag (such as 'nowcasting'), or simply as an additional variable for forecasting purposes, see Choi and Varian (2012); Askitas and Zimmermann (2006); Suhoy (2009); Ginsberg et al. (2009); Da et al. (2011); D'Amuri and Marcucci (2013); Fantazzini and Fomichev (2014) for some recent applications.

Table 1
Description of economic variables used in the analysis. The second column reports the frequency of publishing: M – monthly data, Q – quarterly data, D – daily data. The abbreviations used in the fourth column represent the data sources: GFB – German Federal Bank (Deutsche Bundesbank), DG ECFIN -Directorate General for Economic and Financial Affairs, FSO – The Federal Statistical Office (Statistisches Bundesamt), EBF – The European Banking Federation, FEA – The Federal Employment Agency (Bundesagentur für Arbeit).

Economic variable	Frequency	Seasonally adjusted	Source	Explanation
Building Construction (BC)	M	Yes	GFB	Volume index of new orders for residential buildings construction
Consumer Confidence Indicator (CCI)	M	No	DG ECFIN	Consumer survey that reflects consumer expectations
Consumer Price Index (CPI)	M	No	FSO	Measure of the ratio of a price of fixed set of consumer goods and services in current period to its price in a basic period
Euro Interbank Offered Rate (EURIBOR)	D	No	EBF	Calculated as an average rate of lending rate of the banks which participate in the survey. For the current research EURIBOR for long-term credits (1 year) is considered
Gross Domestic Product (GDP)	Q	No	FSO	Market value of all goods and services produced within a country. In the present work GDP in nominal billions Euro was taken
Production Index (PI)	M	No	FSO	Production Index for durable goods
Unemployment Rate (UR)	M	No	FEA	The registered unemployed population as a percentage of the civilian labor force
Petrol Price (PP)	M	No	FSO	Consumer price for petrol, price index

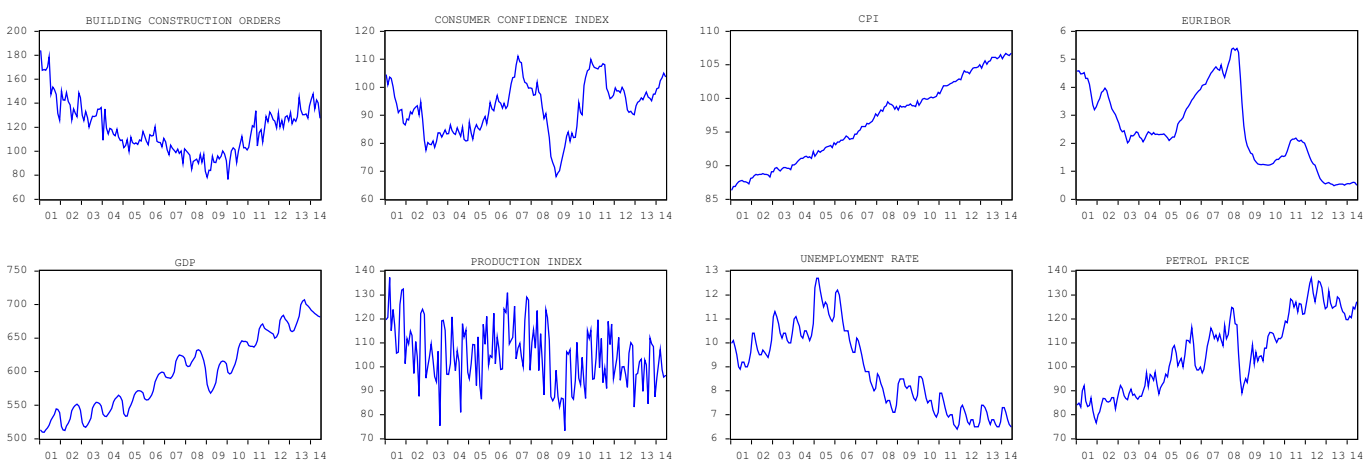


Fig. 2. Economic variables – not seasonally adjusted. Sample: 2001M1–2014M6.

With this in mind, we propose a set of models for the long-term forecasting of car sales in Germany, which consider both economic variables and online search queries. Germany is the third biggest car producer in the world (about 14 million vehicles in 2013 and 20% of the total world production) and the absolute leader in Europe (31% of the total European production), see the reports by the German Association of the Automotive Industry (GTAI, 2014) and the Germany Trade and Invest Organization (VDA, 2014) for more details. As for Internet users, Germany has the second highest number of users in Europe (12.3% of all European users) and the 7th in the world. In June 2014, more than 71 million people in Germany visited the Web at least once a month, representing 88.6% of the adult population (Internet World States, 2014).

The first contribution of this paper is a set of multivariate models which include both Google data and economic variables. So far, the vast majority of the literature has used Google data as an exogenous variable in univariate models for short-term forecasting. Given that the car industry is interested in long-term forecasting, simple univariate models are not sufficient, and multivariate models are required to produce multi-step ahead forecasts for all variables, Google data included. Moreover, we consider multivariate models for both deseasonalized data, the usual approach in the economic literature, and for data not seasonally adjusted, which is more common in practice, since planning and production departments tend to work with raw data.[2]

The second contribution of our paper is a large-scale forecasting exercise for ten car brands in Germany, where we compute out-of-sample forecasts ranging from 1 month to 24 months ahead. Our results show that models including car sales, Google data and economic variables outperform the competing models in the medium term for most of the car brands, while multivariate models including only car sales and Google data outperform the other models for long-term forecasts up to 24 steps ahead. The use of parsimonious models is crucial to obtain precise forecasts in the long run, and the use of Google search data represents a simple and powerful way to summarize the large amount of information available (see also Fantazzini and Fomichev, 2014).

The third contribution of the paper is a set of robustness checks to verify that our results also hold when considering nonlinear models, different out-of-sample forecasts, the use of directional accuracy as the main evaluation tool, Google data downloaded on different days, and additional car brands.

[2] The authors wish to thank an anonymous director of marketing and sales for pointing out this issue.

Table 2
Unit root tests: RUR = Range Unit Root test by Aparicio et al. (2006); FB=Forward–Backward RUR test by Aparicio et al. (2006); LS=Unit Root test by Lee and Strazicich (2003). Null hypothesis: the time series has a unit root.

Variable	RUR *Test statistic*	FB *Test statistic*	LS 1 break *Test statistic*	LS 2 breaks *Test statistic*	The null hypothesis is rejected by all tests?
Car sales					
BMW	0.71*	1.16	−5.08*	−11.14*	No
Citroen	1.34	1.95	−5.12*	−6.09*	No
Fiat	0.79*	1.89	−4.75*	−6.31*	No
Jaguar	0.87*	1.39	−4.47	−6.98*	No
Kia	1.42	2.01	−4.94*	−5.89*	No
Mitsubishi	0.79*	1.34	−5.05*	−5.79*	No
Opel	0.87*	1.56	−6.17*	−6.87*	No
Suzuki	1.02*	1.67	−4.91*	−6.47*	No
Toyota	1.50	1.95	−4.92*	−5.86*	No
Volkswagen	0.87*	1.73	−6.66*	−7.52*	No
Economic variables					
BUILD	1.34	2.17	−2.33	−8.68*	No
CCI	1.18	2.23	−3.60	−4.07	No
CPI	9.14*	13.15*	−3.53	−4.10	No
EURIBOR	3.07	3.73*	−3.46	−4.29	No
PP	2.68	3.96*	−3.65	−5.26	No
GDP	6.30*	8.75*	−3.67	−4.53	No
PI	1.42	1.67	−3.88	−4.80	No
UR	5.28*	7.30*	−3.42	−5.66	No
Google data					
BMW GI	1.34	1.77	−5.24*	−8.59*	No
Citroen GI	1.97	2.34	−5.98*	−6.71*	No
Fiat GI	1.43	2.34	−4.59*	−7.07*	No
Jaguar GI	1.52	1.90	−7.12*	−8.10*	No
Kia GI	0.80*	1.39	−7.45*	−8.12*	No
Mitsubishi GI	2.68	2.97	−9.26*	−9.83*	No
Opel GI	1.25	2.53	−4.51*	−5.24	No
Suzuki GI	1.88	2.09	−7.18*	−8.24*	No
Toyota GI	1.34	1.90	−4.67*	−5.17	No
Volkswagen GI	1.34	1.83	−4.96*	−5.55	No

* Significance at the 5% level.

The paper is organized as follows. Section 2 describes the data and the in-sample analysis, and the forecasting models and their out-of-sample performance are reported in Section 2.1. Robustness checks are discussed in Sections 4 and 5 briefly concludes.

2. Data and in-sample analysis

We analyze new car registrations in the Federal Republic of Germany, as provided in press releases by the Federal Motor Transport Authority (Kraftfahrt–Bundesamt). These data cover the period from January 2001 to June 2014, for a total of 162 observations. The data consist of monthly numbers of new vehicle registrations by vehicle type and new registrations of passenger cars by brand starting from 2001. For different reasons, the information for some car brands was truncated: certain brands were present only after 2001; others stopped being observed well before 2014; or the registration statistics were not published due to the small number of registrations per month. Our car brands were selected based on the availability of a long time series for new car registrations and their presence in the "`Vehicle Brands`" Google subcategory. Moreover, car brands were chosen to reflect both foreign and domestic car producers.

There were only 22 brands which had both monthly data continuously available since 2001 and were present in Google Trends. We divided these brands into clusters by taking the average sales for each brand and using the method of *k*-means with Euclidean distance. We wanted to determine large, medium and small car manufacturers, and assign all brands into three clusters. The method of *k*-means allowed us to define the number of clusters a priori and minimize the within-cluster distance while maximizing the between-cluster distance (see e.g. Hartigan, 1975). The initial *k* cluster centers are chosen to maximize the initial distance. The data are arranged to the nearest cluster center, therefore *k* clusters are formed. Next, new cluster centers are chosen as centers of mass for the clusters. After recalculation, the data are again assigned to the nearest cluster centers. The procedure ends when all centers of mass are stabilized. We found three clusters consisting of the following brands:

- *Large sellers*: Volkswagen, Opel, Ford, BMW, Audi (average monthly sales between 19523 and 53820);
- *Medium-sized sellers*: Renault, Toyota, Peugeot, Hyundai, Fiat, Mazda, Citroen, Nissan (average monthly sales between 4976 and 14074);
- *Small sellers*: Jaguar, Kia, Land Rover, Porsche, Subaru, Honda, Volvo, Mitsubishi, Suzuki (average monthly sales between 355 and 3351).

We also used the method of *k*-means with the monthly sales data from January 2001 to June 2014 and we obtained the same division into three clusters.

For the sake of space, interest and to keep the empirical analysis computationally tractable, throughout the paper we will consider three large sellers (Volkswagen, Opel, BMW), three medium-sized sellers (Toyota, Fiat, Citroen), and four small sellers (Jaguar, Kia, Mitsubishi, Suzuki). The remaining 12 brands will be examined as a robustness check in Section 4.5.

The plots of the monthly sales are reported in Fig. 1 (right vertical axis). Car sales are subject to seasonal fluctuations and all car brands tend to show several peaks during the year, with the biggest one taking place at the end of spring. In general, car sales decline during winter. The Census X-12 tests for seasonality

Table 3
Periodic Unit root tests by Boswijk and Franses (1996) and Franses and Paap (2004). NC=Not Converged. The second step is performed only if the first step numerically converged and did not reject the null hypothesis. *p*-values smaller than 0.05 are in bold.

Variable	Sample: 2001–2014		Sample: 2008–2014	
	1st step H_0: periodic unit root	**2nd step H_0: non periodic unit root**	**1st step H_0: periodic unit root**	**2nd step H_0: non periodic unit root**
Car Sales				
BMW	NC	NC	NC	NC
Citroen	18.66*	/	7.21	0.46
Fiat	16.60*	/	4.43	**0.00**
Jaguar	42.41*	/	NC	NC
Kia	10.46*	/	4.96	0.08
Mitsubishi	22.97*	/	16.96*	/
Opel	15.38*	/	10.66*	/
Suzuki	24.85*	/	15.95*	/
Toyota	10.19*	/	15.81*	/
Volkswagen	58.20*	/	NC	NC
Economic variables				
BUILD	7.99	0.09	2.32	0.11
CCI	3.23	0.06	1.02	0.14
CPI	0.13	**0.00**	0.30	0.44
EURIBOR	0.37	0.66	1.99	0.15
PP	1.97	0.88	1.36	0.10
GDP	0.01	**0.00**	0.15	**0.00**
PI	36.79*	/	22.07*	/
UR	0.52	0.56	NC	NC
Google data				
BMW GI	8.93	0.49	2.71	0.53
Citroen GI	4.90	0.47	4.46	0.13
Fiat GI	4.47	**0.04**	1.84	0.11
Jaguar GI	12.02*	/	5.17	**0.01**
Kia GI	16.82*	/	8.07	0.76
Mitsubishi GI	3.91	0.99	2.19	0.35
Opel GI	6.06	0.64	6.69	0.53
Suzuki GI	3.60	**0.02**	3.63	**0.04**
Toyota GI	5.86	0.46	5.15	**0.01**
Volkswagen GI	11.20*	/	5.38	0.39

* Significance at the 5% level.

detected that all brands exhibit stable seasonality, with no evidence of moving seasonality.

The second source of data consists of Google Trends data, which can be downloaded from `www.google.com/trends/`, using the specific "`Autos and Vehicles`" category and its "`Vehicle Brands`" subcategory. The Google Index (GI) is the ratio of the number of queries relative to a particular category (in our case the car brand), with respect to all queries in the selected region at a given point of time. The data were collected for the whole of Germany for the period January 2004–June 2014. The data have a weekly frequency and were converted to a monthly series by taking average values. While the GIs for a keyword are normalized to be bounded between 0 and 100, where 100 is the peak of the search queries, the GIs for a category are expressed in terms of percentage change from their first observation in January 2004, so that they can be both positive and negative. Their plots are reported in Fig. 1 (left vertical axis): it is interesting to note that the turning points in the GIs anticipate those in the car sales for all car brands. This initial evidence suggests that Google data may be of some help for medium- and long-term forecasting.

Additionally, we included a number of economic variables related to car sales, based on recent works by Shahabuddin (2009) and Sa-ngasoongsong et al. (2012). These variables are assumed to reflect the state of the national economy, and the factors that can influence a consumer's decision to purchase a car. The selected economic variables and their descriptions are presented in Table 1. The data were collected for the period January 2001 to June 2014. All data, with the exception of building construction orders (which were available only seasonally adjusted), show some form of seasonality, with peaks during the summer season and troughs at the end of the year. The quarterly GDP data were converted to monthly data via the quadratic match average procedure, while the daily data for Euribor rates were transformed into monthly data by taking their average. Their plots are reported in Fig. 2.

Data with seasonal behavior were seasonally adjusted with the Census X-12 adjustment program developed by US Census Bureau. However, we also considered the raw data, since they are more common in practice and of greater interest for production planners and marketing managers, who base their decisions on real data which exhibit seasonality.

All data were transformed into logarithms to reduce variability and convert nonlinear patterns to linear patterns[3] (see Sa-ngasoongsong et al., 2012). The descriptive statistics for the car registrations, the Google data and the economic variables (both seasonally adjusted and raw data) are not reported for the sake of space and are available from the authors upon request.

To select the best multivariate model for each car brand, we follow the structural relationship identification methodology discussed by Sa-ngasoongsong et al. (2012) for the case of the US car market. Briefly, the first step is to identify the order of integration

[3] The GIs were linearly re-scaled to positive numbers and then transformed into logarithms.

Table 4
Single-equation and multivariate cointegration tests with and without structural break(s) for seasonally adjusted data. The null hypothesis for all tests is the absence of cointegration. The tests considered the case of a level shift. The table cells for the Johansen tests report the number of CEs selected at the 5% level. NC=not converged.

Variable		Gregory and Hansen (1996) one (endogenous) break		Hatemi (2008) two (endogenous) breaks	
		Z–t statistic	*Break date*	*Z–t statistic*	*Break dates*
Single-equation cointegration tests					
BMW		−10.61*	2010M02	−11.14*	2006M09 2008M07
Citroen		−7.38*	2009M02	−8.35	2005M08 2007M07
Fiat		−7.54*	2006M01	−8.27	2005M11 2007M08
Jaguar		−14.54*	2012M09	−14.30*	2007M10 2011M02
Kia		−8.27*	2006M09	−8.61	2006M09 2011M01
Mitsubishi		−10.98*	2009M03	−10.79*	2008M04 2008M12
Opel		−8.72*	2009M02	−7.60	2009M09 2010M10
Suzuki		−10.85*	2009M02	−10.14	2006M09 2007M06
Toyota		−7.95*	2009M12	−8.40	2006M09 2009M07
Volkswagen		−9.96*	2009M03	−9.35	2005M08 2007M08
	Johansen (1995) **No breaks**	Johansen et al. (2000) **One(exogenous) break**		Johansen et al. (2000) **Two (exogenous) breaks**	
	N. of CEs at 5% level	*N. of CEs at 5% level*	*Break date (GH, 1996)*	*N. of CEs at 5% level*	*Break dates (H,2008)*
Multivariate cointegration tests					
BMW	5 CE	5 CE	2010M02	5 CE	2006M09 2008M07
Citroen	5 CE	4 CE	2009M02	5 CE	2005M08 2007M07
Fiat	7 CE	5 CE	2006M01	7 CE	2005M11 2007M08
Jaguar	5 CE	4 CE	2012M09	5 CE	2007M10 2011M02
Kia	5 CE	3 CE	2006M09	4 CE	2006M09 2011M01
Mitsubishi	4 CE	0 CE	2009M03	NC	2008M04 2008M12
Opel	5 CE	4 CE	2009M02	5 CE	2009M09 2010M10
Suzuki	5 CE	5 CE	2009M02	NC	2006M09 2007M06
Toyota	5 CE	5 CE	2009M12	5 CE	2006M09 2009M07
Volkswagen	5 CE	5 CE	2009M03	5 CE	2005M08 2007M08

* Significance at the 5% level.

using unit root tests; if all variables are stationary, VAR and VARX (Vector Autoregressive with exogenous variables) models are used. The second step determines the exogeneity of each variable using the sequential reduction method for weak exogeneity by Hall et al. (2002), who consider weakly exogenous each variable for which the test is not rejected and re-test the remaining variables until all weakly exogenous variables are identified. For non-stationary variables, cointegration rank tests are employed to determine the presence of a long-run relationship among the endogenous variables: if this is the case, VECM or VECMX (Vector Error Correction model with exogenous variables) models are used, otherwise VAR or VARX models in differences are applied. The last step is to compute the impulse response functions from the chosen model to trace the effect of a unit shock in one of the variables on the future values of car sales, and to compute out-of-sample forecasts (see Sa-ngasoongsong et al., 2012 for more details). Our approach differs from the one proposed by Sa-ngasoongsong et al. (2012) in two respects: first, we employ unit root tests and cointegration tests allowing for structural breaks, given the possible break in the years 2008–2009 during the global financial crisis. Second, we employ the previous identification methodology for both the seasonally adjusted data and the raw data.

2.1. Stationarity

2.1.1. Seasonally adjusted data

The stationarity of our variables is analyzed using several unit root tests allowing for potential endogenous structural break(s), both under the null of a unit root and under the alternative. We justify this choice considering the strong influence the global financial crisis in the years 2007–2009 had on the German economy, which is visible when looking at Figs. 1 and 2. As for the Google data, we remark that the statistical effects of dividing the original search data by the total number of web searches in the same week and area are unknown, so that we cannot say a priori whether they are stationary or not (see also Fantazzini and Fomichev, 2014 for a discussion on this issue). More specifically, we employed four unit root tests: the Lee and Strazicich (2003) unit root tests allowing for one and two breaks, respectively, and the Range Unit Root (RUR) and the Forward–Backward RUR tests suggested by Aparicio et al. (2006), which are non-parametric tests robust against nonlinearities, error distributions, structural breaks and outliers. A brief description of these tests is reported in the Technical Appendix A accompanying this paper and can be found on the authors' websites.

The results in Table 2 show that the majority of our time series are not stationary. However, the Lee and Strazicich (2003) tests show a stronger evidence of unit roots for economic variables, while the Aparicio et al. (2006) tests show the same for car sales and Google data. If we follow a conservative approach and analyze when all four tests reject the null hypothesis (see the last column in Table 2), then all car brands can be deemed non-stationary.

2.1.2. Raw data

To test the null hypothesis of a periodic unit root, we follow the two-step strategy suggested by Boswijk and Franses (1996) and Franses and Paap (2004). In the first step, a likelihood ratio test for testing a single unit root in a Periodic Auto-Regressive (PAR) model of order p is performed. Since there is no version of this test with endogenous breaks, we estimated it both with the full sample starting in 2001, and with a smaller sample starting in 2008. The year 2008 was chosen following the previous evidence of a

possible break in this year, which emerged with the unit root tests allowing for breaks in the case of seasonally adjusted data. If the null of a periodic unit root cannot be rejected, Boswijk and Franses (1996) and Franses and Paap (2004) suggest to test in a second step whether the process contains a non-periodic unit root equal to 1 for all seasons. A description of these tests is reported in the Technical Appendix B.

Table 5
Single-equation and multivariate cointegration tests with and without structural break(s) for raw data. The null hypothesis for all tests is the absence of cointegration. The tests considered the case of a level shift. The table cells for the Johansen tests report the number of CEs selected at the 5% level. NC=not converged.

Variable		Gregory and Hansen (1996) one (endogenous) break		Hatemi (2008) two (endogenous) breaks	
		Z-t statistic	*Break date*	*Z-t statistic*	*Break dates*
Single-equation cointegration tests					
BMW		−10.78*	2010M02	11.35*	2006M09 2008M07
Citroen		−7.70*	2009M02	8.60	2005M08 2007M07
Fiat		−7.63*	2005M10	8.64	2005M10 2007M08
Jaguar		−13.10*	2006M11	NC	NC
Kia		−8.71*	2006M09	9.25	2009M09 2011M01
Mitsubishi		−11.54*	2009M02	10.88*	2008M03 2008M12
Opel		−8.48*	2009M02	7.30	2009M09 2010M12
Suzuki		−11.00*	2009M02	9.64	2006M09 2007M07
Toyota		−7.44*	2009M12	8.03	2009M10 2010M12
Volkswagen		−10.67*	2009M02	9.63	2005M08 2007M07
	Johansen (1995) **No Breaks**	Johansen et al. (2000) **one(exogenous) break**		Johansen et al. (2000) **two(exogenous) breaks**	
	N. of CEs at 5% level	*N. of CEs at 5% level*	*Break date (GH, 1996)*	*N. of CEs at 5% level*	*Break dates (H,2008)*
Multivariate cointegration tests					
BMW	5 CE	4 CE	2010M02	5 CE	2006M09 2008M07
Citroen	5 CE	5 CE	2009M02	5 CE	2005M08 2007M07
Fiat	5 CE	6 CE	2005M10	7 CE	2005M10 2007M08
Jaguar	3 CE	0 CE	2006M11	NC	NC
Kia	5 CE	5 CE	2006M09	5 CE	2009M09 2011M01
Mitsubishi	4 CE	4 CE	2009M02	NC	NC
Opel	5 CE	4 CE	2009M02	5 CE	2009M09 2010M12
Suzuki	5 CE	6 CE	2009M02	NC	NC
Toyota	5 CE	5 CE	2009M12	5 CE	2009M10 2010M12
Volkswagen	5 CE	6 CE	2009M02	6 CE	2005M08 2007M07

* Significance at the 5% level.

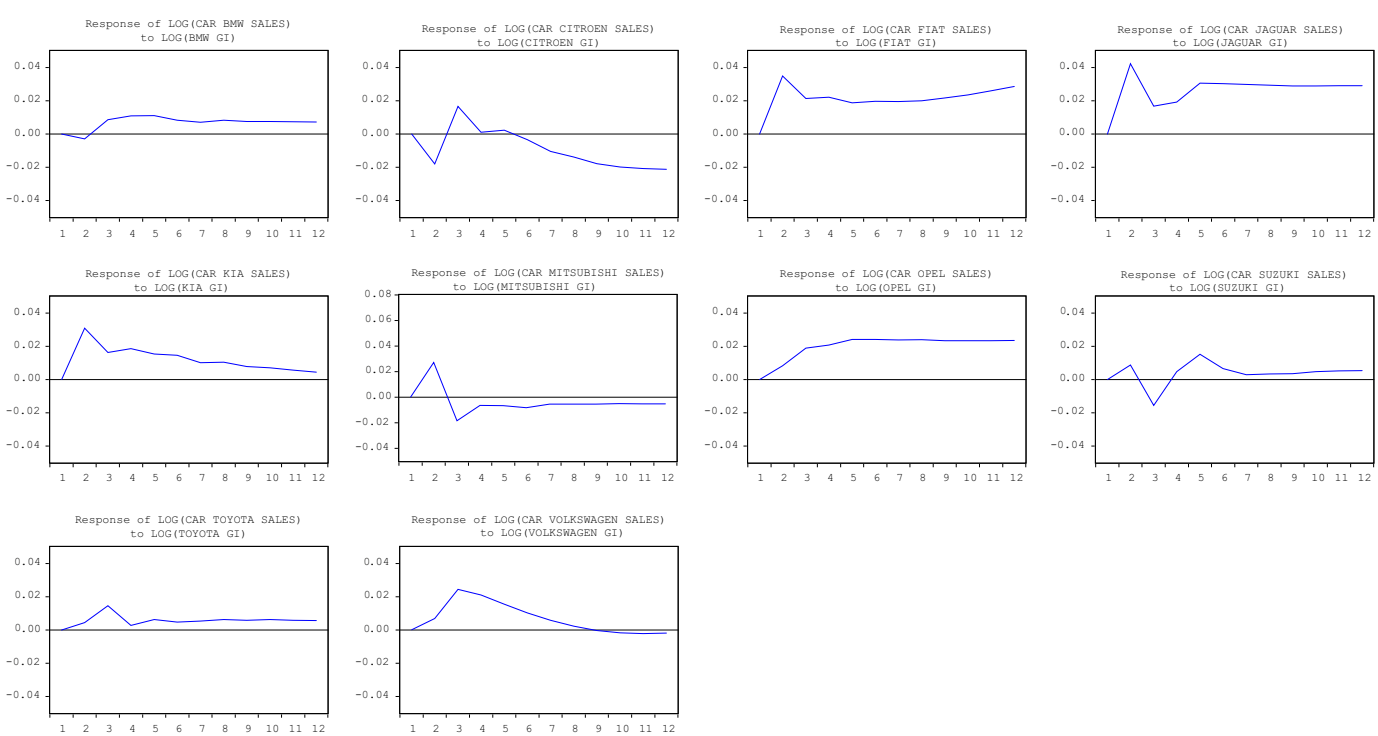


Fig. 3. Impulse response functions: response of car sales (in logs) to generalized one standard deviation innovations in the Google Indexes.

Table 6
Long-run parameters and adjustment coefficients for the Volkswagen car sales equation left table). Misspecification tests on the residuals from the Volkswagen VECMX model (right table). *t*-statistics are reported in brackets, while *p*-values smaller than 5% are reported in bold.

Long-run parameters (β)						Misspecification tests		
Variable	CE 1	CE 2	CE 3	CE 4	CE 5		*p*-values	*p*-values
Log(SALES(−1))	1	0	0	0	0	*Multi. LM(1)*	0.06	*Uni. JB test*
Log(BC(−1))	0	1	0	0	0	*Multi. LM(2)*	0.76	SALES **0.02**
Log(CCI(−1))	0	0	1	0	0	*Multi. LM(3)*	0.22	BC 0.77
Log(CPI(−1))	0	0	0	1	0	*Multi. LM(4)*	0.35	CCI 0.12
Log(EURIBOR(−1))	0	0	0	0	1	*Multi. LM(5)*	**0.02**	CPI 0.09
Log(PP(−1))	0.11	−0.71	−0.29	−0.03	0.77	*Multi. LM(6)*	0.65	EURIBOR 0.62
	[0.35]	[−2.63]	[−0.91]	[−3.34]	[1.20]	*Multi. LM(7)*	0.75	PP 0.43
Log(PI(−1))	2.03	1.90	1.57	0.14	−8.97	*Multi. LM(8)*	0.09	PI **0.03**
	[5.19]	[5.60]	[3.97]	[14.10]	[−11.10]	*Multi. LM(9)*	0.52	UR 0.54
Log(UR(−1))	0.98	−1.12	−0.27	0.05	6.53	*Multi. LM(10)*	0.41	GOOGLE **0.01**
	[3.51]	[−4.63]	[−0.95]	[6.33]	[11.31]	*Multi. LM(11)*	0.06	GDP 0.51
Log(GOOGLE(−1))	−1.95	−0.77	0.08	−0.03	−3.16	*Multi. LM(12)*	0.33	*Multi. JB test* **0.01**
	[−6.82]	[−3.08]	[0.26]	[−3.83]	[−5.34]			*Multi. White* 0.30
Log(GDP(−1))	2.16	−2.96	−0.83	−0.46	19.78			
	[2.91]	[−4.61]	[−1.11]	[−24.16]	[12.95]			
Constant	−27.59	14.91	−4.73	−2.14	−89.07			
	[−6.64]	[4.13]	[−1.13]	[−20.05]	[−10.39]			
Adjustment coefficients (α) – car sales equation								
	−0.72	−0.04	0.64	5.55	0.18			
	[−6.33]	[−0.23]	[4.60]	[1.70]	[4.25]			

Table 3 shows that car sales offer different results depending on the sample used: if the full sample is considered, non-stationarity is rejected for all car brands but BMW (for which the estimates did not reach numerical convergence); if the smaller sample starting from 2008 is used, the test failed to converge for several brands, while for two brands (Citroen and Kia) the null of a non-periodic unit root cannot be rejected. This evidence again highlights the possible presence of a structural break in 2008 during the global financial crisis. Economic variables and GIs are mostly non-stationary with a non-periodic unit root and the results do not change substantially with the sample used.

2.2. Weak exogeneity and cointegration tests

2.2.1. Seasonally adjusted data

The next step in the structural relationship identification methodology discussed by Sa-ngasoongsong et al. (2012) is to determine the exogeneity of each variable using the sequential reduction method for weak exogeneity proposed by Hall et al. (2002). This method exogenizes all weakly exogenous variables and re-tests the remaining variables until all weakly exogenous variables are identified. The variables that reject the null of weak exogeneity after re-testing are reported in Table 12 in Appendix A: the Euribor series can be considered weakly exogenous for four car brands, while almost all other variables are deemed endogenous (with some exceptions for Mitsubishi).

We then proceeded to test for cointegration using the variables which were deemed endogenous according to the previous sequential test procedure by Hall et al. (2002). We test for cointegration using a set of cointegration tests allowing for the presence of structural break(s):

- (Gregory and Hansen, 1996) single-equation cointegration test allowing for one endogenous break;
- (Hatemi, 2008) single-equation cointegration test allowing for two endogenous breaks;
- (Johansen et al., 2000) multivariate test allowing for the presence of one or two exogenous break(s), where the dates of the breaks are the ones selected by Gregory and Hansen (1996) and Hatemi (2008) tests, respectively.

A description of these cointegration tests is reported in the Technical Appendix C. For the sake of generality, we also considered the multivariate cointegration test by Johansen (1995) without breaks. The main advantage of single-equation approaches is that they allow for endogenous breaks. However, these tests are not suitable when the right-hand variables in the cointegration vector are not weakly exogenous (as in our case) and when there is more than one cointegrating vector. In this case, multivariate cointegration tests should be used. The only problem with the multivariate tests by Johansen et al. (2000) is that they allow only for exogenous breaks. Accordingly, we followed a 2-step strategy: we first estimated the single-equation tests to obtain an indication of the structural break dates. We then used these dates to compute the tests by Johansen et al. (2000). Finally, we remark that the number of lags for the Johansen tests were chosen to minimize the Schwartz criterion and to make the residuals approximately white noise.

Table 4 shows that there is strong evidence for cointegration for all considered car brands. However, structural breaks seem to have a non-negligible effect, particularly when considering Johansen multivariate tests. Moreover, the effects of breaks appear to be much stronger for foreign brands than for domestic brands (BMW, Volkswagen and, to a lesser extent, Opel), for which the cointegration tests do not change substantially when breaks are taken into account.

Table 7
Models used for forecasting (baseline case).

Type	Log-levels /log-returns	Exogenous variables	Google data	Notes	Short cut notation (seas. adj. data)	Short cut notation (raw data)
VEC MODELS						
VECM	Log-lev/log-ret	No	Yes		VECM	VECMP
VECMX	Log-lev/log-ret	Yes	Yes		VECMX	VECMXP
VECM	Log-lev/log-ret	No	No		VECMNOGO	VECMPNOGO
VECM	Log-lev/log-ret	No	No	Sample starts in 2004	VECMNOGO4	VECMPNOGO4
VECMX	Log-lev/log-ret	Yes	No		VECMXNOGO	VECMXPNOGO
VECMX	Log-lev/log-ret	Yes	No	Sample starts in 2004	VECMXNOGO4	VECMXPNOGO4
VECM	Log-lev/log-ret	No	Yes	Only sales and GI. Lags: 1,12	VECongo112	VEPongo112
VECM	Log-lev/log-ret	No	Yes	Only sales and GI. Lags: 1-12	VECongo12	VECPongo12
VAR MODELS						
VAR	Log-levels	No	Yes		VAR	VARP
VAR	Log-returns	No	Yes		VARD	VARPD
VAR	Log-levels	Yes	Yes		VARX	VARXP
VAR	Log-returns	Yes	Yes		VARXD	VARXPD
VAR	Log-levels	No	No		VARNOGO	VARPNOGO
VAR	Log-levels	No	No	Sample starts in 2004	VARNOGO4	VARPNOGO4
VAR	Log-returns	No	No		VARDNOGO	VARPDNOGO
VAR	Log-returns	No	No	Sample starts in 2004	VARDNOGO4	VARPDNOGO4
VAR	Log-levels	Yes	No		VARXNOGO	VARXPNOGO
VAR	Log-levels	Yes	No	Sample starts in 2004	VARXNOGO4	VARXPNOGO4
VAR	Log-returns	Yes	No		VARXDNOGO	VARXPDNOGO
VAR	Log-returns	Yes	No	Sample starts in 2004	VARXDNOGO4	VARXPDNOGO4
VAR	Log-levels	No	Yes	Only sales and GI. Lags: 1,12	VARongo112	VARongo112
VAR	Log-levels	No	Yes	Only sales and GI. Lags: 1-12	VARongo12	VARongo12
VAR	Log-returns	No	Yes	Only sales and GI. Lags: 1,12	VADongo112	VADongo112
VAR	Log-returns	No	Yes	Only sales and GI. Lags: 1-12	VADongo12	VADongo12
BVAR MODELS						
BVAR	Log-levels	Yes	Yes		BVAR	BVARP
BVAR	Log-returns	Yes	Yes		BVARD	BVARPD
BVAR	Log-levels	Yes	No		BVARNOGO	BVARPNOGO
BVAR	Log-levels	Yes	No	Sample starts in 2004	BVARNOGO4	BVARPNOGO4
BVAR	Log-returns	Yes	No		BVARDNOGO	BVARPDNOGO
BVAR	Log-returns	Yes	No	Sample starts in 2004	BVARDNOGO4	BVARPDNOGO4
UNIVARIATE TIME SERIES MODELS						
AR(12)	Log-returns	No	No		AR12	AR12
AR(12)	Log-returns	No	No	Sample starts in 2004	AR124	AR124
R. w.	Log-returns	No	No		RW	RW
R. w.	Log-returns	No	No	Sample starts in 2004	RW4	RW4
PERIODIC ERROR CORRECTION MODELS						
Periodic ECM	Log-lev/log-ret	No	Yes	Only sales and GI. Lags: 1-12	/	PECM

2.2.2. *Raw data*

To determine the exogeneity of variables with potential seasonal behavior, we extend the previous sequential reduction method for weak exogeneity by including centered seasonal dummies: they sum to zero over time and therefore do not affect the asymptotic distributions of the tests (see Johansen, 1995, 2006). The variables that reject the null of weak exogeneity after re-testing are reported in Table 13 in Appendix A: the results for raw data are not too dissimilar to the seasonally adjusted data, even though there are less variables which are weakly exogenous. We then tested for cointegration using the variables which were found to be endogenous, and the previous cointegration tests augmented with centered seasonal dummies, see Table 5.

In the case of raw data, the evidence for cointegration appears to be quite similar to that of seasonally adjusted data, particularly when considering the Johansen test without breaks and with one break. Moreover, the fact that the Johansen test with two breaks failed to converge for some car brands indicates that our sample is too small for two breaks and that only tests with one break should be considered.

Periodic cointegration tests using all variables could not be implemented due to the high number of parameters being estimated (the so-called "curse of dimensionality"). However, we wanted to consider a restricted bivariate periodic error correction model including only car sales and Google data. Even though such a specification is definitely biased – missing several important economic variables – this parsimonious model can nevertheless be of interest for forecasting purposes. Moreover, the capacity of Google data to summarize a wealth of information should not be underestimated. In this regard, we implemented the single-equation periodic cointegration test discussed in Franses and Paap (2004), which is an extension of the Boswijk (1994) cointegration test. The null hypothesis is the absence of cointegration against the alternative of periodic cointegration and the right-

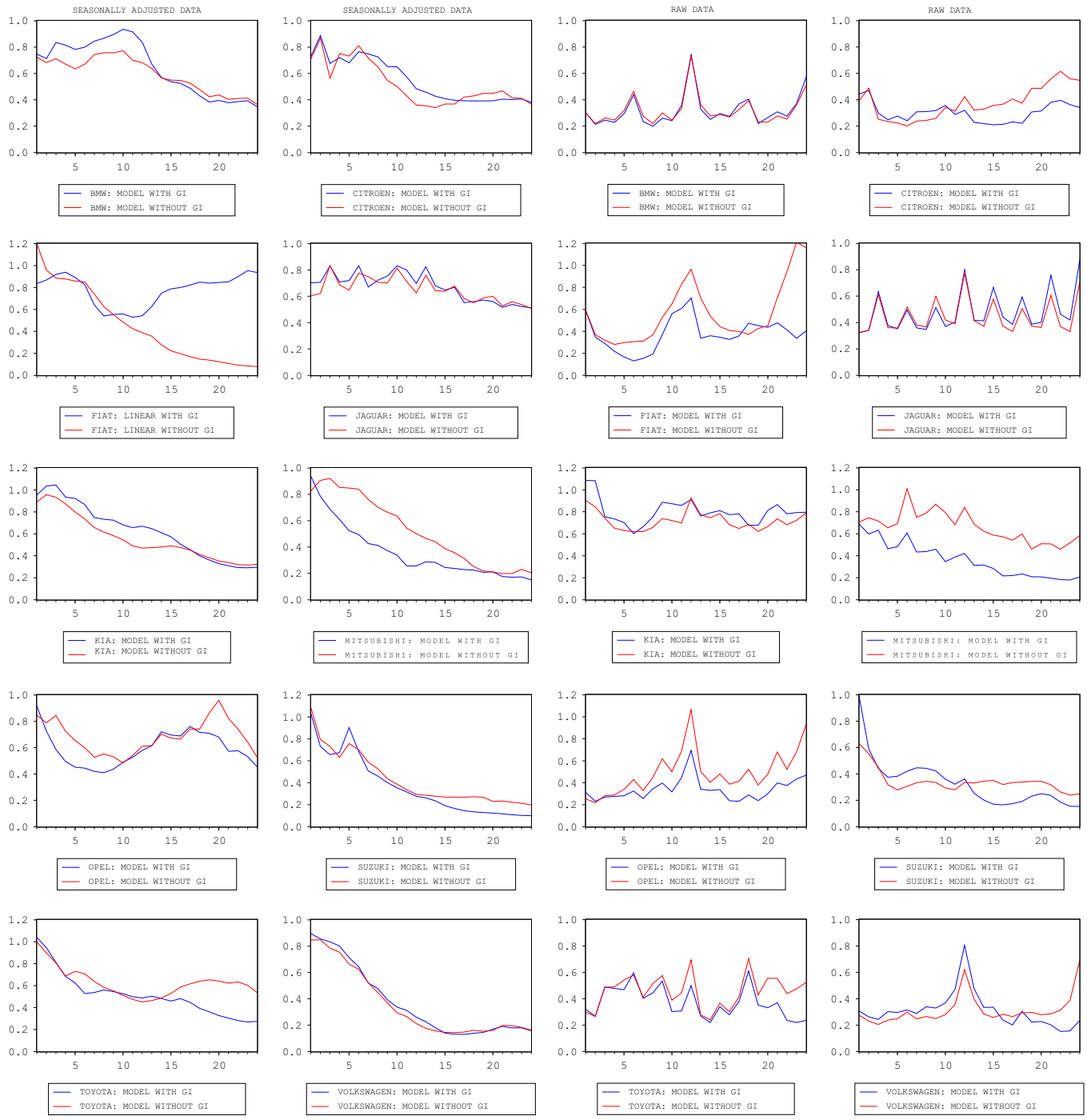


Fig. 4. Ratios of the MSPEs of the best models with and without Google data and the Random Walk model across all forecasting horizons. The first two columns show results for seasonally adjusted data, and the last two for raw data.

hand variables should be weakly exogenous. A description of this test as well as the test for weak exogeneity in the case of periodic variables by Boswijk (1994) is reported in the Technical Appendix D. Since we are not aware of any extension of this test allowing for structural breaks, we estimated it using both the full sample and a reduced sample starting in 2008 to take any potential break into account and the results are reported in Table 14 in Appendix A: the evidence in favor of periodic cointegration is fairly strong, but the results of the Boskwijk test statistics change partially when the smaller sample starting in 2008 is considered. Caution should therefore be exercised when dealing with this restricted model. Interestingly, the GIs are weakly exogenous with respect to car sales for almost all brands at the 5% level and this outcome does not change substantially with the sample used.

2.3. Impulse response functions

After the VECM (or VECMX) models were selected for each car brand, we proceeded to compute the impulse response functions (IRFs) in order to trace the effects of a one-time shock in one of the variables on current and future values of car sales. More specifically, we computed the generalized impulse response functions by Pesaran and Shin (1998), which do not depend on the ordering of the variables. For the sake of interest and space, we report here

Table 8
List of the best models for each model class, the corresponding MSPE and ranking. Forecast horizon: 6 steps ahead.

Model class	*Model*	*MSE*	*Ranking*	*MCS*	*Model*	*MSE*	*Ranking*	*MCS*	*Model*	*MSE*	*Ranking*	*MCS*
Seasonally adjusted data												
	Bmw				**Citroen**				**Fiat**			
Linear w. GI	BVAR	7,255,188	36	Yes	VARD	1,937,867	4	Yes	VARongo112	11,439,630	10	Yes
Linear w/o GI	AR12	608,0991	4	Yes	VARDNOGO	2,054,348	6	No	BVARNOGO	11,763,929	11	Yes
Nonlinear	AAR(7)dlog	5,698,093	1	Yes	LSTAR(3)log	1,688,104	1	Yes	SETAR(2)log	9,458,281	1	Yes
RW		9,073,312	54	Yes		2,534,232	36	No		13,839,395	26	Yes
	Mitsubishi				**Opel**				**Suzuki**			
Linear w. GI	VECongo112	285,910	1	Yes	VECongo112	16,265,040	1	Yes	VARongo112	1,206,941	9	Yes
Linear w/o GI	VECMNOGO4	484,285	32	Yes	VECMXNOGO	21,934,786	6	No	VECMXNOGO	1,233,267	10	Yes
Nonlinear	SETAR(11)log	347,695	3	Yes	AAR(2)log	24,373,967	9	No	SETAR(3)log	812,440	1	Yes
RW		577,041	62	Yes		36,590,894	40	No		1,760,865	27	No
Raw data												
	Bmw				**Citroen**				**Fiat**			
Linear w. GI	VADongo112	9,546,849	1	Yes	VARongo112	913,022	3	Yes	VARongo112	1,225,557	1	Yes
Linear w/o GI	BVARPDNOGO	10,099,266	2	Yes	BVARPNOGO	763,696	1	Yes	BVARPNOGO	2,874,716	13	Yes
Nonlinear	AAR(8)log	17,435,149	24	No	AAR(3)log	1,121,818	6	No	LSTAR(2)log	2,615,865	2	Yes
RW		21,780,883	62	No		3,800,014	70	No		9,377,139	62	No
	Mitsubishi				**Opel**				**Suzuki**			
Linear w. GI	VEPongo112	193,905	1	Yes	VEPongo112	11,342,673	1	Yes	BVARP	523,703	12	Yes
Linear w/o GI	BVARPDNOGO	322,456	26	Yes	BVARPNOGO	15,033,303	5	No	BVARPNOGO4	380,224	10	Yes
Nonlinear	LSTAR(3)log	234,530	2	Yes	SETAR(9)dlog	31,266,013	29	No	SETAR(2)log	269,229	1	Yes
RW		319,260	24	Yes		34,959,425	38	No		1,244,803	34	No

	Model	*MSE*	*Ranking*	*MCS*	*Model*	*MSE*	*Ranking*	*MCS*
Seasonally adjusted data								
	Jaguar				**Kia**			
Linear w. GI	VADongo112	7409	22	Yes	VARongo112	1,220,217	18	Yes
Linear w/o GI	AR12	6913	11	Yes	AR12	1,036,132	12	Yes
Nonlinear	SETAR(5)log	6478	1	Yes	AAR(4)log	909,397	1	Yes
		8907	61	Yes		1,406,331	28	Yes
	Toyota				**Volkswagen**			
Linear w. GI	VECongo112	4,040,621	1	Yes	VARongo112	60,790,600	2	Yes
Linear w/o GI	VECMXNOGO	5,429,724	3	Yes	AR124	58,970,200	1	Yes
Nonlinear	AAR(1)log	5,653,546	6	Yes	AAR(1)log	71,730,620	9	No
		7,663,277	29	No		94,350,150	36	No
Raw data								
	Jaguar				**Kia**			
Linear w. GI	BVARPD	7852	1	Yes	BVARPD	731,755	1	Yes
Linear w/o GI	BVARPDNOGO4	8209	2	Yes	BVARPDNOGO	753,717	3	Yes
Nonlinear	AAR(3)log	12,342	30	No	AAR(9)log	828,305	6	Yes
		15,886	90	No		1,218,423	33	No
	Toyota				**Volkswagen**			
Linear w. GI	VEPongo112	3,433,042	2	Yes	VEPongo112	48,248,128	2	Yes
Linear w/o GI	VECMXPNOGO4	3,333,000	1	Yes	BVARPNOGO	45,734,023	1	Yes
Nonlinear	SETAR(1)dlog	4,165,033	6	Yes	SETAR(4)log	72,437,319	12	No
		5,758,188	30	No		153,257,703	63	No

only the IRFs for the seasonally adjusted sales data (Fig. 3) with respect to a generalized one standard deviation innovation in the Google Indexes. Moreover, we report in Table 6 the estimated long-run parameters in the cointegration equations and their adjustment coefficients for the Volkswagen car sales equation, noting that Volkswagen is the biggest car maker and seller in Germany. A battery of misspecification tests computed on the VECMX model residuals is reported in the same table as well: we computed multivariate LM test statistics for residual serial correlation up to a specified order, univariate and multivariate Jarque-Bera residual normality tests, and the multivariate White heteroskedasticity test (see Johansen, 1995 and Lutkepohl, 2005 for more details about these tests). The full results are available from the authors upon request.

As expected, a unit shock in the Google Index has a rather long and positive effect for almost all car brands. Similarly, the model estimates in Table 6 show that the Google Index enters almost all cointegration equations with significant positive coefficients,[4] while the residual tests do not signal any serious misspecification.

3. Out-of-sample forecasting analysis

The last step in the structural relationship identification methodology discussed by Sa-ngasoongsong et al. (2012) is to compare the forecasting performances of the selected VECM (or VECMX) models with a set of competitors.

[4] The signs of the long-run parameters in Table 6 are switched due to the error correction representation.

Table 9
List of the best models for each model class, the corresponding MSPE and ranking. Forecast horizon: 12 steps ahead.

Model class	*Model*	*MSE*	*Ranking*	*MCS*	*Model*	*MSE*	*Ranking*	*MCS*	*Model*	*MSE*	*Ranking*	*MCS*
Seasonally adjusted data												
	Bmw				**Citroen**				**Fiat**			
Linear w. GI	VARongo112	11,630,497	20	Yes	VECongo112	1,746,615	9	Yes	VARongo112	8,070,599	15	No
Linear w/o GI	AR12	9,491,265	5	Yes	BVARNOGO	1,302,593	6	Yes	BVARNOGO	5,773,194	10	No
Nonlinear	SETAR(1)log	8,506,115	1	Yes	SETAR(4)log	1,170,021	1	Yes	AAR(1)log	2,142,114	1	Yes
RW		13,909,104	43	Yes		3,606,090	43	No		14,889,930	30	No
	Mitsubishi				**Opel**				**Suzuki**			
Linear w. GI	VECongo12	227,696	1	Yes	VECongo112	21470315	3	Yes	BVAR	921,548	12	No
Linear w/o GI	AR12	446,751	5	Yes	BVARNOGO	22,609,378	5	Yes	BVARNOGO4	979,593	14	No
Nonlinear	SETAR(11)log	393,336	12	Yes	SETAR(7)dlog	20,607,165	1	Yes	AAR(3)log	222,809	1	Yes
RW		888,768	70	Yes		37,088,135	43	No		3,319,004	44	No
Raw data												
	Bmw				**Citroen**				**Fiat**			
Linear w. GI	VADongo112	11,298,025	3	Yes	VARongo112	872,250	1	Yes	PECM	2,688,297	15	Yes
Linear w/o GI	AR124	11,133,748	1	Yes	AAR(3)log	1,109,012	3	Yes	BVARPNOGO4	3,686,731	23	No
Nonlinear	SETAR(10)log	13,316,723	9	No	BVARPNOGO	1,156,313	4	Yes	SETAR(1)log	1,732,960	1	Yes
RW		15,260,733	16	No		2,734,442	47	No		3,822,802	24	Yes
	Mitsubishi				**Opel**				**Suzuki**			
Linear w. GI	VEPongo112	207,765	1	Yes	VEPongo112	12,940,989	1	Yes	BVARP	585,175	13	No
Linear w/o GI	AR12	414,701	26	Yes	BVARPNOGO	19,896,671	4	No	BVARPNOGO4	541,203	12	Yes
Nonlinear	AAR(1)log	273,074	2	Yes	AAR(1)dlog	24,701,172	13	No	AAR(4)log	240,509	1	Yes
RW		493,513	41	Yes		18,710,646	3	Yes		1,613,001	38	No

	Model	*MSE*	*Ranking*	*MCS*	*Model*	*MSE*	*Ranking*	*MCS*
Seasonally adjusted data								
	Jaguar				**Kia**			
Linear w. GI	VECongo112	9089	15	Yes	VARongo112	1,226,001	24	
Linear w/o GI	AR12	8158	3	Yes	AR12	859,701	14	
Nonlinear	SETAR(3)log	7758	1	Yes	SETAR(3)log	709,664	1	
		13,024	63	Yes		1,830,155	29	
	Toyota				**Volkswagen**			
Linear w. GI	BVAR	5,198,460	6	Yes	VARongo112	34,715,966	2	Yes
Linear w/o GI	BVARNOGO4	4,832,954	1	Yes	AR124	28,352,989	1	Yes
Nonlinear	SETAR(7)dlog	4,977,505	2	Yes	SETAR(7)log	37,548,975	4	Yes
		10,675,830	36	No		133,013,321	47	No
Raw data								
	Jaguar				**Kia**			
Linear w. GI	BVARPD	9021	2	Yes	VADongo112	859,092	3	Yes
Linear w/o GI	BVARPDNOGO4	8749	1	Yes	BVARPDNOGO	871,338	5	Yes
Nonlinear	SETAR(7)dlog	11,973	13	Yes	AAR(9)log	765,639	1	Yes
		11,294	9	Yes		942,287	10	Yes
	Toyota				**Volkswagen**			
Linear w. GI	VEPongo112	2,398,770	1	Yes	BVARP	78,930,479	7	Yes
Linear w/o GI	BVARPDNOGO4	3,334,235	6	Yes	BVARPNOGO	60,857,150	1	Yes
Nonlinear	SETAR(6)log	3,292,246	5	Yes	SETAR(6)log	72,922,923	2	Yes
		4,803,964	26	No		98,132,979	22	Yes

3.1. Seasonally adjusted data

We compared a set of 34 models, which allow for different degrees of model flexibility, parsimonious specifications and numerical tractability. More specifically, three types of multivariate models were employed:

- *Vector Error Correction*(*VEC*) *models*: We considered both VECM and VECMX models, as well as models with and without Google data, to better examine their effects on forecasting performance. The number of lags was selected to minimize the Schwartz criteria and to make the residuals approximately white noise. We also considered a set of parsimonious bivariate specifications including only car sales and Google data, which may be of interest for long-term forecasting.
- *Vector Auto-Regressive* (*VAR*) *models*: We considered VAR models with variables in log-levels and in log-differences, to consider both cases of stationarity and non-stationarity. Moreover, models with and without exogenous variables and with and without Google data were also considered. Finally, a set of parsimonious bivariate VAR models including only car sales and Google data was included.
- *Bayesian Vector Auto-Regressive* (*BVAR*) *models*: When there are a lot of variables and a high number of lags, estimating the parameters of a VAR model can be very difficult, if not impossible. One way to solve this issue is to shrink the parameters using Bayesian methods. Bayesian VAR models have recently enjoyed a lot of success in macroeconomic forecasting (see Koop and Korobilis, 2010 for a recent review and Fantazzini and Fomichev, 2014 for a recent application with Google data). In this regard, we used the so-called Litterman/Minnesota prior, which was developed by researchers at the University of Minnesota and at the Federal Reserve Bank of Minneapolis, and which is a common choice in empirical applications due to its computational speed and forecasting success (see Doan et al., 1984; Litterman, 1986; Koop and Korobilis, 2010). A brief

Table 10
List of the best models for each model class, the corresponding MSPE and ranking. Forecast horizon: 18 steps ahead.

Model class	*Model*	*MSE*	*Ranking*	*MCS*	*Model*	*MSE*	*Ranking*	*MCS*	*Model*	*MSE*	*Ranking*	*MCS*
Seasonally adjusted data												
	Bmw				**Citroen**				**Fiat**			
Linear w. GI	VARongo112	8,464,643	2	Yes	VECongo112	1,920,350	7	Yes	VECongo112	13,037,790	21	No
Linear w/o GI	AR12	9,384,237	3	Yes	BVARNOGO	2,109,563	9	Yes	AR124	2,271,788	4	Yes
Nonlinear	AAR(1)log	8,018,508	1	Yes	AAR(1)log	1,266,893	1	Yes	AAR(1)log	783,045	1	Yes
RW		19,689,847	42	No		4,907,445	38	No		15,368,830	27	No
	Mitsubishi				**Opel**				**Suzuki**			
Linear w. GI	VECongo112	174,546	1	Yes	VECongo112	18,394,010	11	Yes	VARongo112	634,907	9	No
Linear w/o GI	VECMNOGO4	195,035	4	Yes	AR12	19,024,430	12	Yes	VECMXNOGO	1,279,649	15	No
Nonlinear	SETAR(10)log	288,480	5	No	SETAR(7)dlog	11,826,920	1	Yes	SETAR(2)log	191,776	1	Yes
RW		777,574	66	No		25,704,740	29	Yes		4,680,547	37	No
Raw data												
	Bmw				**Citroen**				**Fiat**			
Linear w. GI	VARPD	15,598,460	11	Yes	VARongo112	1,402,351	2	Yes	VEPongo112	5,156,512	20	Yes
Linear w/o GI	AR124	15,171,520	8	Yes	BVARPNOGO	2,374,012	15	Yes	AR12	4,056,693	18	Yes
Nonlinear	LSTAR(9)log	12,761,820	1	Yes	AAR(3)log	1,151,259	1	Yes	LSTAR(2)log	2,217,764	1	Yes
RW		38,548,350	62	No		6,342,691	66	No		10,898,232	44	No
	Mitsubishi				**Opel**				**Suzuki**			
Linear w. GI	VEPongo112	159,969	1	Yes	VARongo112	10,159,473	1	Yes	VARongo112	463,961	13	Yes
Linear w/o GI	BVARPNOGO4	395,301	24	Yes	BVARPDNOGO	18,345,564	3	Yes	BVARPNOGO4	816,324	17	No
Nonlinear	AAR(1)log	257,037	4	Yes	SETAR(10)dlog	29,080,284	11	No	AAR(2)log	237,811	1	Yes
RW		685,032	60	No		35,161,323	16	No		2,418,065	36	No

	Model	*MSE*	*Ranking*	*MCS*	*Model*	*MSE*	*Ranking*	*MCS*
Seasonally adjusted data								
	Jaguar				**Kia**			
Linear w. GI	VECongo112	9462	2	Yes	VARongo112	940,676	19	Yes
Linear w/o GI	AR12	9313	1	Yes	AR124	970,566	21	Yes
Nonlinear	LSTAR(5)log	10,319	6	Yes	SETAR(8)log	780,965	1	Yes
		16,920	50	No		2,358,597	29	No
	Toyota				**Volkswagen**			
Linear w. GI	VECongo112	3,866,818	1	Yes	VARongo112	14,865,460	1	Yes
Linear w/o GI	BVARNOGO4	6,302,790	19	Yes	AR124	17,283,870	2	Yes
Nonlinear	SETAR(3)dlog	4,001,607	2	Yes	AAR(1)log	26,144,200	4	Yes
		9,851,512	36	No		107,189,500	43	No
Raw data								
	Jaguar				**Kia**			
Linear w. GI	VEPongo112	13,929	2	Yes	VEPongo112	1,249,080	5	Yes
Linear w/o GI	VARPNOGO	11,879	1	Yes	AR124	1,262,323	6	Yes
Nonlinear	AAR(3)log	14,774	4	Yes	AAR(9)log	967,518	1	Yes
		23,496	67	Yes		1,841,665	29	No
	Toyota				**Volkswagen**			
Linear w. GI	VEPongo112	3,849,544	1	Yes	VECPongo12	58,221,750	2	Yes
Linear w/o GI	BVARPDNOGO	4,438,429	6	Yes	BVARPNOGO	55,863,450	1	Yes
Nonlinear	SETAR(8)log	4,084,996	3	Yes	SETAR(2)log	83,379,030	10	Yes
		6,310,415	29	No		190,773,000	38	No

description of BVAR models can be found in the Technical Appendix E. Similarly to the VAR and VECM models, we considered models with and without exogenous variables, with and without Google data and with variables both in log-levels and in log-differences.

Besides these models, we also considered a set of standard univariate time series models:

- The Random Walk with drift.
- An AR(12) model for the log-returns of car sales.

Moreover, all models without Google data were estimated using both a long sample starting in 2001 and a short one starting in 2004, in the hope that this will show more clearly the advantages of Google data. The full details of all 34 multivariate models are reported in Table 7. For ease of reference, we also report in the sixth column a short-cut notation for identifying each model in the tables reporting the models forecasting performances.

We used the data between 2001M1 and 2008M9 as the first initialization sample for the models without Google data, and data from 2004M1 till 2008M9 for the models with Google data and those without Google data but estimated on a shorter sample. The evaluation period ranged from 2008M10 till 2014M6 and was used to compare forecasts from 1 step ahead up to 24 steps ahead. The top three models in terms of the Mean Squared Prediction Error (MSPE) for each forecasting horizon and each car brand are reported in Table 15, while the full results are available from the authors upon request.

Table 15 shows that there is no single model which outperforms all competitors for all horizons and all car brands. However, some general indications can be retrieved:

- The MSPEs of the competing models with forecasting horizons up to 8–10 steps ahead are relatively close (results not reported)

Table 11
List of the best models for each model class, the corresponding MSPE and ranking. Forecast horizon: 24 steps ahead.

Model class	*Model*	*MSE*	*Ranking*	*MCS*	*Model*	*MSE*	*Ranking*	*MCS*	*Model*	*MSE*	*Ranking*	*MCS*
Seasonally adjusted data												
	Bmw				**Citroen**				**Fiat**			
Linear w. GI	VARongo112	8,099,346	2	Yes	VARongo112	2,770,683	9	No	VECongo112	16,937,190	22	No
Linear w/o GI	AR124	8,529,251	3	Yes	AR124	2,670,345	8	No	AR124	1,425,390	2	Yes
Nonlinear	AAR(1)log	7,144,053	1	Yes	AAR(1)log	1,476,196	1	Yes	AAR(1)log	906,669	1	Yes
RW		23,423,336	41	No		7,290,366	31	No		18,093,150	24	No
	Mitsubishi				**Opel**				**Suzuki**			
Linear w. GI	VECongo112	145,162	1	Yes	VECongo112	13,706,410	2	Yes	VARongo112	711,434	9	No
Linear w/o GI	VECMNOGO4	198,324	3	Yes	VECMXNOGO	15,895,950	6	Yes	VECMXNOGO	1,414,340	13	No
Nonlinear	SETAR(10)log	175,800	2	Yes	SETAR(8)dlog	11,794,440	1	Yes	SETAR(2)log	208,338	1	Yes
RW		972,338	62	No		30,265,060	32	No		7,152,681	34	No
Raw data												
	Bmw				**Citroen**				**Fiat**			
Linear w. GI	VADongo112	19,015,100	29	Yes	VARongo112	2,029,587	10	Yes	VARongo112	1,822,430	5	Yes
Linear w/o GI	AR12	16,978,200	10	Yes	BVARPNOGO	3,248,731	28	Yes	VARPDNOGO	5,203,134	24	No
Nonlinear	LSTAR(8)log	12,820,190	1	Yes	AAR(3)log	987,284	1	Yes	LSTAR(5)log	1,600,311	1	Yes
RW		33,076,160	51	Yes		5,935,189	55	No		4,488,673	21	No
	Mitsubishi				**Opel**				**Suzuki**			
Linear w. GI	VARongo112	154,734	1	Yes	VARongo112	9,654,207	1	Yes	VARongo112	640328	14	No
Linear w/o GI	AR12	436,153	42	No	VECMXPNOGO4	19,153,863	3	Yes	BVARPNOGO4	1,044,408	15	No
Nonlinear	AAR(1)log	216,117	3	Yes	LSTAR(10)dlog	23,307,805	10	No	AAR(2)log	211,698	1	Yes
RW		748,481	65	No		20,556,112	7	Yes		4,180,500	47	No

	Model	*MSE*	*Ranking*	*MCS*	*Model*	*MSE*	*Ranking*	*MCS*
Seasonally adjusted data								
	Jaguar				**Kia**			
Linear w. GI	VECongo112	10,310	2	Yes	VARongo112	965,237	7	No
Linear w/o GI	AR12	10,164	1	Yes	AR124	1,056,006	13	No
Nonlinear	AAR(2)log	11325	5	Yes	SETAR(6)log	816,700	1	Yes
		20,103	41	No		3,274,180	30	No
	Toyota				**Volkswagen**			
Linear w. GI	VECongo112	4,005,582	1	Yes	VARongo112	17,513,320	1	Yes
Linear w/o GI	VECMXNOGO	7,803,276	18	No	AR124	18,042,460	2	Yes
Nonlinear	SETAR(6)dlog	5,568,402	2	Yes	AAR(1)log	27,725,190	4	Yes
		14,568,460	42	No		111,008,400	40	No
Raw data								
	Jaguar				**Kia**			
Linear w. GI	VEPongo112	15,424	3	Yes	VARongo12	1,256,239	5	No
Linear w/o GI	VARPNOGO	12,523	1	Yes	AR124	1,244,649	4	No
Nonlinear	LSTAR(10)log	16,592	11	Yes	AAR(9)log	1,000,641	1	Yes
		17,635	17	Yes		1,577,128	16	No
	Toyota				**Volkswagen**			
Linear w. GI	VARongo112	2,192,391	1	Yes	VECPongo12	23,921,080	1	Yes
Linear w/o GI	VECMPNOGO	4,851,103	7	Yes	BVARPNOGO	70,306,330	4	No
Nonlinear	SETAR(6)log	4,730,786	4	Yes	SETAR(3)log	75,637,050	5	No
		9,257,854	37	Yes		101,935,900	21	No

and the Random Walk and the AR(12) models are sometimes ranked among the top three models.

- Bayesian VAR models, particularly in differences and without Google data, perform rather well across all car brands and for short and medium forecasts (up to 12 steps ahead).
- Bivariate models including only car sales and Google models and using only the first and the 12th lags perform extremely well across most of the car brands examined, particularly for long-term forecasts. The parsimonious specifications of these models clearly allow for efficiency gains where forecasting is of concern.
- The forecasting power of the best models using Google data increases with the length of the forecast horizon, particularly with forecast horizons higher than 12 steps ahead. This evidence is similar to that found in D'Amuri and Marcucci (2013) and Fantazzini and Fomichev (2014).
- Models without Google data estimated with the long sample starting in 2001 tend to perform better than those estimated with a shorter sample starting in 2004.
- There are no particular differences between large, medium-sized and small sellers and between foreign and German manufacturers.

So as to provide an idea about how prediction errors evolve over time, Fig. 4 (columns 1 and 2 for seasonally adjusted data) shows the ratios of the MSPE of the best model with Google data and the Random Walk model across all forecasting horizons, together with the ratios of the MSPE of the best model without Google data and the Random Walk model. We remark that the best models tend to vary across different horizons.

The ratios in Fig. 4 show that it is difficult to outperform the random walk model in the case of short-term forecasts. Moreover,

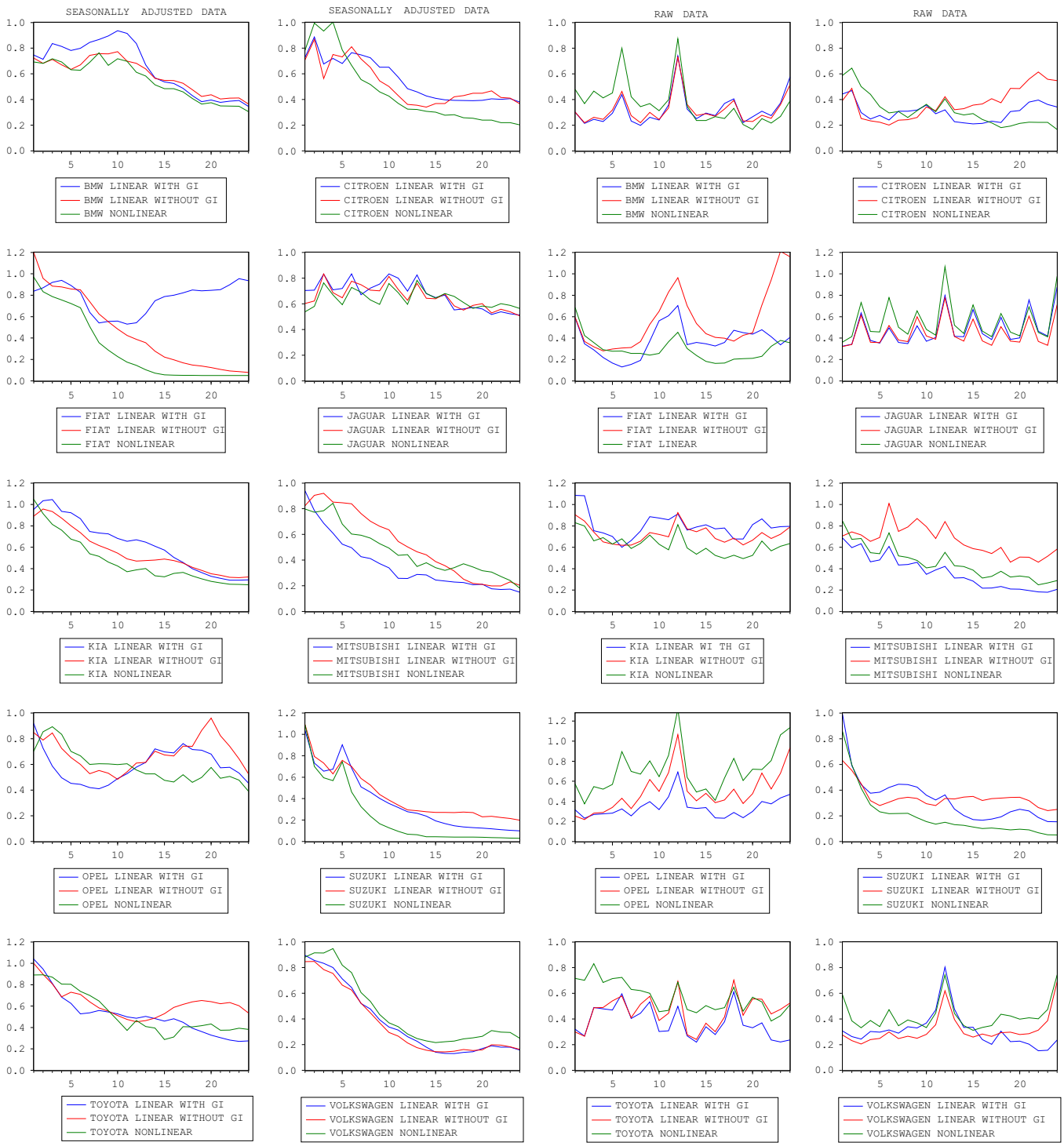


Fig. 5. Ratios of the MSPEs of the best models with and without Google data with those of the Random Walk model, together with the ratios of the MSPEs of the best nonlinear models and the Random Walk model across all forecasting horizons. The first two columns show results for seasonally adjusted data, and the last two for raw data.

the best models without Google data tend to perform better than the best models with Google data for short and medium forecasts, whereas in general models using Google data show lower MSPEs for long-term forecasts with horizons higher than 12 steps ahead. This evidence suggests that potential gains in terms of forecasting performance may be achieved by using forecast combination methods. The development of these methods is beyond the scope of this paper and will be the subject of future studies.

Model rankings in terms of the MSPE do not show whether the competing forecasts are statistically different or not. We therefore tested for significant differences in forecast accuracy using the Model Confidence Set (MCS) approach proposed by Hansen et al. (2011). The MCS is a sequential test of equal predictive ability, with the starting hypothesis that all models considered have equal forecasting performance. Given an initial set of forecasts, it tests the null that no forecast is distinguishable from any other and discards any inferior forecasts if they exist. The MCS procedure yields a model confidence set containing the best forecasting models at a given confidence level. Since our dataset is not too large and the number of forecasting models is moderate, we employed the semiquadratic test statistic (T_{SQ}), which is

more computationally intensive but more selective, see e.g. Rossi and Fantazzini (2016). The loss function used was the MSPE, while the p-values for the test statistic were obtained using a stationary block bootstrap with a block length of 12 months and 1000 re-samples. If the p-value was lower than a defined confidence level α, the model was not included in the MCS and viceversa. A brief description of the MCS approach is reported in the Technical Appendix F.

The models included in the MCS at the 10% level for all car brands and forecast horizons are reported in Table 16[5]: for the sake of space and interest, we report only the total number of selected models, the total number of selected Google-based models, and whether the Random Walk model was included or not. The full set of results is available from the authors upon request.

Table 16 shows that most, if not all, models are selected in the case of forecasts up to 10–12 steps ahead for five car brands out of ten: the differences in forecasting performances are not large enough to distinguish between them, meaning that the MCS contains a large number of models. Moreover, the Random Walk model is often included. Instead, for long-term forecasts (12 steps ahead and higher), only a small number of models is selected, most of them bivariate models including only car sales and GIs, Bayesian VARs with GIs and sometimes the AR(12). Besides, the Random Walk model is seldom included. Here, the data are much more informative and it is possible to select a limited number of models which statistically outperform their competitors.

3.2. Raw data

We compared the same 34 models used for seasonally adjusted data, but augmented with centered seasonal dummies to model potential seasonal behavior. Moreover, we also considered the bivariate Periodic Error Correction Model PECM(1,12) which includes only car sales and Google data, as discussed in Section 2.2.2. To account for the possible endogeneity of regressors and improve the efficiency of the parameter estimates in small samples, we estimated the error correction term using the method of dynamic OLS (see Boswijk and Franses, 1995; Hayashi, 2000; Franses and Paap, 2004). A short-cut notation for identifying each model in the subsequent tables reporting their forecasting performances is reported in the last column of Table 7.

We used the data between 2001M1 and 2009M6 as the first initialization sample for the models without Google data, while we used the initialization sample 2004M1-2009M6 for the models with Google data and for those without Google data but estimated on a shorter sample. The evaluation period ranged from 2009M7 till 2014M6 and was used to compare forecasts from 1 step ahead up to 24 steps ahead. The top three models in terms of the Mean Squared Prediction Error (MSPE) for each forecasting horizon and each car brand are reported in Table 17, while a summary of the models included in the MCS is reported in Table 18. The ratios of the MSPE of the best model with Google data and the Random Walk model across all forecasting horizons, together with the ratios of the MSPE of the best model without Google data and the Random Walk model are shown in the last two columns of Fig. 4.

The results are somewhat similar to those which emerged from seasonally adjusted data, but there are also some important differences. Models without Google data now perform better, with respect to the case of seasonally adjusted data. Moreover, the number of models selected in the MCS is now much smaller (often no more than 2–6 models): Bayesian VARs (with and without Google data) and parsimonious bivariate models including only sales and GIs again represent the majority of models included in the MCS at the 10% level.

[5] We set $\alpha = 0.10$ as in Hansen et al. (2011).

4. Robustness checks

We wanted to verify that the superior performance of Google-based models also holds under alternative forecasting. We performed a series of robustness checks, considering alternative nonlinear models, alternative out-of-sample intervals, evaluating the directional accuracy of the competing forecasting models, checking whether Google data downloaded on different days can affect the models' forecasting performances, and examining additional car brands.

4.1. Nonlinear models

A part of the economic and financial literature has suggested the use of nonlinear models for forecasting purposes (for instance, see Franses and Dijk, 2000 and Terasvirta et al., 2011 for a discussion at the textbook level). Given this evidence, we estimated a set of nonlinear models and compared their forecasting performances with the models in Section 3. More specifically, we considered three nonlinear models:

- the SETAR model with 2 regimes (see Tong, 1990 for a discussion at the textbook level);
- the logistic smooth transition autoregressive (LSTAR) model, which is a generalization of the SETAR model (see Tong, 1990);
- the additive autoregressive model (AAR), also known as generalized additive model (GAM), since it combines generalized linear models and additive models (see Wood, 2006 for a discussion at the textbook level).

A description of these nonlinear models is given in the Technical Appendix G. See D'Amuri and Marcucci (2013) and Fantazzini and Fomichev (2014) for a discussion of robustness checks using these nonlinear models.

The top three models in terms of the MSPE for each forecasting horizon and each car brand are reported in Table 19 for seasonally adjusted data and in Table 21 for raw data. A summary of the models included in the MCS is reported in Table 20 for seasonally adjusted data and in Table 22 for raw data.

In general, nonlinear models are very competitive, thus confirming past literature dealing with car sales forecasting (see Da et al., 2003; Kunhui et al., 2007; Brühl et al., 2009; Hulsmann et al., 2012). Particularly, parsimonious AAR and SETAR models involving only a few lags are often ranked among the top models in terms of MSPE. Moreover, AAR models with log-prices performed very well for medium- and long-term forecasts, similarly to what was found in Fantazzini and Fomichev (2014) when forecasting the real price of oil. However, nonlinear models were difficult to estimate, and specifications with a large number of lags failed to converge. Particularly, the LSTAR proved to be the most challenging and computationally intensive (see Franses and Dijk, 2000 for a discussion of this issue). The results of the MCS confirm this evidence and most of the models included at the 10% level are nonlinear, whereas the only selected linear models are mostly Google-based. This evidence therefore seems to suggest that Google data may explain a good portion of the nonlinearity displayed by sales data.

In the case of raw data, nonlinear models are less competitive than linear models, particularly for forecasting horizons up to 12 steps ahead, whereas Bayesian VAR models and bivariate linear models including car sales and GIs are often the top ranked models across most of the car brands. However, for long-term forecasts, more than half of the models included in the MCS are nonlinear, while the remaining selected models are mainly bivariate Google-based models.

Tables 8–11 report the MSPEs, rankings, and eventual inclusion in the MCS of the best models in the case of 6, 12, 18, 24 step-ahead forecasts, respectively, for four model classes: linear models

with GI, linear models without GI, nonlinear models and Random Walk models. Parsimonious bivariate models including only car sales and GIs are the best in the first class; AR(12) models and Bayesian models usually top the second class, while AAR and SETAR models with few lags are the best nonlinear models. The Random Walk has low rankings in long-term forecasts, but fares better for short-term forecasts.

The previous evidence is confirmed and summarized by Fig. 5 in Appendix C, which shows the ratios of the MSPEs of the best models with and without Google data with those of the Random Walk model, together with the ratios of the MSPEs of the best nonlinear models and the Random Walk model across all forecasting horizons: nonlinear models tend to perform better with seasonally adjusted data and medium- and long-term forecasts.

Finally, for the sake of interest (given the importance of long-term forecasts for car manufacturers) and space, we report in Tables 23 and 24 the list of models included in the Model Confidence Set for each car brand for 24 step-ahead forecasts, for seasonally adjusted data and raw data, respectively. In the latter case, the number of models selected is higher on average than for seasonally adjusted data, which was expected given the more noisy nature of raw data.

4.2. Alternative out-of-sample periods

Our baseline out-of-sample interval includes the global financial crisis which started in 2007 and had a strong effect on car sales. Moreover, our in-sample analysis highlighted a potential structural break in the years 2008–2009. Therefore, we want to verify that our results continue to hold with different business cycle conditions, as recently highlighted by D'Amuri and Marcucci (2013). We considered the following two alternative out-of-samples:

- 2008M10-2009M6: this sample includes the official period of recession in Germany.
- 2009M7-2014M6: this sample starts after the end of the recession.

Due to the dimensionality of these new out-of-samples, we considered forecasts up to only 8 steps ahead. Moreover, this robustness check was performed only with seasonally adjusted data, since the first forecast with raw data takes place after the end of the recession.[6] The top three models in terms of the MSPE for each forecasting horizon and each car brand are reported in Table 25 for the recession period, and in Table 26 for the expansion period.

The results are somewhat mixed and change substantially according to the car brand which is examined. However, some general indications can still be gained: Google-based models and linear models without Google data were the best models during the recession, while Google-based models and nonlinear models performed (slightly) better during the economic expansion. These results therefore provide further evidence of a structural break in the years 2008–2009. In general, Google-based models had forecasting performances which were more robust across different business cycles than their competitors, thus confirming similar evidence found by D'Amuri and Marcucci (2013) and Fantazzini and Fomichev (2014).

4.3. Directional accuracy

The analysis has so far only considered the accuracy of forecasts in terms of magnitude, but directional accuracy is also important: forecasts with the correct direction of change may still provide useful information even with large forecast errors. This is particularly important when predicting a turning point, which is a special case of directional accuracy and represents a change in the direction of movement of the analyzed variable (Theil, 1961; Naik and Leuthold, 1986).

The top three models in terms of average directional accuracy (in %) for each car brand, for short-term forecasts (1–6 steps ahead), medium-term forecasts (7–12 steps ahead), and long-term forecasts (13–24 steps ahead) are reported in Table 27 (top part) for seasonally adjusted data and in Table 27 (bottom part) for raw data.

In the case of seasonally adjusted data, parsimonious bivariate models, including only car sales and GIs, as well as AAR models had the higher percentage of correct forecasts of the direction of change for most of the car brands and forecasting horizons. As for raw data, similarly to what we saw in Section 4.1, nonlinear models are, in general, less competitive than linear models. More specifically, linear models without Google data performed better than with seasonally adjusted data (particularly for short-term directional forecasts), while nonlinear models were competitive only for medium- to long-term directional accuracy. Instead, Google-based models performed relatively well and simple bivariate models with car sales, GIs and centered seasonal dummies provided very precise forecasts of the direction of change for most of the car brands and forecasting horizons.

The somewhat differing results between seasonally adjusted data and raw data could be due to two reasons. Firstly, the procedure of seasonal adjustment changes the statistical properties of the data and can affect considerably the models' forecasting performances (Zellner, 1978; Franses and Paap, 2004). Secondly, Boivin and Ng (2006) and Stock and Watson (2006) have shown that small models may outperform models with a larger number of parameters because they allow for a better extraction of relevant signals than models overloaded with parameters and complex specifications. In this regard, Google data allow us to summarize a lot of information and reduce model complexity.

4.4. Sampling variability of google data

Google data does not refer to the population of searches, but only to a sample. As a consequence, the time series of Google data can vary substantially from one download to another.[7] We downloaded the GIs for a number of subsequent days to check how sampling variability can affect the models' forecasting performances. More specifically, we compared the forecasts computed in our baseline case with GIs downloaded on 15/08/2014, with forecasts computed with the average GIs downloaded between the 15/08/2014 and the 02/09/2014. We used the average GIs following the approach recently proposed by Carriere-Swallow and Labbé (2013). Table 28 shows the average ratio – averaged across all forecasting horizons – of the MSPE for the forecasts computed with GIs downloaded on the 15/08/2014, with respect to MSPE for the forecasts computed with the average GIs downloaded between 15/08/2014 and 02/09/2014.

Almost all models have ratios close to 1, with the notable exception of high-dimensional VEC models, which did not reach convergence for a couple of car brands (Toyota and Kia). The large variance of estimators for cointegrated models in small-medium samples is a well-known issue in the econometric literature (Stock and Watson (1993); Maddala and Kim (1998) (Section 5.7) and Hayashi (2000) (Section 10.4)): most likely, the sampling noise of Google data exacerbates this inference problem. Using average GIs can solve this issue to some extent, but not completely: the high-dimensional VECM models still did not reach convergence in some cases. Moreover, the rankings of Google-based models in the case of averaged data are very close, if not identical, to the rankings of Google-based models in the baseline case for all car brands (results not reported). Therefore, the

[6] Raw data required a larger initialization sample due to the inclusion of centered seasonal dummies.

[7] The authors want to thank an anonymous referee for pointing out this issue.

most advisable solution is probably either to use parsimonious VEC models or revert to Bayesian methods.

4.5. Additional car brands

In the baseline section, we analyzed 10 car brands out of the 22 car brands which both have monthly data continuously available since 2001 and are present in Google Trends. We briefly examine here the forecasting performances of the remaining 12 car brands:

- *Large sellers*: Ford, Audi;
- *Medium-sized sellers*: Hyundai, Mazda, Nissan, Peugeot, Renault;
- *Small sellers*: Honda, Land Rover, Porsche, Subaru, Volvo.

Tables 29 and 30 report the top three models in terms of MSPE for each forecasting horizon and each car brand, in the case of seasonally adjusted data and raw data, respectively.

The results are similar to those of the baseline case: parsimonious bivariate linear models involving only GIs and car sales and nonlinear models (with few lags) are the best models for all brands examined. Bayesian models are a valid alternative for short-term forecasting in the case of seasonally adjusted data.

5. Conclusions

This paper proposed a set of multivariate models for forecasting car sales using both Google data and economic variables. Moreover, we considered multivariate models for both deseasonalized data and for raw data. We performed a forecasting exercise for ten car brands in Germany, and we computed out-of-sample forecasts ranging from 1 month to 24 months ahead. Our results showed that Bayesian VAR models performed rather well for all car brands and for short- and medium-term forecasts, while parsimonious bivariate models including only car sales and Google models outperformed the competing models in the case of long-term forecasts for several brands. Furthermore, the forecasting power of the best Google-based models increased with the length of the forecast horizon, particularly with forecast horizons higher than 12 steps ahead. Apart from this, no particular differences between large, medium-sized and small sellers and between foreign and German manufacturers were found. In case of raw data, models without Google data performed better than in the case of seasonally adjusted data. However, Bayesian VARs (with and without Google data) and parsimonious bivariate models including only sales and Google data represented again the majority of models included in the MCS at the 10% level. Finally, we performed a set of robustness checks to verify that our results also hold under different forecasting setups. We found out that nonlinear AAR and SETAR models were very competitive and were included in the MCS together with Google-based models, thus suggesting that Google data may explain a part of the nonlinearity displayed by sales data. However, nonlinear models were difficult to estimate and on several occasions failed to converge. Alternative out-of-sample intervals highlighted that Google-based models performed better during the recession (which is of particular importance for car manufacturers) and, in general, they had forecasting performances which were more robust across different business cycles than their competitors. Our previous results also held in the case of directional accuracy, which showed that Google-based models provided the most precise forecasts of the direction of change. We found that the sampling variability of Google data can be problematic for high-dimensional VEC models. Using the averaged Google data over several days can solve this issue to some extent, but parsimonious VEC models and Bayesian methods are valid alternatives as well. The results in the baseline case also held for twelve additional car brands.

Even though we considered a very large set of models, we had to restrict their potential range in order to keep the forecasting exercise computationally tractable. An avenue of future research would be to consider additional models such as fractional cointegration, exponential smoothing methods in state space form, and many others.

Appendix A. In-sample analysis

See Tables 12, 13 and 14.

Appendix B. Forecasting performances: Baseline case

B.1. Seasonally adjusted data

See Tables 15 and 16.

Table 12
Weak exogeneity of seasonally adjusted data: variables for which the null hypothesis of weak exogeneity can be rejected after re-testing at the 5% probability level.

Car brand	BC	CCI	CPI	EURIBOR	PP	GDP	PI	UR	Google
BMW	V	V	V		V	V	V	V	V
Citroen	V	V	V	V	V	V	V	V	V
Fiat	V	V	V	V	V	V	V	V	V
Jaguar	V	V	V		V	V	V	V	V
Kia	V	V	V		V	V	V	V	V
Mitsubishi	V		V				V	V	V
Opel	V	V	V	V	V	V	V	V	V
Suzuki	V	V	V	V	V	V	V	V	V
Toyota	V	V	V	V	V	V	V	V	V
Volkswagen	V	V	V	V	V	V	V	V	V

Table 13
Weak exogeneity of raw data: variables for which the null hypothesis of weak exogeneity can be rejected after re-testing at the 5% probability level.

Car brand	BC	CCI	CPI	EURIBOR	PP	GDP	PI	UR	Google
BMW	V	V	V		V	V	V	V	V
Citroen	V	V	V	V	V	V	V	V	V
Fiat	V	V	V	V	V	V	V	V	V
Jaguar	V		V					V	V
Kia	V	V	V	V	V	V	V	V	V
Mitsubishi	V	V	V	V	V	V	V		V
Opel	V	V	V	V	V	V	V	V	V
Suzuki	V	V	V	V	V	V	V	V	V
Toyota	V	V	V	V	V	V	V	V	V
Volkswagen	V	V	V	V	V	V	V	V	V

Table 14
The null hypothesis is the absence of cointegration against the alternative of periodic cointegration. The Boskwijk test considered the case with seasonal intercepts.

Car brand	Sample: 2001–2014		Sample: 2008-2014	
	Boskwijk joint Wald test statistic	Weak exogeneity test of GIs (*p*-value)	Boskwijk joint Wald test statistic	Weak exogeneity test of GIs (*p*-value)
BMW	46.90*	0.86	39.08*	0.85
Citroen	45.75*	0.63	58.38*	0.30
Fiat	65.87*	0.15	116.11*	**0.00**
Jaguar	22.26	0.05	19.83	0.49
Kia	48.45*	0.06	31.83*	0.95
Mitsubishi	37.46*	0.66	29.17	0.62
Opel	49.26*	0.53	29.70	0.54
Suzuki	41.72*	0.61	67.46*	0.25
Toyota	69.64*	**0.01**	47.02*	**0.02**
Volkswagen	41.10*	0.56	48.33*	0.37

p-values smaller than 5% are reported in bold.

* Significance at the 5% level.

Table 15
Top three models in terms of MSPE for each forecasting horizon and each car brand.

Car brand	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8	Step 9	Step 10	Step 11	Step 12
BMW	AR12	AR12	AR12	AR12	AR12	AR12	AR12	AR12	AR12	AR12	AR12	AR12
	BVARNOGO4	BVAR	AR124	AR124	AR124	AR124	AR124	AR124	AR124	AR124	AR124	AR124
	BVAR	AR124	BVAR	BVAR	BVAR	BVAR	VADongo1112	VADongo1112	BVAR	BVARDNOGO4	VADongo1112	VARongo1112
OPEL	BVARDNOGO	VADongo12	VADongo12	VECongo1112	VECongo1112	VECongo1112	VECM	VECongo1112	VECongo1112	VECMXNOGO	VECongo1112	VECongo1112
	BVARDNOGO4	VECongo1112	VECongo1112	VADongo12	VECM	VECM	VECMX	VARongo1112	VARongo1112	VECMNOGO	VECMXNOGO	BVARNOGO
	BVARD	VARongo12	VARongo12	VECongo12	VECMX	VECMX	VECongo1112	VECMXNOGO	VECMXNOGO	VECongo1112	VECMNOGO	AR12
VOLKSWAGEN	BVARDNOGO	BVARDNOGO	VARNOGO	VARDNOGO	AR124	AR124	VARongo1112	AR124	AR124	AR124	AR124	AR124
	BVARDNOGO4	BVARD	VARXNOGO	VARXDNOGO	VARongo1112	VARongo1112	AR124	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	BVARD	BVARDNOGO4	VARDNOGO	VARNOGO	VECongo1112	AR12	AR12	AR12	AR12	AR12	AR12	AR12
CITROEN	VARDNOGO4	VECMXNOGO4	VECMXNOGO4	VARD	VARD	VARD	BVARNOGO	BVARNOGO	BVARNOGO	BVARNOGO	BVARNOGO	BVARNOGO
	VARXDNOGO4	VECMNOGO4	VECMNOGO4	VARXD	VARXD	VARXD	VARD	VARongo1112	VARongo1112	VARongo1112	VECongo1112	VECongo1112
	VECMXNOGO4	BVARDNOGO	VARD	VARDNOGO	VARDNOGO	VARDNOGO	VARXD	VAR	BVAR	VECongo1112	VARongo1112	VARongo1112
FIAT	VECongo1112	VADongo1112	AR12	AR12	AR12	VARongo1112	VARongo1112	VARongo1112	BVARNOGO	BVARNOGO	BVARNOGO	BVARNOGO
	VARongo1112	VECongo1112	VADongo1112	VADongo1112	BVARNOGO	BVARNOGO	BVARNOGO	BVARNOGO	VARongo1112	VARongo1112	AR12	AR124
	VADongo1112	AR12	VECongo1112	VECongo1112	VECongo1112	AR12	AR12	AR12	AR12	AR12	AR124	AR12
TOYOTA	RW	VECMXNOGO	VARNOGO	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	BVAR	BVARNOGO4	BVARNOGO4	BVARNOGO4
	VARNOGO	VECMNOGO	VARXNOGO	VARNOGO	VARNOGO	BVAR	BVAR	BVAR	BVARNOGO4	BVAR	BVAR	BVAR
	VARXNOGO	VECongo1112	VECongo1112	VARXNOGO	VARXNOGO	VECMXNOGO	BVARNOGO4	BVARNOGO4	VECongo1112	VECongo1112	VECongo1112	VECongo1112
JAGUAR	AR12	AR12	VADongo1112	AR12	AR12	AR12	VECongo1112	AR12	AR12	AR12	AR12	AR12
	AR124	AR124	AR12	VECongo1112	AR124	VADongo1112	AR12	VECongo1112	VECongo1112	VECongo1112	AR124	AR124
	BVARDNOGO	BVARDNOGO	BVARDNOGO4	AR124	VECongo1112	BVARDNOGO4	AR124	AR124	BVARDNOGO4	BVARD	VECongo1112	VECongo1112
KIA	BVARDNOGO	AR12	AR124	BVARNOGO4	AR12	AR12	AR12	AR12	AR12	AR12	AR12	AR12
	BVARDNOGO4	AR124	VARDNOGO4	AR12	AR124	AR124	AR124	AR124	AR124	AR124	AR124	AR124
	BVARD	BVARDNOGO	AR12	AR124	BVARNOGO4	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
MITSUBISHI	BVARDNOGO	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo12	VECongo12	VECongo12
	BVARDNOGO4	BVARNOGO4	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VECongo1112	VARongo1112	VECongo1112
	BVARNOGO4	VADongo1112	VADongo1112	VADongo1112	VECongo12	VECongo12	VARongo12	VARongo12	VARongo12	VARongo1112	VARongo12	VARongo1112
SUZUKI	RW	VARongo1112	VARongo1112	VECMXNOGO	VECMXNOGO	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VECongo1112	VECongo1112	BVAR
	RW4	BVARNOGO	BVARNOGO	VECMNOGO	VECMNOGO	VECMXNOGO	BVAR	VECongo1112	VECongo1112	BVARNOGO	BVAR	VECongo1112
	VARongo1112	BVARNOGO4	BVARNOGO4	VARongo1112	BVARNOGO	VECMNOGO	BVARNOGO	BVARNOGO	BVARNOGO	BVAR	BVARNOGO4	BVARNOGO4

Car brand	Step 13	Step 14	Step 15	Step 16	Step 17	Step 18	Step 19	Step 20	Step 21	Step 22	Step 23	Step 24
BMW	AR12	AR12	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	AR124	VARongo1112	AR12	AR12	AR12	AR12	AR124	AR124	AR124	AR124	AR124	AR124
	VARongo1112	AR124	AR124	AR124	AR124	AR124	AR12	AR12	AR12	AR12	AR12	AR12
OPEL	VECongo1112	AR12	AR12	AR12	AR12	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112
	AR12	VECongo1112	VECongo1112	VECongo1112	VECongo1112	AR12	VADongo12	VARongo1112	VECMXNOGO	VECMXNOGO	VECMXNOGO	VECMXNOGO
	BVARNOGO	AR124	AR124	AR124	AR124	VADongo1112	AR12	VADongo1112	VECMNOGO	VECMNOGO	VECMNOGO	VECMNOGO
VOLKSWAGEN	AR124	AR124	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	AR124	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	AR12	VARongo1112	AR124	AR124	AR124	AR124	AR124	AR12	AR124	AR124	AR124	AR124
	VARongo1112	AR12	AR12	AR12	AR12	AR12	AR12	VARongo1112	AR12	AR12	AR12	AR12
CITROEN	BVARNOGO	BVARNOGO	BVARNOGO	BVARNOGO	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	AR124
	VECongo1112	VECongo1112	VECongo1112	VECongo1112	BVARNOGO	BVARNOGO	BVARNOGO	BVARNOGO	VARongo1112	AR124	AR124	VARongo1112
	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	AR124	VARongo1112	VARongo1112	VECongo1112
FIAT	BVARNOGO	AR124	AR124	AR124	AR124	AR124	AR124	AR124	AR124	AR124	AR124	AR124
	AR124	BVARNOGO	AR12	AR12	AR12	AR12	AR12	AR12	AR12	AR12	AR12	AR12
	AR12	AR12	BVARNOGO	VARDNOGO	VARDNOGO	VARDNOGO	VARDNOGO	VARDNOGO	VARDNOGO	VARDNOGO	VARDNOGO	VARDNOGO
TOYOTA	BVARNOGO4	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112
	VECongo1112	BVARNOGO4	BVARNOGO4	BVARNOGO4	BVARNOGO4	BVARNOGO4	AR12	AR12	AR12	VECMXNOGO	VECMXNOGO	VECMXNOGO
	BVAR	BVAR	BVAR	BVAR	BVAR	AR124	BVARNOGO4	BVARNOGO4	AR124	VECMNOGO	VECMNOGO	VECMNOGO
JAGUAR	AR12	AR12	AR12	VECongo1112	VECongo1112	AR12	VECongo1112	VECongo1112	VARongo1112	VARongo12	VECongo1112	AR12
	AR124	VECongo1112	VECongo1112	AR12	AR12	VECongo1112	AR12	AR12	VARongo12	AR12	AR12	VECongo1112
	VECongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	AR124	VARongo1112	AR12	VARongo1112	VARongo1112	AR124

Table 15 (*continued*)

	Step 13	Step 14	Step 15	Step 16	Step 17	Step 18	Step 19	Step 20	Step 21	Step 22	Step 23	Step 24
KIA	AR12	AR12	AR12	AR124	AR124	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	AR124	AR124	AR124	AR12	VARongo1112	AR124	AR124	AR124	AR124	AR124	AR124	AR124
	VARongo1112	VARongo1112	VARongo1112	VARongo1112	AR12	AR12	AR12	AR12	AR12	AR12	AR12	AR12
MITSUBISHI	VECongo1112	VECongo1112	VARongo1112	VARongo1112	VARongo1112	VECongo1112	VECongo1112	VECMNOGO4	VECongo1112	VECongo1112	VECongo1112	VECongo1112
	VARongo1112	VARongo1112	VECongo1112	VECongo1112	VECongo1112	VARongo1112	VECMNOGO4	VARongo12	VECMNOGO4	VECMNOGO4	VECMNOGO4	VECMNOGO4
	VECongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo1112	VECongo1112	VARongo1112	VARongo1112	VARongo12	VARongo12
SUZUKI	BVAR	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	BVARNOGO4	BVAR	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112
	BVARNOGO	BVARNOGO4	BVAR	BVAR	BVAR	BVAR	BVAR	VECMXNOGO	VECMXNOGO	VECMXNOGO	VECMXNOGO	VECMXNOGO

B.1.1. Raw data

See Tables 17 and 18.

Appendix C. Robustness checks: Nonlinear models

C.0.2. Seasonally adjusted data

See Tables 19 and 20.

C.0.3. Raw data

See Tables 21 and 22.

C.0.4. List of models included in the Model Confidence Set. Forecast horizon: 24 steps ahead. Seasonally adjusted data

See Table 23.

C.0.5. List of models included in the Model Confidence Set. Forecast horizon: 24 steps ahead. Raw data

See Table 24.

Appendix D. Robustness checks: Alternative out-of-sample periods

D.0.6. Seasonally adjusted data – Recession

See Table 25.

D.0.7. Seasonally adjusted data – Expansion

See Table 26.

Appendix E. Robustness checks: Directional accuracy

See Table 27.

Appendix F. Robustness checks: Variability of Google data

See Table 28.

Appendix G. Robustness checks: Additional car brands

G.0.8. Seasonally adjusted data

See Table 29.

G.0.9. Raw data

See Table 30.

Table 16
Models included in the Model Confidence Set for each forecast horizon (from 1 step up to 24 steps ahead) and for each car brand.

BMW	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	15	34	32	34	34	33	22	21	22	32	31	31	28	21	23	22	6	7	6	6	6	6	6	7
Google models	6	14	13	14	14	13	10	10	10	13	13	13	13	10	12	11	3	3	3	3	3	3	3	3
Random Walk included?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No
OPEL	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	5	4	7	8	2	2	1	5	5	1	2	4	8	9	9	8	9	10	9	8	4	2	2	2
Google models	1	3	6	6	2	2	1	3	3	0	1	1	4	3	4	3	4	4	4	4	3	1	1	1
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No
VOLKSWAGEN	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	6	4	1	1	3	4	2	1	2	2	2	2	2	3	5	5	4	3	4	4	4	3	5	4
Google models	2	1	0	0	1	1	1	0	1	1	1	1	1	1	3	3	2	1	2	2	2	1	3	2
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
CITROEN	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	1	1	1	34	34	34	2	4	4	4	4	4	4	4	5	7	10	9	7	8	9	9	13	13
Google models	0	0	0	14	14	14	0	1	2	2	2	2	2	2	3	4	4	4	3	4	5	5	6	6
Random Walk included?	No	No	No	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No	No	No
FIAT	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	6	7	8	11	12	16	10	10	11	9	6	5	4	4	4	4	4	2	2	1	1	1	1	1
Google models	3	2	3	4	4	7	4	3	4	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0
Random Walk included?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No
TOYOTA	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	2	1	1	2	2	3	4	4	4	4	5	5	6	7	7	11	9	2	2	2	1	2	2	3
Google models	0	0	0	1	1	2	2	2	2	2	2	2	2	3	3	5	3	1	1	2	1	1	1	2
Random Walk included?	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No
JAGUAR	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	27	34	9	21	22	24	22	23	29	22	11	15	17	7	5	6	4	7	4	1	7	6	7	7
Google models	11	14	4	9	11	10	9	10	14	10	5	8	9	4	3	3	3	5	2	1	4	3	4	4
Random Walk included?	No	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No
KIA	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	20	32	34	31	26	28	33	34	34	23	25	10	21	19	19	19	16	10	10	8	8	34	30	4
Google models	8	13	14	14	10	11	14	14	14	10	10	5	12	9	9	9	8	6	5	3	3	14	13	1
Random Walk included?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
MITSUBISHI	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	28	33	34	33	33	34	32	31	7	11	9	8	11	11	6	4	4	5	5	7	6	5	8	9
Google models	10	14	14	14	14	14	14	13	5	6	6	6	7	7	4	3	3	4	3	4	4	3	5	6
Random Walk included?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
SUZUKI	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	9	9	3	1	1	2	6	6	6	6	7	7	7	6	5	7	5	4	2	2	2	2	2	2
Google models	4	4	1	0	0	1	3	3	3	3	4	4	4	3	3	4	3	3	2	2	2	2	2	2
Random Walk included?	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No

Table 17
Top three models in terms of MSPE for each forecasting horizon and each car brand.

Car brand	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8	Step 9	Step 10	Step 11	Step 12
BMW	BVARPNOGO4	BVARPD	VADongo1112	VADongo1112	VADongo1112	VADongo1112	VADongo1112	VADongo1112	VADongo1112	VADongo1112	AR12	AR124
	BVARP	BVARP	BVARPDNOGO	BVARPDNOGO	AR12	BVARPDNOGO	BVARPD	BVARPD	BVARPDNOGO	AR12	AR124	AR12
	BVARPNOGO	BVARPDNOGO	BVARPDNOGO4	BVARPD	BVARPDNOGO	AR12	BVARPDNOGO4	BVARPDNOGO4	AR12	AR124	VADongo1112	VADongo1112
OPEL	BVARPDNOGO	BVARPNOGO	VARongo1112	VARP	VARP	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112
	BVARPNOGO	VEPongo1112	VEPongo1112	VARXP	VARXP	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VADongo1112	VADongo1112
	BVARPDNOGO4	BVARPDNOGO	BVARPD	VEPongo1112	VEPongo1112	VARP	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	VARongo1112	RW
VOLKSWAGEN	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO
	VECMXPNOGO	BVARP	VARPNOGO	VARPNOGO	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	BVARP	VEPongo1112	VEPongo1112	BVARP
	VECMPNOGO	BVARPDNOGO	VARXPNOGO	VARXPNOGO	VARPNOGO	VADongo1112	VARPNOGO	BVARP	VEPongo1112	BVARP	VADongo1112	VADongo1112
CITROEN	VARPNOGO4	VARongo1112	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	VARongo1112	VARongo1112
	VARXPNOGO4	BVARPNOGO	BVARPNOGO4	VARongo1112	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	VARongo1112	BVARPNOGO	VARongo12
	BVARPNOGO4	BVARPNOGO4	VARongo1112	BVARPNOGO4	VARPNOGO	VARongo1112	VARPNOGO	VARPNOGO	VARongo1112	BVARPNOGO4	BVARPNOGO4	BVARPNOGO
FIAT	BVARP	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	BVARP	VEPongo1112	PECM
	BVARPDNOGO4	VEPongo1112	VADongo1112	VADongo1112	VADongo1112	VADongo12	VECPongo12	VADongo1112	VEPongo1112	VEPongo1112	PECM	VEPongo1112
	BVARPNOGO	VARPDNOGO	BVARPDNOGO4	BVARPNOGO	BVARPNOGO	BVARPNOGO	VARongo12	BVARP	VADongo1112	PECM	VADongo1112	BVARP
TOYOTA	BVARPNOGO4	BVARPDNOGO	BVARPDNOGO	VARongo1112	VEPongo1112	VECMXPNOGO4	VEPongo1112	BVARP	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112
	BVARPDNOGO	BVARPD	BVARPD	VEPongo1112	BVARP	VEPongo1112	VECMXPNOGO4	VARongo1112	BVARP	BVARP	BVARP	BVARP
	BVARP	BVARPDNOGO4	BVARPDNOGO4	BVARPNOGO4	VARongo1112	BVARP	VARongo1112	VEPongo1112	VADongo12	VARongo1112	VARongo1112	VARongo1112
JAGUAR	BVARPNOGO4	BVARPNOGO4	VARPNOGO	VARPNOGO	VARongo12	BVARPD	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VARPNOGO	BVARPDNOGO4
	BVARP	BVARP	VARongo1112	VARongo12	VARPNOGO	BVARPDNOGO4	VARongo12	VARPNOGO	VARongo12	VARPNOGO	BVARPDNOGO4	BVARPD
	VARongo1112	VEPongo1112	VARPNOGO4	VARPNOGO4	BVARPNOGO4	VADongo1112	VARPNOGO	VARongo12	VARongo1112	VARongo12	BVARPD	BVARPDNOGO
KIA	AR12	BVARPDNOGO	BVARPDNOGO	AR12	AR12	BVARPD	AR12	AR12	AR12	AR12	AR12	VADongo1112
	VARPNOGO	BVARPNOGO	VARongo1112	VARongo1112	BVARPDNOGO	VADongo1112	BVARP	AR124	BVARPDNOGO	VARongo1112	BVARPDNOGO	BVARPDNOGO
	VARXPNOGO	AR12	BVARPNOGO	AR124	BVARP	BVARPDNOGO	VECMXPNOGO	VARongo1112	BVARPDNOGO4	AR124	AR124	BVARPDNOGO4
MITSUBISHI	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112
	BVARPDNOGO	BVARPD	BVARPD	BVARPD	BVARPDNOGO	VADongo1112	PECM	PECM	PECM	PECM	VECPongo12	PECM
	BVARPD	VADongo1112	BVARPDNOGO	BVARPDNOGO	VADongo1112	RW4	BVARPDNOGO	BVARPDNOGO	VECPongo12	VECPongo12	PECM	VECPongo12
SUZUKI	BVARPDNOGO	BVARPNOGO	VARongo1112	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4
	BVARPDNOGO4	BVARPDNOGO	BVARPNOGO4	BVARPNOGO	VARongo1112	BVARP	BVARP	BVARP	BVARP	BVARP	BVARP	BVARP
	BVARPD	BVARPNOGO4	BVARPNOGO	VARongo1112	BVARP	VARongo1112	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO

	Step 13	Step 14	Step 15	Step 16	Step 17	Step 18	Step 19	Step 20	Step 21	Step 22	Step 23	Step 24
BMW	VADongo1112	VADongo1112	AR12	AR12	AR124	AR124	VARPD	AR124	AR124	AR124	VECMPNOGO	AR12
	AR12	AR12	VADongo1112	AR124	AR12	AR12	AR124	AR12	AR12	AR12	VECMPNOGO4	AR124
	AR124	BVARPD	AR124	VADongo1112	VADongo1112	VARPD	AR12	VECMP	VECMPNOGO	VECMP	AR12	VARPNOGO
OPEL	VEPongo1112	VEPongo1112	VEPongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	VADongo1112	VADongo1112	VADongo12	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112
	BVARPD	BVARPDNOGO	VARongo1112	VADongo12	BVARPDNOGO	BVARPDNOGO	BVARPDNOGO	BVARPDNOGO	BVARPDNOGO	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO4
VOLKSWAGEN	BVARPNOGO	BVARPNOGO	BVARPNOGO	VECPongo12	VECPongo12	BVARPNOGO	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VECPongo12
	BVARP	AR124	AR124	BVARPNOGO	BVARPNOGO	VECPongo12	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	VADongo1112	VEPongo1112
	VEPongo1112	BVARP	AR12	VARongo1112	VARongo1112	VEPongo1112	BVARPDNOGO	BVARP	VADongo1112	VADongo1112	VECPongo12	VADongo1112
CITROEN	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	BVARPNOGO	BVARPNOGO	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VARongo12	VEPongo1112	VEPongo1112
	BVARPNOGO4	VARongo12	BVARPNOGO	BVARPNOGO	VADongo1112	VADongo1112	VADongo1112	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO
FIAT	PECM	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	AR12	AR12	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	VEPongo1112	PECM	VADongo1112	VADongo1112	AR12	VEPongo1112	VARongo1112	VARPDNOGO	VARPDNOGO	VADongo12	VEPongo1112	VEPongo1112
	BVARP	VADongo1112	PECM	BVARPDNOGO4	BVARPDNOGO4	VARPDNOGO	BVARPDNOGO	VARXPDNOGO	VARXPDNOGO	VEPongo1112	VADongo12	RW
TOYOTA	BVARP	BVARPD	VARongo1112	BVARPD	VARongo1112	VEPongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	BVARPD	BVARPDNOGO4	BVARPD	BVARPDNOGO	BVARPD	VARongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112
	VEPongo1112	BVARPDNOGO	BVARPDNOGO	VARongo1112	BVARPDNOGO	BVARPD	VARPNOGO4	VARPNOGO4	VARPNOGO4	VECMXPNOGO	BVARPD	VADongo12
JAGUAR	BVARPDNOGO4	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO
	VEPongo1112	VARPNOGO4	VARPNOGO4	VARPNOGO4	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VARPNOGO4	VEPongo1112	VARPNOGO4	VARPNOGO4

	BVARPD	VEPongo1112	VEPongo1112	VEPongo1112	VARPNOGO4	VARPNOGO4	VARPNOGO4	VARPNOGO4	VEPongo1112	VARPNOGO4	VADongo12	VEPongo1112
KIA	VARongo1112	AR12	AR124	AR124	AR124	VEPongo1112	AR124	AR124	AR124	AR124	AR124	AR124
	AR12	AR124	AR12	AR12	AR12	AR124	AR12	AR12	AR12	AR12	VARongo12	VARongo12
	VADongo1112	VADongo1112	VARongo1112	VARongo1112	VARongo1112	AR12	VARongo1112	VARongo1112	VARongo12	VARongo12	AR12	AR12
MITSUBISHI	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VARongo1112	VARongo1112	VEPongo1112	VARongo1112
	PECM	PECM	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VEPongo1112	VEPongo1112	VARongo1112	VEPongo1112
	VECPongo12	VARongo1112	VECPongo12	VECPongo12	VECPongo12	VECPongo12	VECPongo12	VECPongo12	VECPongo12	VECPongo12	VARongo12	VECPongo12
SUZUKI	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPD	BVARPD	BVARPD	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4
	VADongo1112	BVARPDNOGO	BVARPDNOGO	BVARPD	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	BVARPD	VADongo1112	VARPNOGO	BVARPNOGO	BVARPD

Table 18
Models included in the Model Confidence Set for each forecast horizon (from 1 step up to 24 steps ahead) and for each car brand.

BMW	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	22	14	16	18	20	12	18	16	17	10	5	4	6	7	5	5	8	18	17	22	21	21	19	16
Google models	10	5	8	8	10	5	9	8	8	8	2	1	2	2	1	2	4	7	9	11	10	10	9	6
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
OPEL	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	8	8	10	1	1	3	5	8	8	8	6	5	10	12	13	10	10	10	4	5	4	4	4	5
Google models	4	5	6	1	1	3	3	4	4	4	4	3	5	6	8	6	6	6	3	3	3	2	2	2
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	Yes
VOLKSWAGEN	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	2	4	2	2	3	4	3	4	4	5	4	5	6	7	14	15	15	13	11	11	10	5	2	2
Google models	0	1	0	0	1	2	1	2	2	3	2	3	4	3	7	8	8	8	7	6	6	4	2	2
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
CITROEN	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	1	4	4	5	3	4	4	4	3	4	5	7	8	8	11	9	8	8	9	11	13	10	11	10
Google models	0	1	1	1	0	1	1	1	1	1	2	3	4	4	6	4	4	4	5	6	7	6	6	6
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
FIAT	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	5	4	8	6	9	10	12	10	13	12	12	8	9	9	8	9	8	3	4	2	3	1	2	4
Google models	2	2	4	3	6	8	8	7	7	6	6	4	5	5	5	5	3	1	1	1	1	1	2	2
Random Walk included?	No	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	Yes
TOYOTA	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	12	11	7	5	7	8	13	16	15	9	11	9	10	9	13	11	11	9	3	5	3	8	8	11
Google models	5	5	2	3	4	5	5	7	6	4	4	4	5	5	4	5	4	3	2	3	2	3	3	5
Random Walk included?	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No	Yes
JAGUAR	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	15	25	25	25	29	12	19	20	20	26	22	11	18	15	9	12	14	11	13	17	9	3	6	14
Google models	8	14	12	12	15	6	10	10	10	13	10	5	10	9	4	6	7	4	6	9	4	1	2	5
Random Walk included?	No	No	Yes	No	No	No	No	No	No	No	No	Yes	No	No	Yes	No	No	No	No	No	No	No	No	Yes
KIA	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	2	11	10	8	13	12	3	6	11	11	12	12	15	13	14	15	13	14	14	11	11	9	7	7
Google models	0	2	4	2	5	5	1	2	4	3	4	5	7	5	6	6	5	6	7	4	4	4	2	2
Random Walk included?	No	Yes	No	No	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
MITSUBISHI	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	12	7	12	12	12	15	14	1	4	4	6	6	5	5	4	4	4	4	2	2	2	3	3	4
Google models	6	4	5	6	7	8	8	1	4	4	5	5	5	5	4	4	4	4	2	2	2	3	1	3
Random Walk included?	No	No	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
SUZUKI	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	2	10	8	5	9	6	11	10	9	11	12	13	12	13	11	12	10	9	9	9	5	4	4	5
Google models	0	4	3	2	3	3	4	6	3	5	5	6	5	5	6	6	6	4	4	5	3	2	2	2
Random Walk included?	No	No	No	No	No	Yes	Yes	No	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No

Table 19
Top three models in terms of MSPE for each forecasting horizon and each car brand.

Car brand	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8	Step 9	Step 10	Step 11	Step 12
BMW	AAR(6)log	SETAR(1)log	AR12	AR12	SETAR(8)log	AAR(7)dlog	AAR(6)dlog	AR12	AAR(5)dlog	SETAR(1)log	SETAR(1)log	SETAR(1)log
	AAR(8)log	SETAR(2)log	SETAR(6)dlog	SETAR(1)log	AR12	AAR(6)dlog	SETAR(3)dlog	SETAR(1)log	AAR(6)dlog	AAR(5)dlog	AR12	AAR(1)log
	AAR(7)log	AR12	SETAR(5)dlog	SETAR(6)log	SETAR(1)log	AAR(9)dlog	AR12	SETAR(2)log	AR12	SETAR(2)log	AAR(1)log	LSTAR(1)log
OPEL	LSTAR(3)dlog	VADongo12	VADongo12	VECongo1112	VECongo1112	VECongo1112	VECM	VECongo1112	VECongo1112	VECMXNOGO	VECongo1112	SETAR(7)dlog
	LSTAR(2)dlog	VECongo1112	VECongo1112	VADongo12	VECM	VECM	VECMX	VARongo1112	VARongo1112	VECMNOGO	VECMXNOGO	AAR(8)dlog
	LSTAR(4)dlog	VARongo12	VARongo12	VECongo12	VECMX	VECMX	VECongo1112	VECMXNOGO	VECMXNOGO	VECongo1112	VECMNOGO	VECongo1112
VOLKSWAGEN	BVARDNOGO	BVARDNOGO	VARNOGO	VARDNOGO	AR124	AR124	VARongo1112	AR124	AR124	AR124	AR124	AR124
	BVARDNOGO4	BVARD	VARXNOGO	VARXDNOGO	VARongo1112	VARongo1112	AR124	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	AAR(1)dlog	BVARDNOGO4	VARDNOGO	VARNOGO	VECongo1112	AR12	AR12	AR12	AR12	AR12	AR12	AR12
CITROEN	VARDNOGO4	VECMXNOGO4	VECMXNOGO4	VARD	VARD	LSTAR(3)log	LSTAR(3)log	LSTAR(3)log	SETAR(4)log	SETAR(4)log	SETAR(4)log	SETAR(4)log
	VARXDNOGO4	VECMNOGO4	VECMNOGO4	VARXD	VARXD	SETAR(4)log	SETAR(4)log	SETAR(4)log	LSTAR(3)log	SETAR(5)log	LSTAR(3)log	LSTAR(3)log
	VECMXNOGO4	BVARDNOGO	VARD	VARDNOGO	VARDNOGO	SETAR(5)log	SETAR(5)log	SETAR(5)log	SETAR(5)log	LSTAR(3)log	SETAR(3)log	SETAR(3)log
FIAT	VECongo1112	AAR(1)log	AAR(1)log	SETAR(4)log	SETAR(2)log	SETAR(2)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log
	VARongo1112	VADongo1112	SETAR(4)log	AAR(1)log	AAR(1)log	AAR(1)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log
	VADongo1112	VECongo1112	SETAR(3)log	SETAR(5)log	SETAR(4)log	SETAR(4)log	LSTAR(2)log	SETAR(1)log	SETAR(1)log	SETAR(1)log	SETAR(1)log	SETAR(1)log
TOYOTA	AAR(1)log	AAR(1)log	VARNOGO	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	BVAR	SETAR(6)dlog	SETAR(6)dlog	BVARNOGO4
	SETAR(2)log	VECMXNOGO	VARXNOGO	VARNOGO	VARNOGO	BVAR	BVAR	BVAR	BVARNOGO4	BVARNOGO4	BVARNOGO4	SETAR(7)dlog
	SETAR(1)log	VECMNOGO	VECongo1112	VARXNOGO	VARXNOGO	VECMXNOGO	BVARNOGO4	BVARNOGO4	SETAR(6)dlog	BVAR	BVAR	SETAR(5)dlog
JAGUAR	AAR(4)log	SETAR(4)dlog	AAR(3)log	AAR(3)log	AAR(4)log	SETAR(5)log	VECongo1112	SETAR(3)log	LSTAR(4)log	LSTAR(4)log	LSTAR(4)log	SETAR(3)log
	AAR(3)log	AAR(3)log	AAR(4)log	LSTAR(3)log	SETAR(5)log	LSTAR(4)log	SETAR(5)log	LSTAR(4)log	LSTAR(5)log	SETAR(3)log	SETAR(3)log	LSTAR(5)log
	AAR(5)log	SETAR(3)dlog	LSTAR(1)dlog	AAR(4)log	SETAR(3)log	SETAR(4)log	AAR(4)log	SETAR(4)log	SETAR(3)log	AAR(4)log	AAR(4)log	AR12
KIA	BVARDNOGO	AAR(7)log	AAR(7)log	AAR(7)log	AAR(7)log	AAR(4)log	SETAR(3)log	SETAR(3)log	AAR(7)log	SETAR(3)log	SETAR(3)log	SETAR(3)log
	BVARDNOGO4	SETAR(9)log	AAR(4)log	AAR(4)log	AAR(4)log	AAR(2)log	AAR(4)log	AAR(4)log	SETAR(3)log	LSTAR(1)log	SETAR(1)log	LSTAR(1)log
	BVARD	SETAR(1)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(7)log	AAR(2)log	LSTAR(1)log	LSTAR(1)log	AAR(7)log	LSTAR(1)log	AAR(4)log
MITSUBISHI	LSTAR(2)log	LSTAR(2)log	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo12	VECongo12	VECongo12
	BVARDNOGO	VECongo1112	AAR(4)dlog	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VECongo1112	VARongo1112	VECongo1112
	LSTAR(5)log	AAR(2)dlog	AAR(5)dlog	AAR(4)log	SETAR(1)log	SETAR(1)log	VARongo12	VARongo12	VARongo12	VARongo1112	VARongo12	VARongo1112
SUZUKI	RW	SETAR(3)log	SETAR(3)log	SETAR(3)log	SETAR(3)log	SETAR(3)log	SETAR(3)log	SETAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log
	RW4	SETAR(2)log	SETAR(5)log	SETAR(5)log	VECMXNOGO	SETAR(2)log	SETAR(2)log	AAR(3)log	SETAR(3)log	SETAR(3)log	SETAR(3)log	SETAR(3)log
	VARongo1112	VARongo1112	AAR(4)log	VECMXNOGO	VECMNOGO	LSTAR(1)log	AAR(3)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log

	Step 13	Step 14	Step 15	Step 16	Step 17	Step 18	Step 19	Step 20	Step 21	Step 22	Step 23	Step 24
BMW	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log
	AR12	AR12	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	SETAR(1)log	VARongo1112	AR12	AR12	AR12	AR12	AR124	AR124	AR124	AR124	AR124	AR124
OPEL	SETAR(7)dlog	AAR(8)dlog	AAR(8)dlog	AAR(8)dlog	SETAR(7)dlog	SETAR(7)dlog	AAR(8)dlog	AAR(8)dlog	SETAR(7)dlog	SETAR(7)dlog	SETAR(8)dlog	SETAR(8)dlog
	AAR(8)dlog	SETAR(7)dlog	SETAR(7)dlog	SETAR(7)dlog	AAR(8)dlog	AAR(8)dlog	SETAR(7)dlog	SETAR(7)dlog	SETAR(8)dlog	SETAR(8)dlog	SETAR(7)dlog	VECongo1112
	SETAR(8)dlog	SETAR(8)dlog	SETAR(8)dlog	SETAR(8)dlog	AAR(7)dlog	AAR(7)dlog	AAR(3)dlog	SETAR(5)dlog	AAR(8)dlog	AAR(8)dlog	VECongo1112	SETAR(7)dlog
VOLKSWAGEN	AR124	AR124	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	AR124	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	AR12	VARongo1112	AR124	AR124	AR124	AR124	AR124	AR12	AR124	AR124	AR124	AR124
	VARongo1112	AR12	AR12	AR12	AR12	AR12	AR12	VARongo1112	AR12	AR12	AR12	AR12
CITROEN	SETAR(4)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log
	LSTAR(3)log	BVARNOGO	LSTAR(3)log	SETAR(4)log	LSTAR(3)log	SETAR(4)log	LSTAR(3)log	SETAR(3)log	SETAR(4)log	SETAR(3)log	SETAR(4)log	SETAR(4)log
	AAR(1)log	SETAR(4)log	SETAR(4)log	LSTAR(3)log	SETAR(4)log	SETAR(5)log	SETAR(4)log	LSTAR(3)log	LSTAR(3)log	LSTAR(3)log	LSTAR(3)log	SETAR(3)log
FIAT	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log
	SETAR(2)log	SETAR(2)log	LSTAR(2)log	LSTAR(2)log	LSTAR(2)log	LSTAR(2)log	SETAR(2)log	AR124	AR124	AR124	AR124	AR124
	SETAR(1)log	LSTAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	AR124	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log
TOYOTA	SETAR(5)dlog	SETAR(5)dlog	SETAR(6)dlog	SETAR(6)dlog	SETAR(3)dlog	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112
	SETAR(4)dlog	SETAR(4)dlog	SETAR(5)dlog	SETAR(3)dlog	SETAR(5)dlog	SETAR(3)dlog	SETAR(3)dlog	SETAR(3)dlog	SETAR(6)dlog	SETAR(6)dlog	SETAR(7)dlog	SETAR(6)dlog
	SETAR(3)dlog	SETAR(3)dlog	SETAR(4)dlog	SETAR(5)dlog	SETAR(4)dlog	SETAR(6)dlog	SETAR(7)dlog	SETAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog
JAGUAR	AR12	AR12	AR12	VECongo1112	VECongo1112	AR12	SETAR(1)log	VECongo1112	VARongo1112	VARongo12	VECongo1112	AR12
	SETAR(3)log	VECongo1112	SETAR(3)log	AR12	AR12	VECongo1112	VECongo1112	SETAR(1)log	VARongo12	AR12	AR12	VECongo1112

Table 19 (*continued*)

Car brand	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**	**Step 7**	**Step 8**	**Step 9**	**Step 10**	**Step 11**	**Step 12**
	AR124	SETAR(1)log	VECongo1112	SETAR(1)log	VARongo1112	VARongo1112	AR12	AR12	AR12	VARongo1112	VARongo1112	AR124
KIA	LSTAR(1)log	AAR(9)log	AAR(9)log	AAR(9)log	AAR(9)log	SETAR(8)log	SETAR(8)log	SETAR(8)log	SETAR(8)log	SETAR(6)log	SETAR(6)log	SETAR(6)log
	SETAR(3)log	SETAR(3)log	AAR(8)log	SETAR(6)log	SETAR(5)log	SETAR(7)log	SETAR(7)log	SETAR(6)log	SETAR(6)log	SETAR(8)log	LSTAR(1)log	LSTAR(1)log
	AAR(4)log	LSTAR(1)log	SETAR(1)log	SETAR(5)log	SETAR(8)log	LSTAR(1)log	SETAR(6)log	LSTAR(1)log	LSTAR(1)log	LSTAR(1)log	SETAR(8)log	SETAR(8)log
MITSUBISHI	VECongo1112	VECongo1112	VARongo1112	VARongo1112	VARongo1112	VECongo1112	VECongo1112	VECMNOGO4	VECongo1112	VECongo1112	VECongo1112	VECongo1112
	VARongo1112	VARongo1112	VECongo1112	VECongo1112	VECongo1112	VARongo1112	VECMNOGO4	VARongo12	VECMNOGO4	VECMNOGO4	VECMNOGO4	SETAR(1)log
	VECongo12	VARongo12	SETAR(1)log	VARongo12	VARongo12	VARongo12	VARongo1112	VECongo1112	VARongo1112	VARongo1112	VARongo12	VECMNOGO4
SUZUKI	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	SETAR(2)log
	SETAR(3)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	AAR(1)log
	AAR(3)log	SETAR(3)log	AAR(2)log	SETAR(3)log	AAR(2)log	SETAR(3)log	AAR(2)log	AAR(3)log	SETAR(3)log	SETAR(3)log	SETAR(3)log	SETAR(3)log

Table 20
Models included in the Model Confidence Set for each forecast horizon (from 1 step up to 24 steps ahead) and for each car brand.

BMW	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	49	72	65	75	71	71	73	72	73	73	63	57	37	37	37	36	8	3	7	3	6	6	8	6
Google models	6	13	12	13	13	13	13	12	12	12	9	6	1	4	4	4	2	1	2	1	2	1	1	1
Nonlinear models	36	42	42	43	43	42	42	43	43	43	41	38	31	29	29	27	4	1	3	1	2	2	4	2
Random Walk included?	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No
OPEL	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	16	4	8	8	2	2	1	4	5	1	2	10	31	40	31	31	40	34	39	31	18	13	7	6
Google models	1	3	5	6	2	2	1	3	3	0	1	2	4	4	3	3	4	4	5	3	2	2	1	1
Nonlinear models	12	0	2	0	0	0	0	0	0	0	0	5	22	31	24	24	30	25	27	22	15	10	5	4
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	No	No	Yes	Yes	Yes	Yes	No	No	No	No
VOLKSWAGEN	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	7	4	1	1	3	4	6	1	3	3	4	6	4	5	4	4	4	4	4	5	4	4	4	4
Google models	2	1	0	0	1	1	2	0	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1
Nonlinear models	1	0	0	0	0	0	1	0	1	1	2	3	1	2	1	1	1	1	1	1	1	1	1	1
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
CITROEN	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	1	1	1	80	81	4	10	12	13	14	12	8	13	12	13	21	15	34	27	8	10	8	7	6
Google models	0	0	0	14	14	0	0	1	1	2	1	1	2	2	3	3	2	6	3	1	1	1	1	0
Nonlinear models	0	0	0	46	47	3	7	9	10	10	9	5	9	8	8	15	12	23	20	7	9	7	6	6
Random Walk included?	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
FIAT	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	14	16	32	34	40	42	45	35	36	25	16	1	1	1	1	1	2	2	2	2	2	2	2	2
Google models	6	2	2	3	4	4	5	4	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Nonlinear models	4	12	25	25	29	30	31	24	26	18	10	1	1	1	1	1	1	1	1	1	1	1	1	1
Random Walk included?	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No
TOYOTA	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	3	2	1	2	2	5	5	8	9	13	23	22	25	34	30	36	21	30	22	31	23	22	22	7
Google models	0	0	0	1	1	2	2	2	2	2	2	2	3	3	2	5	2	3	2	2	2	2	2	2
Nonlinear models	2	1	0	0	0	1	1	3	5	9	17	16	19	24	21	25	16	22	16	23	18	17	18	5
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No
JAGUAR	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	65	69	64	74	70	76	71	72	76	74	75	59	44	46	35	33	25	32	33	27	29	29	32	23
Google models	6	7	8	9	8	10	8	8	9	10	10	7	9	7	7	6	4	5	6	5	5	4	5	4
Nonlinear models	49	50	46	53	53	54	53	53	53	51	53	43	27	34	24	25	18	21	24	20	22	23	22	15
Random Walk included?	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No
KIA	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	86	86	85	85	84	85	85	85	85	85	64	52	28	24	79	69	36	24	23	17	13	10	12	10
Google models	14	14	14	14	14	14	14	14	14	14	10	9	2	3	13	9	6	2	2	2	2	1	1	0
Nonlinear models	52	52	51	51	51	51	51	51	51	51	46	37	21	19	50	47	24	20	20	14	10	9	11	10
Random Walk included?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	No	No	No	No	No	No	No
MITSUBISHI	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	98	98	98	98	98	98	98	98	98	97	82	65	39	56	40	34	24	6	6	13	12	6	14	17
Google models	14	14	14	14	14	14	14	14	14	14	11	10	6	8	5	6	5	3	3	5	4	3	5	5
Nonlinear models	64	64	64	64	64	64	64	64	64	63	55	44	27	39	28	22	15	1	1	5	6	1	6	8
Random Walk included?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No
SUZUKI	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	11	18	16	4	2	13	27	23	20	16	12	8	10	8	8	7	7	8	7	7	7	7	6	5
Google models	4	4	3	0	0	2	4	4	3	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Nonlinear models	2	11	11	3	1	10	20	15	13	13	10	8	10	8	8	7	7	8	7	7	7	7	6	5
Random Walk included?	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No

Table 21
Top three models in terms of MSPE for each forecasting horizon and each car brand.

Car brand	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**	**Step 7**	**Step 8**	**Step 9**	**Step 10**	**Step 11**	**Step 12**
BMW	BVARPNOGO4	BVARPD	VADongo1112	VADongo1112	VADongo1112	VADongo1112	VADongo1112	VADongo1112	VADongo1112	VADongo1112	AR12	AR124
	BVARP	BVARP	BVARPDNOGO	BVARPDNOGO	AR12	BVARPDNOGO	BVARPD	BVARPD	BVARPDNOGO	AR12	AR124	AR12
	BVARPNOGO	BVARPDNOGO	BVARPDNOGO4	BVARPD	BVARPDNOGO	AR12	BVARPDNOGO4	BVARPDNOGO4	AR12	AR124	VADongo1112	VADongo1112
OPEL	BVARPDNOGO	BVARPNOGO	VARongo1112	VARP	VARP	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112
	BVARPNOGO	VEPongo1112	VEPongo1112	VARXP	VARXP	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VADongo1112	VADongo1112
	BVARPDNOGO4	BVARPDNOGO	BVARPD	VEPongo1112	VEPongo1112	VARP	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	VARongo1112	RW
VOLKSWAGEN	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO
	VECMXPNOGO	BVARP	VARPNOGO	VARPNOGO	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	BVARP	SETAR(7)log	SETAR(4)log	SETAR(6)log
	VECMPNOGO	BVARPDNOGO	VARXPNOGO	VARXPNOGO	VARPNOGO	VADongo1112	VARPNOGO	BVARP	VEPongo1112	SETAR(3)log	SETAR(3)log	SETAR(2)log
CITROEN	VARPNOGO4	VARongo1112	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	VARongo1112	VARongo1112
	VARXPNOGO4	BVARPNOGO	BVARPNOGO4	VARongo1112	BVARPNOGO4	BVARPNOGO4	BVARPNOGO4	AAR(3)log	AAR(3)log	VARongo1112	AAR(3)log	VARongo12
	BVARPNOGO4	BVARPNOGO4	VARongo1112	BVARPNOGO4	VARPNOGO	VARongo1112	VARPNOGO	BVARPNOGO4	BVARPNOGO4	AAR(3)log	BVARPNOGO	AAR(3)log
FIAT	BVARP	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	AAR(6)log	AAR(6)log	AAR(6)log	SETAR(1)log
	BVARPDNOGO4	VEPongo1112	VADongo1112	VADongo1112	LSTAR(2)log	LSTAR(2)log	SETAR(1)log	AAR(6)log	AAR(5)log	AAR(5)log	SETAR(1)log	SETAR(4)log
	BVARPNOGO	VARPDNOGO	BVARPDNOGO4	LSTAR(2)log	LSTAR(3)log	AAR(1)log	AAR(1)log	SETAR(1)log	AAR(3)log	SETAR(4)log	SETAR(4)log	AAR(1)log
TOYOTA	BVARPNOGO4	BVARPDNOGO	BVARPDNOGO	VARongo1112	VEPongo1112	VECMXPNOGO4	VEPongo1112	BVARP	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112
	BVARPDNOGO	BVARPD	BVARPD	VEPongo1112	BVARP	VEPongo1112	VECMXPNOGO4	VARongo1112	BVARP	BVARP	BVARP	BVARP
	BVARP	BVARPDNOGO4	BVARPDNOGO4	BVARPNOGO4	VARongo1112	BVARP	VARongo1112	VEPongo1112	VADongo12	VARongo1112	VARongo1112	VARongo1112
JAGUAR	BVARPNOGO4	BVARPNOGO4	VARPNOGO	VARPNOGO	VARongo12	BVARPD	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VARPNOGO	BVARPDNOGO4
	BVARP	BVARP	VARongo1112	VARongo12	VARPNOGO	BVARPDNOGO4	VARongo12	VARPNOGO	VARongo12	VARPNOGO	BVARPDNOGO4	BVARPD
	VARongo1112	VEPongo1112	VARPNOGO4	VARPNOGO4	BVARPNOGO4	VADongo1112	VARPNOGO	VARongo12	VARongo1112	VARongo12	BVARPD	BVARPDNOGO
KIA	AAR(4)log	AAR(4)log	AAR(7)log	AR12	AAR(9)log	BVARPD	AAR(7)log	AAR(7)log	AAR(7)log	AAR(9)log	AAR(9)log	AAR(9)log
	AAR(6)log	BVARPDNOGO	BVARPDNOGO	AAR(9)log	AR12	VADongo1112	AAR(9)log	AAR(9)log	AAR(9)log	AAR(7)log	AAR(8)log	AAR(7)log
	AAR(2)log	AAR(6)log	VARongo1112	AAR(7)log	AAR(7)log	BVARPDNOGO	AR12	SETAR(5)log	AR12	AR12	AAR(7)log	VADongo1112
MITSUBISHI	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112
	BVARPDNOGO	AAR(1)log	AAR(1)log	LSTAR(3)log	LSTAR(2)log	LSTAR(3)log	LSTAR(2)log	LSTAR(2)log	PECM	LSTAR(2)log	LSTAR(2)log	AAR(1)log
	BVARPD	AAR(2)log	BVARPD	LSTAR(4)log	LSTAR(3)log	AAR(2)log	LSTAR(4)log	AAR(1)log	LSTAR(2)log	AAR(1)log	LSTAR(3)log	AAR(2)log
SUZUKI	BVARPDNOGO	BVARPNOGO	SETAR(2)log	AAR(1)log	SETAR(2)log	SETAR(2)log	SETAR(3)log	AAR(2)log	SETAR(2)log	SETAR(3)log	AAR(4)log	AAR(4)log
	BVARPDNOGO4	BVARPDNOGO	AAR(1)log	SETAR(2)log	AAR(1)log	SETAR(3)log	SETAR(2)log	SETAR(3)log	AAR(2)log	AAR(2)log	SETAR(6)log	SETAR(3)log
	BVARPD	BVARPNOGO4	VARongo1112	AAR(3)log	AAR(3)log	AAR(3)log	AAR(2)log	SETAR(2)log	AAR(3)log	SETAR(2)log	AAR(3)log	AAR(3)log
	Step 13	**Step 14**	**Step 15**	**Step 16**	**Step 17**	**Step 18**	**Step 19**	**Step 20**	**Step 21**	**Step 22**	**Step 23**	**Step 24**
BMW	VADongo1112	LSTAR(1)log0	LSTAR(1)log0	LSTAR(1)log0	LSTAR(8)log	LSTAR(9)log	LSTAR(1)log1	LSTAR(9)log	SETAR(7)log	LSTAR(1)log1	LSTAR(1)log1	LSTAR(8)log
	LSTAR(9)log	VADongo1112	LSTAR(9)log	AR12	SETAR(1)log	SETAR(9)log	LSTAR(9)log	SETAR(1)log	SETAR(1)log	SETAR(1)log	LSTAR(7)log	LSTAR(1)log1
	LSTAR(1)log0	LSTAR(9)log	AR12	LSTAR(8)log	LSTAR(1)log1	SETAR(1)log	LSTAR(1)log0	LSTAR(8)log	LSTAR(1)log0	SETAR(6)log	LSTAR(8)log	SETAR(1)log
OPEL	VEPongo1112	VEPongo1112	VEPongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	VADongo1112	VADongo1112	VADongo12	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112
	BVARPD	BVARPDNOGO	VARongo1112	VADongo12	BVARPDNOGO	BVARPDNOGO	BVARPDNOGO	BVARPDNOGO	BVARPDNOGO	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO4
VOLKSWAGEN	BVARPNOGO	BVARPNOGO	BVARPNOGO	VECPongo12	VECPongo12	BVARPNOGO	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VECPongo12
	SETAR(2)log	AR124	AR124	BVARPNOGO	BVARPNOGO	VECPongo12	BVARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	VADongo1112	VEPongo1112
	SETAR(3)log	BVARP	SETAR(3)log	VARongo1112	VARongo1112	VEPongo1112	BVARPDNOGO	BVARP	VADongo1112	VADongo1112	VECPongo12	VADongo1112
CITROEN	VARongo1112	VARongo1112	VARongo1112	VARongo1112	AAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log
	AAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log	VARongo1112	VARongo1112	VARongo1112	SETAR(1)log	SETAR(1)log	SETAR(1)log	SETAR(2)dlog	SETAR(2)dlog
	BVARPNOGO	BVARPNOGO	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	SETAR(1)log	VARongo1112	SETAR(2)dlog	SETAR(2)dlog	LSTAR(2)dlog	LSTAR(2)dlog
FIAT	SETAR(1)log	LSTAR(4)log	LSTAR(2)log	LSTAR(2)log	LSTAR(2)log	LSTAR(2)log	AAR(3)log	AAR(6)log	AAR(5)log	LSTAR(1)log1	VARongo1112	LSTAR(5)log
	LSTAR(4)log	SETAR(1)log	LSTAR(4)log	SETAR(3)log	LSTAR(3)log	LSTAR(3)log	LSTAR(2)log	SETAR(2)log	AAR(6)log	SETAR(1)log	LSTAR(5)log	AAR(3)log
	SETAR(4)log	LSTAR(2)log	SETAR(3)log	LSTAR(3)log	SETAR(4)log	AAR(3)log	LSTAR(3)log	AAR(5)log	SETAR(2)log	LSTAR(5)log	AAR(4)log	AAR(4)log
TOYOTA	BVARP	BVARPD	VARongo1112	BVARPD	VARongo1112	VEPongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	BVARPD	BVARPDNOGO4	BVARPD	BVARPDNOGO	BVARPD	VARongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112
	VEPongo1112	BVARPDNOGO	BVARPDNOGO	VARongo1112	BVARPDNOGO	SETAR(8)log	VARPNOGO4	SETAR(3)log	SETAR(6)log	SETAR(3)log	SETAR(3)log	VADongo12
JAGUAR	BVARPDNOGO4	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO
	VEPongo1112	VARPNOGO4	VARPNOGO4	VARPNOGO4	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	AAR(4)log	LSTAR(1)log0	VARPNOGO4	VARPNOGO4
	BVARPD	VEPongo1112	VEPongo1112	VEPongo1112	VARPNOGO4	VARPNOGO4	VARPNOGO4	LSTAR(1)log0	AAR(3)log	AAR(3)log	LSTAR(1)log0	VEPongo1112
KIA	AAR(9)log	AAR(9)log	AAR(9)log	AAR(9)log	AAR(9)log	AAR(9)log	AAR(9)log	AAR(9)log	AAR(9)log	AAR(9)log	AAR(9)log	AAR(9)log
	SETAR(5)log	AAR(8)log	AAR(7)log	AAR(7)log	AAR(7)log	AAR(7)log	AAR(7)log	AAR(7)log	AAR(7)log	AAR(7)log	AR124	AAR(7)log
	AAR(6)log	SETAR(5)log	AAR(6)log	SETAR(5)log	SETAR(5)log	AAR(8)log	SETAR(7)log	SETAR(5)log	AR124	AR124	AAR(7)log	AAR(8)log

MITSUBISHI	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VARongo1112	VARongo1112	VEPongo1112	VARongo1112
	AAR(1)log	AAR(1)log	AAR(1)log	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VEPongo1112	VEPongo1112	VARongo1112	VEPongo1112
	AAR(2)log	PECM	LSTAR(2)log	AAR(1)log	AAR(1)log	VECPongo12	VECPongo12	VECPongo12	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log
SUZUKI	AAR(2)log	SETAR(7)log	SETAR(6)log	SETAR(3)log	AAR(4)log	AAR(2)log	SETAR(7)log	SETAR(7)log	SETAR(7)log	AAR(2)log	AAR(2)log	AAR(2)log
	SETAR(4)log	AAR(2)log	AAR(2)log	SETAR(4)log	AAR(2)log	SETAR(6)log	AAR(2)log	SETAR(5)log	AAR(2)log	AAR(1)log	AAR(1)log	AAR(1)log
	SETAR(3)log	SETAR(2)log	AAR(4)log	AAR(4)log	SETAR(6)log	AAR(4)log	AAR(4)log	AAR(2)log	AAR(4)log	SETAR(6)log	SETAR(2)log	SETAR(7)log

Table 22
Models included in the Model Confidence Set for each forecast horizon (from 1 step up to 24 steps ahead) and for each car brand.

BMW	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	19	12	14	18	23	12	18	20	30	14	9	4	11	21	22	14	38	48	57	60	55	57	65	59
Google models	7	4	6	8	10	5	9	9	10	3	3	1	2	4	4	2	7	9	13	13	12	13	13	10
Nonlinear models	2	0	0	0	3	0	0	3	10	2	3	0	5	12	14	9	22	30	33	34	32	32	39	36
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes
OPEL	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	8	10	10	1	1	3	5	8	12	6	6	5	13	18	23	15	16	11	4	5	6	4	4	4
Google models	4	6	6	1	1	3	3	4	5	4	4	3	6	7	9	7	6	6	3	3	3	2	2	2
Nonlinear models	0	0	0	0	0	0	0	0	3	0	0	0	3	6	9	5	6	2	0	0	1	0	0	0
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	Yes
VOLKSWAGEN	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	2	4	2	2	3	4	3	4	5	12	13	16	16	16	21	24	28	20	11	17	15	5	2	2
Google models	0	1	0	0	1	2	1	2	2	3	4	4	4	5	7	8	9	7	6	6	6	4	2	2
Nonlinear models	0	0	0	0	0	0	0	0	1	5	6	7	7	5	7	9	11	7	1	7	4	0	0	0
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No
CITROEN	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	1	4	4	5	3	4	5	5	5	7	8	10	19	20	32	40	42	47	48	40	48	49	50	41
Google models	0	1	1	1	0	1	1	1	1	2	2	3	5	5	6	5	5	5	5	6	7	6	7	3
Nonlinear models	0	0	0	0	0	0	1	1	1	2	3	4	11	12	21	30	32	36	38	29	38	38	38	35
Random Walk included?	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	No
FIAT	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	5	4	11	10	20	23	24	27	29	26	31	24	27	17	11	9	18	25	25	22	27	24	25	24
Google models	2	2	5	2	6	7	7	8	7	4	5	4	5	3	2	0	2	2	2	1	2	2	3	2
Nonlinear models	0	0	2	5	11	14	15	16	19	19	21	17	20	14	9	9	15	18	18	20	21	20	20	20
Random Walk included?	No	No	No	No	No	No	No	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No
TOYOTA	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	12	11	7	5	7	11	19	37	50	13	20	14	10	7	25	10	23	21	3	6	3	13	11	23

Table 22 (*continued*)

BMW	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Google models	5	5	2	3	4	3	5	7	7	4	5	4	5	4	5	4	5	4	2	3	2	3	3	6
Nonlinear models	0	0	0	0	0	4	6	21	34	4	9	5	0	0	11	2	10	10	0	2	0	5	4	11
Random Walk included?	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No	Yes
JAGUAR	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	15	32	70	71	35	12	26	73	74	56	35	26	24	51	60	51	51	54	52	49	55	52	51	50
Google models	7	14	14	15	15	6	10	14	13	15	10	9	10	13	13	12	13	12	14	11	11	12	10	11
Nonlinear models	2	6	41	41	6	0	6	47	47	27	14	5	5	27	33	27	27	29	27	27	31	29	27	27
Random Walk included?	No	No	No	No	No	No	No	No	Yes	No	No	Yes	No	No	Yes	No	No	No	No	No	Yes	No	Yes	Yes
KIA	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	11	32	21	12	29	18	6	20	17	35	44	38	44	43	40	8	12	21	16	8	16	3	2	2
Google models	0	1	4	2	5	5	1	2	1	2	4	4	6	6	6	0	1	5	3	1	3	0	0	0
Nonlinear models	9	23	11	4	17	6	3	14	12	26	32	27	30	29	26	7	9	12	10	6	9	2	1	2
Random Walk included?	No	Yes	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	No	Yes	No	No	No
MITSUBISHI	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	22	21	50	74	80	86	25	66	69	72	82	83	56	63	61	26	21	21	28	43	35	26	40	44
Google models	6	6	7	10	11	12	5	8	9	10	12	12	8	10	9	4	5	6	7	7	6	7	6	6
Nonlinear models	10	10	36	55	57	57	17	53	54	55	56	55	43	46	44	21	16	14	20	34	28	19	33	37
Random Walk included?	No	No	Yes	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No
SUZUKI	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**
Total n. of models selected	34	25	24	21	19	21	13	10	8	17	10	12	20	18	7	15	16	9	12	13	17	13	7	11
Google models	4	3	3	3	2	3	1	0	0	1	1	1	3	2	1	1	1	1	1	1	1	1	0	0
Nonlinear models	22	17	16	13	14	16	10	10	8	14	9	10	14	14	6	13	14	8	11	12	13	12	7	11
Random Walk included?	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No

Table 23
List of Models included in the Model Confidence Set for each car brand. Forecast horizon: 24 steps ahead.

Seasonally adjusted data									
BMW	**Citroen**	**Fiat**	**Jaguar**	**Kia**	**Mitsubishi**	**Opel**	**Suzuki**	**Toyota**	**Volkswagen**
AAR(1)log	AAR(1)log	AAR(1)log	AR12	SETAR(6)log	VECongo1112	SETAR(8)dlog	SETAR(2)log	VECongo1112	VARongo1112
VARongo1112	SETAR(4)log	AR124	VECongo1112	LSTAR(1)log	SETAR(1)log	VECongo1112	AAR(1)log	SETAR(6)dlog	AR124
AR124	SETAR(3)log		AR124	SETAR(8)log	VECMNOGO4	SETAR(7)dlog	SETAR(3)log	SETAR(1)dlog	AR12
AR12	LSTAR(3)log		VARongo1112	SETAR(5)log	VARongo12	SETAR(5)dlog	AAR(3)log	SETAR(7)dlog	AAR(1)log
AAR(4)log	SETAR(5)log		AAR(2)log	SETAR(1)log	VARongo1112	AAR(8)dlog	SETAR(4)log	SETAR(3)dlog	
BVARNOGO	AAR(2)log		SETAR(1)log	SETAR(7)log	LSTAR(1)log0	VECMNOGO		SETAR(4)dlog	
			AAR(6)log	SETAR(1)log	SETAR(1)log				
			AAR(1)log	SETAR(9)log	AR12				
			LSTAR(1)log	SETAR(1)log	VECongo12				
			AAR(7)log	SETAR(3)log	VADongo12				
			AAR(3)log		LSTAR(1)log1				
			SETAR(3)log		LSTAR(1)log2				
			AAR(5)log		AR124				
			AAR(4)log		SETAR(4)log				
			VECongo12		AAR(5)log				
			LSTAR(4)log		AAR(3)dlog				
			LSTAR(5)log		VECMNOGO				
			LSTAR(3)log						
			SETAR(4)log						
			VARongo12						
			SETAR(1)log						
			VECMNOGO						
			VECMNOGO4						

Table 24
List of models included in the Model Confidence Set for each car brand. Forecast horizon: 24 steps ahead.

Raw data									
BMW	**Citroen**	**Fiat**	**Jaguar**	**Kia**	**Mitsubishi**	**Opel**	**Suzuki**	**Toyota**	**Volkswagen**
LSTAR(8)log	AAR(3)log	LSTAR(5)log	VARPNOGO	AAR(9)log	VARongo1112	VARongo1112	AAR(2)log	VARongo1112	VECPongo12
LSTAR(1)log1	SETAR(2)dlog	AAR(3)log	VARPNOGO4	AAR(8)log	VEPongo1112	VEPongo1112	AAR(1)log	VEPongo1112	VEPongo1112
SETAR(1)log	LSTAR(2)dlog	AAR(4)log	VEPongo1112		AAR(1)log	VECMPNOGO	SETAR(7)log	VADongo12	
SETAR(3)log	SETAR(1)log	SETAR(4)log	BVARPDNOGO4		AAR(2)log	RW	SETAR(3)log	SETAR(6)log	
LSTAR(1)log0	LSTAR(1)dlog	VARongo1112	BVARPD		LSTAR(5)log		SETAR(2)log	SETAR(8)log	
SETAR(1)log	SETAR(3)log	LSTAR(4)log	VADongo12		LSTAR(4)log		SETAR(4)log	SETAR(3)log	
LSTAR(7)log	LSTAR(4)dlog	SETAR(1)log	VARXPNOGO		VECPongo12		AAR(3)log	VECMPNOGO	
SETAR(7)log	SETAR(3)dlog	AAR(1)log	BVARPDNOGO		LSTAR(3)log		AAR(4)log	SETAR(5)log	
LSTAR(6)log	SETAR(4)dlog	LSTAR(3)log	VADongo1112		SETAR(5)log		SETAR(8)log	SETAR(4)log	
AR12	VARongo1112	SETAR(5)log	VECPongo12		VARongo12		SETAR(6)log	BVARPD	
SETAR(5)log	LSTAR(3)dlog	AAR(2)log	LSTAR(1)log0		AAR(3)log		AAR(7)dlog	SETAR(9)log	
AAR(1)log	SETAR(6)log	LSTAR(2)log	LSTAR(9)log		SETAR(4)log			BVARPDNOGO4	
SETAR(6)log	SETAR(4)log	AAR(6)log	BVARPNOGO		LSTAR(7)log			VECMXPNOGO	
AR124	LSTAR(4)log	SETAR(1)log	AAR(3)log		LSTAR(2)log			BVARPDNOGO	
VARPNOGO	LSTAR(6)log	LSTAR(1)log	SETAR(9)log		AAR(8)dlog			SETAR(1)log	
LSTAR(1)log	SETAR(7)log	AAR(5)log	SETAR(1)log		SETAR(6)log			RW4	
SETAR(8)log	LSTAR(8)log	SETAR(2)log	RW		AAR(4)log			SETAR(1)dlog	
SETAR(1)log	LSTAR(3)log	SETAR(3)log	AR12		LSTAR(6)log			SETAR(4)dlog	
SETAR(4)log	LSTAR(7)log	LSTAR(1)log1	AAR(4)log		SETAR(2)log			VADongo1112	
AAR(3)log	AAR(4)log	VEPongo1112	LSTAR(1)log		LSTAR(1)log			LSTAR(4)dlog	
VECMPNOGO	SETAR(2)log	VARXPDNOGO	AAR(1)log		SETAR(3)log			BVARP	
VECMPNOGO4	SETAR(8)log	AR12	LSTAR(3)log		LSTAR(1)log0			RW	
LSTAR(4)log	SETAR(1)dlog	SETAR(1)log	SETAR(5)log		SETAR(7)log			SETAR(2)dlog	
AAR(8)log	LSTAR(5)dlog	LSTAR(6)log	SETAR(3)log		AAR(6)dlog				
LSTAR(2)log	AAR(1)dlog		SETAR(4)log		AAR(5)dlog				
AAR(2)log	VEPongo1112		SETAR(6)log		AAR(7)dlog				
LSTAR(5)log	SETAR(5)dlog		LSTAR(4)log		LSTAR(6)dlog				
LSTAR(3)log	BVARPNOGO		LSTAR(5)log		LSTAR(1)log1				
VADongo1112	AAR(6)dlog		BVARP		SETAR(6)dlog				
AAR(5)log	LSTAR(5)log		LSTAR(6)log		LSTAR(5)dlog				
AAR(4)log	AAR(8)dlog		AAR(2)log		PECM				
LSTAR(9)log	LSTAR(2)log		SETAR(7)log		SETAR(5)dlog				
SETAR(2)log	BVARPNOGO4		SETAR(2)log		BVARP				
AAR(6)log	VARongo12		SETAR(8)log		AAR(3)dlog				
BVARPNOGO	AAR(2)dlog		LSTAR(7)log		AAR(4)dlog				
VARPNOGO4	AAR(9)dlog		AR124		SETAR(1)log				
BVARPNOGO4	AAR(7)dlog		LSTAR(2)log		SETAR(1)log				
AAR(7)log	LSTAR(8)dlog		LSTAR(8)log		AAR(2)dlog				
SETAR(9)log	LSTAR(11)dlog		AAR(2)dlog		SETAR(4)dlog				
SETAR(5)dlog	LSTAR(9)dlog		BVARPNOGO4		LSTAR(4)dlog				

Table 24 (*continued*)

Raw data									
BMW	**Citroen**	**Fiat**	**Jaguar**	**Kia**	**Mitsubishi**	**Opel**	**Suzuki**	**Toyota**	**Volkswagen**
VEPongo1112	VARXPDNOGO4		VECMPNOGO4		LSTAR(3)dlog				
BVARP			VECMPNOGO		LSTAR(2)dlog				
BVARPD			SETAR(8)dlog		BVARPNOGO4				
BVARPDNOGO			SETAR(5)dlog		SETAR(3)dlog				
BVARPDNOGO4			VARP						
AAR(9)log			VARongo12						
PECM			VECMP						
VECMP			VARPD						
VADongo12			AAR(8)log						
RW			VARXPD						
RW4									
AAR(5)dlog									
AAR(7)dlog									
VARPD									
SETAR(4)dlog									
SETAR(6)dlog									
VECMXP									
VECMXPNOGO4									
VARXP									

Table 25
Top three models in terms of MSPE for each forecasting horizon and each car brand.

Car brand	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**	**Step 7**	**Step 8**
BMW	VARD	VARXDNOGO4	SETAR(1)log	VADongo12	VARXNOGO4	VAR	VECongo12	VARX
	SETAR(5)log	VARDNOGO4	SETAR(8)log	VECongo12	VADongo12	VADongo12	SETAR(3)dlog	VADongo1112
	VARDNOGO4	VARD	VARXNOGO4	SETAR(1)log	VAR	VECongo1112	VARongo12	VAR
OPEL	VARDNOGO4	VADongo12	VADongo12	VECongo12	AAR(1)log	AAR(1)log	VECM	AAR(1)log
	VARXDNOGO4	VARongo12	VECongo12	VECongo1112	VECongo1112	VECM	VECMX	VARNOGO4
	LSTAR(3)dlog	VECongo12	VARongo12	VADongo12	VECM	VECMX	AAR(1)log	VARXNOGO4
VOLKSWAGEN	AAR(1)dlog	BVARD	VARDNOGO	BVARD	VARDNOGO	VARD	VARD	VARNOGO4
	VADongo1112	VARDNOGO	VARXDNOGO	VARDNOGO	VARXDNOGO	VARXD	VARXD	VARXNOGO4
	VARDNOGO	VARXDNOGO	VADongo12	VARXDNOGO	VADongo1112	VARNOGO4	VARNOGO4	VAR
CITROEN	VARDNOGO4	VECMXNOGO4	VARDNOGO4	BVARD	VARDNOGO4	VARXDNOGO4	VADongo12	VARDNOGO4
	VARXDNOGO4	VECMNOGO4	VARXDNOGO4	VARDNOGO4	VARXDNOGO4	VARDNOGO4	VARDNOGO4	VARXDNOGO4
	VARD	BVARDNOGO4	VECMXNOGO4	VARXDNOGO4	BVARD	VADongo12	VARXDNOGO4	VADongo12
FIAT	VARongo1112	AAR(8)dlog	VADongo12	VARDNOGO4	VARDNOGO4	VARD	VARD	AAR(9)dlog
	VADongo1112	VADongo1112	AAR(8)dlog	VARXDNOGO4	VARXDNOGO4	VARXD	VARXD	VARNOGO4
	VECongo1112	LSTAR(6)log	VECongo12	VARD	AAR(9)dlog	VADongo12	VARNOGO4	VARXNOGO4
TOYOTA	SETAR(2)log	AAR(1)log	AAR(9)log	AAR(9)log	AAR(9)log	VARongo1112	VARongo1112	VARongo1112
	SETAR(4)log	VECMXNOGO	AAR(1)log	AAR(1)log	VARongo1112	AAR(9)log	SETAR(1)log	AAR(9)log
	SETAR(8)log	VECMNOGO	VECMXNOGO	AAR(7)log	VARNOGO	VARNOGO	SETAR(2)log	SETAR(1)log
JAGUAR	AAR(8)log	SETAR(8)dlog	VARongo12	AAR(3)dlog	AAR(8)log	VARongo12	LSTAR(9)dlog	SETAR(3)dlog
	SETAR(1)log	LSTAR(8)dlog	VARDNOGO4	AAR(8)log	AAR(6)log	SETAR(9)dlog	SETAR(9)dlog	LSTAR(2)dlog
	SETAR(9)dlog	SETAR(9)dlog	AAR(8)log	AAR(7)log	SETAR(9)dlog	AAR(3)dlog	LSTAR(1)dlog2	SETAR(2)dlog
KIA	BVARDNOGO	VECMXNOGO4	VECMXNOGO4	BVARNOGO4	BVARDNOGO	BVARDNOGO	AAR(7)log	SETAR(1)log
	VARongo1112	VARXNOGO4	VECMNOGO4	LSTAR(8)dlog	AAR(7)log	AAR(7)log	LSTAR(8)dlog	AAR(3)dlog
	BVARDNOGO4	AAR(7)log	BVARNOGO4	AAR(2)dlog	BVARDNOGO4	SETAR(10)dlog	LSTAR(10)dlog	SETAR(4)log
MITSUBISHI	AAR(9)dlog	AAR(9)dlog	AAR(9)dlog	AAR(9)dlog	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log
	LSTAR(5)log	AAR(4)dlog	LSTAR(1)dlog2	VECMNOGO	VECMNOGO	AAR(7)dlog	SETAR(2)log	LSTAR(7)log
	SETAR(5)log	AAR(8)dlog	AAR(8)log	AAR(8)log	LSTAR(2)log	LSTAR(2)log	LSTAR(7)log	LSTAR(4)log
SUZUKI	VADongo12	SETAR(10)dlog	VADongo12	VECongo12	VECongo12	AAR(5)dlog	AAR(3)log	VADongo1112
	VECongo12	VECM	VECongo12	VADongo12	AAR(5)dlog	VARongo12	AAR(3)dlog	LSTAR(5)dlog
	VECM	VECMX	LSTAR(9)dlog	SETAR(10)dlog	VARDNOGO	AR12	AAR(5)dlog	VARD

Table 26
Top three models in terms of MSPE for each forecasting horizon and each car brand.

Car brand	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**	**Step 7**	**Step 8**
BMW	AAR(6)log	BVAR	SETAR(6)dlog	AR12	AR12	AAR(7)dlog	AAR(6)dlog	AR12
	AAR(7)log	AAR(6)log	AR12	SETAR(6)dlog	SETAR(3)log	AAR(6)dlog	SETAR(3)dlog	SETAR(1)log
	BVARNOGO4	BVARNOGO4	AAR(6)log	SETAR(1)log	SETAR(1)log	AAR(5)dlog	AAR(5)dlog	SETAR(3)dlog
OPEL	LSTAR(3)dlog	VECMXNOGO	VARNOGO4	VARNOGO4	VECM	VECongo1112	VECongo1112	VECongo1112
	LSTAR(2)dlog	VECMNOGO	VARXNOGO4	VARXNOGO4	VECMX	VECMXNOGO	VECMXNOGO	VARongo1112
	LSTAR(4)dlog	SETAR(5)log	VECMXNOGO	VECongo1112	VARNOGO4	VECMNOGO	VECMNOGO	VECMXNOGO
VOLKSWAGEN	AR12	AR124	AR124	AR124	AR124	AR124	VARongo1112	AR124
	AR124	AR12	AR12	AAR(9)log	AR12	VARongo1112	AR124	VARongo1112
	AAR(1)log	VECMXNOGO	VECMXNOGO	AR12	VARongo1112	AR12	AR12	AR12
CITROEN	BVARNOGO	VAR	VARNOGO4	VARXNOGO4	VARNOGO4	VARNOGO	LSTAR(3)log	LSTAR(3)log
	AAR(2)log	VARX	VARXNOGO4	VARNOGO4	VARXNOGO4	VARXNOGO	SETAR(4)log	SETAR(4)log
	SETAR(2)log	VARNOGO	VAR	VARX	VARXNOGO	VAR	SETAR(5)log	SETAR(5)log
FIAT	AAR(1)log	AAR(3)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log
	AAR(3)log	AAR(1)log	AAR(3)log	SETAR(4)log	SETAR(2)log	VARongo1112	LSTAR(2)log	SETAR(1)log
	AR124	AAR(4)log	AAR(5)log	BVARNOGO	LSTAR(2)log	SETAR(2)log	VARongo1112	SETAR(2)log
TOYOTA	VARNOGO	VARNOGO	VARNOGO	VECongo1112	VECongo1112	VECongo1112	BVAR	BVAR
	VARXNOGO	VARXNOGO	VARXNOGO	VARNOGO	BVAR	BVAR	VECongo1112	BVARNOGO4
	BVARNOGO4	BVARD	VARNOGO4	VARXNOGO	BVARNOGO4	BVARNOGO4	BVARNOGO4	VECongo1112
JAGUAR	AAR(4)log	SETAR(3)dlog	AAR(3)log	AAR(3)log	SETAR(5)log	SETAR(5)log	VECongo1112	SETAR(3)log
	AAR(5)log	LSTAR(3)dlog	AAR(4)log	LSTAR(3)log	AAR(4)log	LSTAR(4)log	SETAR(5)log	SETAR(4)log
	AAR(3)log	AAR(3)log	SETAR(2)dlog	AAR(4)log	AAR(3)log	SETAR(4)log	AAR(4)log	SETAR(5)log
KIA	VARDNOGO	LSTAR(1)log	VARDNOGO4	AAR(4)log	AAR(4)log	AAR(4)log	SETAR(3)log	AAR(4)log
	LSTAR(1)log	VARDNOGO4	AAR(4)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(4)log	LSTAR(1)log
	SETAR(9)log	AR12	AAR(7)log	AAR(3)log	AAR(6)log	SETAR(3)log	AAR(2)log	AAR(3)log
MITSUBISHI	BVARDNOGO	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112
	BVARD	LSTAR(1)log	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	BVARDNOGO4	SETAR(1)log	AAR(2)dlog	LSTAR(1)log	SETAR(1)log	SETAR(1)log	SETAR(1)log	VARongo12
SUZUKI	RW	AAR(2)log	SETAR(3)log	SETAR(3)log	SETAR(3)log	SETAR(3)log	SETAR(2)log	SETAR(3)log
	RW4	SETAR(2)log	AAR(3)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(3)log	AAR(3)log
	BVARDNOGO4	BVARDNOGO	AAR(2)log	AAR(4)log	SETAR(1)log	SETAR(1)log	AAR(1)log	SETAR(2)log

Table 27
Top three models in terms of average directional accuracy (in %) for each car brand, for short-term forecasts (1–6 steps ahead), medium-term forecasts (7–12 steps ahead), and long-term forecasts (13–24 steps ahead).

Car brand	**Steps 1–6**	**%**	**Steps 7–12**	**%**	**Steps 13–24**	**%**
Seasonally adjusted data						
BMW	AR124	68.21	AR124	66.18	AAR(1)log	80.51
	SETAR(3)log	66.70	AAR(3)log	65.88	AR124	80.28
	SETAR(8)log	66.20	LSTAR(1)log	65.38	VARongo1112	77.58
OPEL	VADongo12	68.95	VADongo12	69.02	SETAR(11)dlog	77.55
	SETAR(11)dlog	66.51	SETAR(11)dlog	68.13	LSTAR(8)dlog	75.06
	VECM	65.15	LSTAR(8)dlog	67.88	BVARNOGO	74.12
VOLKSWAGEN	AR124	74.53	AR124	82.88	VARongo1112	84.59
	VARongo1112	73.79	VARongo1112	80.79	AR124	84.25
	AR12	72.28	AR12	80.66	AR12	82.25
CITROEN	VECMXNOGO4	65.59	VECongo1112	67.69	VECongo1112	75.12
	VECMNOGO4	65.59	VAR	67.31	AAR(4)dlog	66.61
	VECM	65.15	VARX	67.31	AAR(5)dlog	61.16
FIAT	VECongo1112	61.02	VECongo1112	74.54	AAR(1)dlog	85.16
	VADongo12	60.41	VECongo12	67.31	AAR(2)dlog	83.39
	VARongo1112	59.99	LSTAR(12)dlog	64.59	AAR(3)dlog	83.21
TOYOTA	VECongo1112	66.56	VECongo12	71.59	BVAR	74.53
	VARongo1112	65.27	VADongo12	71.57	BVARNOGO4	72.44
	VECMXNOGO	64.97	BVARNOGO4	71.06	VADongo12	71.03
JAGUAR	LSTAR(3)log	68.15	SETAR(4)log	66.42	VARongo1112	70.79
	AAR(4)log	68.13	LSTAR(3)log	66.17	AR12	70.10
	LSTAR(4)log	68.12	LSTAR(4)log	65.63	VARongo12	69.98
KIA	AAR(7)log	67.52	AAR(2)log	71.68	SETAR(5)log	78.77
	AAR(4)log	66.06	AAR(3)log	71.40	SETAR(6)log	78.77
	VARXNOGO4	65.80	AAR(6)log	71.11	AAR(2)log	78.46
MITSUBISHI	VECongo1112	69.07	VECongo12	81.34	VECongo1112	84.41
	VECongo12	67.31	VARongo12	77.16	VECongo12	82.43
	AAR(4)dlog	65.86	VECongo1112	77.16	VARongo12	79.91
SUZUKI	AAR(1)log	71.10	SETAR(3)log	75.20	AAR(1)log	84.56
	SETAR(3)log	68.56	SETAR(2)log	74.67	SETAR(2)log	79.60

Table 27 (*continued*)

Car brand	**Steps 1–6**	**%**	**Steps 7–12**	**%**	**Steps 13–24**	**%**
	SETAR(2)log	68.08	AAR(1)log	73.82	SETAR(3)log	77.50
Raw data						
BMW	VADongo1112	84.45	VADongo1112	82.27	LSTAR(10)log	81.52
	VARongo12	84.38	BVARPD	79.96	LSTAR(11)log	80.57
	BVARPD	83.89	LSTAR(11)log	79.80	LSTAR(9)log	80.44
OPEL	VADongo1112	87.02	VEPongo1112	81.69	VARongo1112	81.99
	VARongo1112	86.41	VADongo1112	81.13	VADongo1112	77.82
	VADongo12	85.48	VADongo12	79.77	VEPongo1112	77.76
VOLKSWAGEN	BVARPNOGO	82.85	VEPongo1112	78.75	VECPongo12	84.64
	VECMP	82.61	BVARPNOGO	78.42	VEPongo1112	81.13
	VECMXP	82.61	VECMP	76.48	VARongo1112	77.99
CITROEN	VARPNOGO4	80.54	VEPongo1112	76.72	AAR(3)log	78.02
	VARXPNOGO4	80.54	BVARPNOGO4	76.52	AAR(4)log	77.40
	VARPNOGO	80.29	BVARPNOGO	76.49	LSTAR(3)dlog	76.97
FIAT	VARongo1112	85.77	PECM	84.58	VARongo1112	83.02
	VADongo1112	84.65	VADongo12	79.70	VEPongo1112	80.62
	BVARP	84.37	VADongo1112	78.25	VADongo12	79.48
TOYOTA	VARongo1112	81.73	BVARP	79.23	VEPongo1112	79.42
	VEPongo1112	79.64	VADongo12	78.56	VARongo1112	79.03
	BVARPNOGO4	77.02	VECPongo12	78.52	BVARP	75.00
JAGUAR	VARXPNOGO4	80.60	VARXPNOGO4	79.57	VARPNOGO	80.45
	VARXPNOGO	80.05	VARPNOGO	79.12	VARPNOGO4	77.87
	AR12	80.04	VARPNOGO4	77.58	AR12	77.32
KIA	BVARPDNOGO	72.52	VARPNOGO4	67.93	VARPNOGO4	70.62
	VARongo1112	71.97	VARXPNOGO4	67.93	VARXPNOGO4	70.62
	VADongo1112	71.96	VEPongo1112	66.56	RW	69.21
MITSUBISHI	VEPongo1112	75.45	VEPongo1112	77.87	VEPongo1112	82.60
	VARongo1112	74.56	VECPongo12	76.31	VECPongo12	82.07
	VECPongo12	72.98	LSTAR(2)log	73.21	VARongo1112	80.77
SUZUKI	VARongo1112	75.88	SETAR(6)log	77.95	AAR(2)log	78.95
	BVARPNOGO	73.27	SETAR(5)log	77.55	SETAR(8)log	78.90
	AR12	72.68	AAR(4)log	76.77	VARongo1112	78.75

Table 28
Average ratio – averaged across all forecasting horizons– of the MSE for the forecasts computed with GIs downloaded on the 15/08/2014, with respect to MSE for the forecasts computed with the average GIs downloaded between 15/08/2014 and 02/09/2014. NC=not converged.

Model	**BMW**	**Opel**	**Volkswagen**	**Citroen**	**Fiat**	**Toyota**	**Jaguar**	**Kia**	**Mitsubishi**	**Suzuki**
Seasonally adjusted data										
VECM	1.08	3.65	0.94	1.00	0.65	NC	1.03	0.95	1.06	1.95
VECMX	1.07	3.65	0.94	1.00	0.65	NC	0.85	NC	0.77	1.95
VAR	0.92	0.66	1.00	1.00	1.15	0.96	1.00	0.98	0.95	1.18
VARD	1.14	1.61	0.79	0.22	1.49	1.16	0.94	0.92	0.99	1.26
VARX	0.97	0.66	1.00	1.00	1.07	0.86	1.42	0.97	0.95	1.18
VARXD	1.07	1.61	0.79	1.00	1.41	1.05	1.00	1.28	1.00	1.26
VECongo1112	1.02	0.94	1.03	0.99	0.99	1.13	1.02	1.05	0.95	1.05
VECongo12	1.00	0.99	1.04	1.05	1.48	0.41	0.87	1.14	1.01	1.10
VARongo1112	1.00	1.02	0.98	0.99	1.03	1.04	0.98	1.04	1.03	0.89
VARongo12	1.01	0.97	1.04	1.09	1.01	1.19	0.83	1.05	1.03	0.91
VADongo1112	1.02	0.97	1.00	1.00	0.99	1.03	0.96	2.15	1.00	0.98
VADongo12	1.08	0.95	1.04	0.95	0.97	1.02	1.13	3.58	0.93	1.19
BVAR	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.97	1.00	1.00
BVARD	1.00	0.99	1.00	1.00	1.00	1.00	1.00	2.06	0.99	0.99
Raw data										
VECMP	0.61	0.99	1.54	1.20	0.25	NC	1.03	NC	1.57	0.58
VECMXP	1.16	0.99	1.54	1.20	0.25	NC	0.81	NC	1.10	0.58
VEPongo1112	1.00	0.99	0.97	0.98	0.97	0.83	1.04	1.00	0.96	1.04
VECPongo12	0.96	0.93	0.92	1.59	1.27	0.90	1.04	1.07	1.06	2.73
VARP	0.96	0.97	0.98	1.07	0.96	0.99	0.97	1.00	1.08	0.95
VARPD	1.06	1.00	0.93	1.01	1.17	0.57	1.00	0.80	1.12	1.22
VARXP	1.12	0.97	0.98	1.07	0.96	0.88	0.98	1.00	0.97	0.95
VARXPD	1.09	1.00	0.93	1.01	1.17	1.26	1.00	0.80	1.25	1.22
VADongo1112	1.02	1.00	0.96	1.00	0.96	1.49	1.00	1.00	1.00	1.05
VADongo12	1.04	0.91	1.04	1.21	0.98	1.10	0.98	1.03	1.00	3.69
VARongo1112	0.94	0.96	0.96	0.99	0.96	0.88	1.00	0.98	0.99	1.04
VARongo12	1.02	0.93	1.08	1.47	1.00	1.25	0.97	1.02	0.97	1.20
BVARP	1.00	1.00	1.00	1.00	0.99	0.95	1.00	1.00	1.00	1.00
BVARPD	1.00	1.00	1.00	1.00	0.98	0.99	1.00	1.00	1.00	1.00
PECM	0.97	1.46	1.70	1.13	1.00	3.10	1.04	0.63	1.10	1.25

Table 29
Top three models in terms of MSPE for each forecasting horizon and each car brand.

Car brand	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8	Step 9	Step 10	Step 11	Step 12
AUDI	BVARNOGO	AAR(3)log	BVARNOGO	VARXNOGO	LSTAR(6)log	AAR(5)log	AAR(5)log	AAR(5)log	AAR(5)log	SETAR(3)log	SETAR(3)log	SETAR(2)log
	BVARNOGO4	BVARNOGO	AAR(3)log	BVARNOGO	LSTAR(7)log	LSTAR(7)log	SETAR(1)log	SETAR(6)log	SETAR(1)log	SETAR(2)log	SETAR(2)log	AAR(7)log
	BVAR	VARXNOGO	BVAR	BVAR	SETAR(2)log	SETAR(2)log	SETAR(6)log	SETAR(5)log	LSTAR(4)log	AAR(5)log	AAR(6)log	LSTAR(4)log
FORD	BVARDNOGO	BVARDNOGO	BVARDNOGO	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	VARongo1112	VARongo1112
	BVARDNOGO4	BVARDNOGO4	BVARD	AAR(3)log	AAR(5)log	AAR(5)log	AAR(5)log	VARongo1112	AAR(5)log	VARongo1112	AAR(2)log	AAR(2)log
	BVARD	BVARD	BVARDNOGO4	AAR(5)log	AAR(4)log	AAR(4)log	VARongo1112	AAR(5)log	VARongo1112	AAR(5)log	AAR(5)log	AAR(5)log
HYUNDAI	VARXDNOGO	VARDNOGO	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo12	VADongo12	VECongo1112
	LSTAR(1)dlog	VARXDNOGO	VARDNOGO	VARDNOGO	AR12	AR12	AR12	AR12	VADongo12	VADongo12	VECongo12	VADongo12
	VARDNOGO	VARDNOGO4	LSTAR(1)dlog	AAR(2)log	VADongo1112	VADongo1112	AAR(1)log	VARDNOGO	AR12	VECongo1112	VECongo1112	VECongo12
MAZDA	SETAR(1)log	AAR(6)dlog	AAR(6)dlog	SETAR(6)dlog	VECongo12	VECongo12	VECongo12	VECongo12	VECongo12	VECongo12	VECongo12	VECongo12
	SETAR(2)log	AAR(5)dlog	SETAR(6)dlog	LSTAR(6)dlog	VADongo12	VADongo12	VADongo12	VADongo12	VADongo12	VADongo12	VADongo12	VADongo12
	RW	VARongo1112	LSTAR(6)dlog	AAR(6)dlog	SETAR(5)dlog	SETAR(5)dlog	SETAR(6)dlog	SETAR(6)dlog	SETAR(6)dlog	LSTAR(1)dlog2	LSTAR(1)dlog2	LSTAR(1)dlog2
NISSAN	LSTAR(1)log	VADongo1112	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log
	AAR(2)log	AAR(3)log	AAR(3)log	LSTAR(1)log	LSTAR(1)log	AAR(1)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log
	AAR(3)log	AAR(2)log	LSTAR(1)log	AAR(3)log	AAR(3)log	LSTAR(1)log	AAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log
PEUGEOUT	BVARD	BVARD	BVARD	BVARNOGO4	BVARNOGO4	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112
	BVARDNOGO4	BVARDNOGO	BVARDNOGO	BVAR	BVAR	BVAR	AAR(3)log	AAR(3)log	VADongo1112	AAR(5)dlog	AAR(5)dlog	SETAR(3)dlog
	VADongo1112	BVARDNOGO4	BVARDNOGO4	BVARNOGO	BVARNOGO	BVARNOGO4	AAR(2)log	VADongo1112	SETAR(3)dlog	VADongo1112	SETAR(3)dlog	LSTAR(8)dlog
RENAULT	LSTAR(1)log	LSTAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	VECMXNOGO4	VARXNOGO	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log
	SETAR(1)log	SETAR(2)log	VARNOGO	VARNOGO	VARXNOGO	VARXNOGO	VECMXNOGO4	VARXNOGO	SETAR(8)log	SETAR(8)log	LSTAR(6)log	LSTAR(6)log
	SETAR(2)log	VARDNOGO	VARDNOGO	VARXNOGO	VARNOGO	AAR(1)log	AAR(1)log	BVARNOGO	LSTAR(6)log	LSTAR(6)log	SETAR(7)log	SETAR(7)log
HONDA	BVARNOGO	VECMXNOGO	VECongo1112	BVARNOGO	BVARNOGO	BVAR	BVAR	SETAR(9)log	SETAR(9)log	SETAR(9)log	VECongo1112	VECongo1112
	BVARNOGO4	VECongo1112	BVARNOGO	BVARNOGO4	BVARNOGO4	BVARNOGO4	BVARNOGO	VECMXNOGO	SETAR(8)log	VECMXNOGO	VECMXNOGO	VECMXNOGO
	VARXNOGO	BVARNOGO	VARongo1112	VECongo1112	BVAR	BVARNOGO	BVARNOGO4	SETAR(8)log	VECMXNOGO	VECongo1112	SETAR(9)log	SETAR(9)log
LAND ROVER	VARDNOGO	VARDNOGO	VARDNOGO	VARDNOGO	VARDNOGO	VARDNOGO	AAR(3)dlog	SETAR(3)dlog	SETAR(3)dlog	AAR(2)dlog	AAR(2)dlog	AAR(1)dlog
	VARDNOGO4	VARDNOGO4	AAR(1)dlog	AAR(1)dlog	SETAR(3)dlog	AAR(3)dlog	SETAR(3)dlog	AAR(1)dlog	LSTAR(1)dlog	SETAR(1)dlog	AAR(1)dlog	SETAR(6)dlog
	VARXDNOGO	VARD	SETAR(3)dlog	SETAR(3)dlog	AAR(1)dlog	AAR(2)dlog	VARDNOGO	VARDNOGO	AAR(1)dlog	AAR(1)dlog	AAR(3)dlog	SETAR(3)dlog
PORSCHE	LSTAR(1)dlog	AAR(4)dlog	LSTAR(1)dlog	SETAR(6)log	SETAR(6)log	SETAR(6)log	SETAR(6)log	SETAR(6)log	AAR(5)dlog	SETAR(6)log	SETAR(6)log	SETAR(7)dlog
	AAR(4)dlog	AAR(5)dlog	SETAR(2)dlog	SETAR(7)log	LSTAR(3)dlog	SETAR(7)log	SETAR(7)log	AAR(4)dlog	AAR(4)dlog	AAR(4)dlog	AAR(4)dlog	AAR(4)dlog
	AAR(3)dlog	AAR(3)dlog	AAR(3)dlog	LSTAR(3)dlog	SETAR(7)log	AAR(4)dlog	AAR(4)dlog	AAR(6)dlog	SETAR(6)log	SETAR(7)dlog	SETAR(7)dlog	SETAR(9)dlog
SUBARU	LSTAR(1)log	LSTAR(1)log	SETAR(1)log	VARNOGO4	VARNOGO4	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log
	SETAR(1)log	SETAR(1)log	LSTAR(1)log	AAR(1)log	AAR(1)log	SETAR(1)log	VARNOGO4	VARNOGO4	VARNOGO4	SETAR(9)log	SETAR(2)log	SETAR(2)log
	AAR(1)log	AAR(3)log	AAR(1)log	LSTAR(1)log	SETAR(1)log	VARNOGO4	SETAR(2)log	SETAR(2)log	SETAR(9)log	VARNOGO	SETAR(7)log	SETAR(7)log
VOLVO	AAR(2)dlog	VECongo1112	SETAR(4)log	VECongo1112	VECongo1112	VECongo1112	VECM	VECM	SETAR(3)log	VECongo1112	VECongo1112	VECongo1112
	AAR(3)dlog	AAR(2)dlog	AAR(2)dlog	VECM	SETAR(6)dlog	SETAR(3)log	VECongo1112	VECongo1112	VECongo1112	VECM	SETAR(3)log	VARXDNOGO
	AR12	SETAR(2)dlog	SETAR(2)dlog	SETAR(2)dlog	SETAR(4)dlog	SETAR(3)dlog	SETAR(3)dlog	SETAR(3)log	VECM	VECMNOGO4	VECM	VARongo1112

Car brand	Step 13	Step 14	Step 15	Step 16	Step 17	Step 18	Step 19	Step 20	Step 21	Step 22	Step 23	Step 24
AUDI	LSTAR(4)log	LSTAR(4)log	AAR(5)log	AAR(6)log	LSTAR(4)log	LSTAR(4)log	LSTAR(4)log	LSTAR(4)log	SETAR(5)log	LSTAR(3)log	LSTAR(3)log	LSTAR(3)log
	LSTAR(3)log	LSTAR(3)log	SETAR(2)log	AAR(7)log	LSTAR(6)log	SETAR(1)log	LSTAR(3)log	LSTAR(3)log	SETAR(2)log	LSTAR(4)log	LSTAR(4)log	LSTAR(4)log
	AAR(3)log	AAR(3)log	LSTAR(4)log	SETAR(2)log	LSTAR(7)log	LSTAR(3)log	SETAR(1)log	SETAR(5)log	LSTAR(4)log	AAR(3)log	LSTAR(2)log	LSTAR(2)log
FORD	AAR(5)log	AAR(4)log	AAR(4)log	AAR(4)log	AAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log	AR124	AR124	AR124	AR124
	VARongo1112	AAR(3)log	AAR(3)log	AAR(3)log	AAR(4)log	AAR(4)log	AR124	AR124	AAR(3)log	AAR(3)log	AAR(3)log	AAR(3)log
	AAR(3)log	AAR(5)log	AAR(5)log	AAR(5)log	AAR(5)log	AR124	AAR(4)log	AAR(4)log	AAR(6)log	AAR(4)log	AAR(6)log	AAR(4)log
HYUNDAI	VADongo12	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112
	VECongo1112	VADongo12	AR12	VADongo12	AR12	VADongo12	AR12	AR12	VADongo12	AR12	AR12	AR12
	AR12	AR12	VADongo12	AR12	VADongo1112	AR12	VADongo12	VADongo12	AR12	LSTAR(2)dlog	LSTAR(1)dlog	VARXDNOGO
MAZDA	VECongo12	SETAR(10)dlog	SETAR(6)dlog	AAR(6)dlog	LSTAR(1)dlog2	VARongo12	VARongo12	LSTAR(1)dlog2	LSTAR(1)dlog2	VECongo12	VECongo12	VECongo12
	AAR(6)dlog	AAR(6)dlog	SETAR(7)dlog	AAR(5)dlog	AAR(6)dlog	LSTAR(1)dlog2	LSTAR(1)dlog2	VECongo12	VECongo12	LSTAR(1)dlog2	LSTAR(1)dlog2	LSTAR(1)dlog2
	VADongo12	AAR(4)dlog	AAR(6)dlog	AAR(4)dlog	AAR(5)dlog	AAR(6)dlog	VECongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo12
NISSAN	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log
	AAR(2)log	AAR(2)log	LSTAR(1)log	LSTAR(1)log	LSTAR(1)log	LSTAR(1)log	LSTAR(1)log	LSTAR(1)log	LSTAR(1)log	LSTAR(1)log	LSTAR(1)log	LSTAR(1)log
	AAR(3)log	AAR(3)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log
PEUGEOUT	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VADongo12	VECongo1112	SETAR(3)dlog	SETAR(3)dlog	SETAR(3)dlog	SETAR(3)dlog	SETAR(3)dlog

	SETAR(3)dlog	SETAR(3)dlog	SETAR(3)dlog	VADongo12	VADongo12	VECongo1112	SETAR(3)dlog	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112
	AAR(5)dlog	SETAR(2)dlog	VADongo12	SETAR(3)dlog	SETAR(3)dlog	SETAR(3)dlog	VADongo12	SETAR(5)dlog	SETAR(5)dlog	SETAR(5)dlog	SETAR(2)dlog	SETAR(5)dlog
RENAULT	LSTAR(6)log	LSTAR(6)log	LSTAR(6)log	LSTAR(6)log	LSTAR(6)log	LSTAR(6)log	LSTAR(6)log	LSTAR(6)log	LSTAR(6)log	LSTAR(6)log	LSTAR(6)log	LSTAR(6)log
	SETAR(7)log	SETAR(7)log	SETAR(7)log	SETAR(7)log	SETAR(7)log	SETAR(7)log	SETAR(7)log	SETAR(7)log	SETAR(7)log	SETAR(7)log	SETAR(7)log	SETAR(7)log
	SETAR(8)log	SETAR(8)log	SETAR(8)log	SETAR(8)log	SETAR(8)log	SETAR(8)log	SETAR(8)log	SETAR(8)log	SETAR(8)log	SETAR(8)log	SETAR(8)log	SETAR(8)log
HONDA	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VARongo12	VARongo12	VARongo12	VARongo12	SETAR(1)log
	SETAR(9)log	VECMXNOGO	SETAR(9)log	SETAR(9)log	SETAR(9)log	SETAR(9)log	VARongo12	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VARongo12
	VARongo12	SETAR(9)log	VECongo12	SETAR(7)log	SETAR(7)log	VECMXNOGO	SETAR(9)log	VECongo12	VECongo12	VECongo12	VECongo12	VECongo1112
LAND ROVER	AAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog	SETAR(1)dlog
	SETAR(1)dlog	AAR(1)dlog	AAR(1)dlog	AAR(7)dlog	AAR(7)dlog	AAR(7)dlog	AAR(7)dlog	AAR(7)dlog	AAR(6)dlog	AAR(7)dlog	AAR(6)dlog	AAR(1)dlog
	SETAR(3)dlog	RW	AAR(7)dlog	AAR(1)dlog	AAR(1)dlog	AAR(1)dlog	AAR(1)dlog	AAR(1)dlog	AAR(7)dlog	SETAR(9)dlog	AAR(1)dlog	AAR(7)dlog
PORSCHE	SETAR(9)dlog	AAR(4)dlog	AAR(4)dlog	AAR(4)dlog	AAR(4)dlog	AAR(4)dlog	AAR(4)dlog	AAR(3)log	SETAR(7)dlog	SETAR(7)dlog	AAR(8)dlog	AAR(8)dlog
	SETAR(7)dlog	SETAR(9)dlog	SETAR(6)dlog	VARX	AAR(2)dlog	AAR(6)dlog	AAR(8)dlog	AAR(8)dlog	AAR(4)log	AAR(8)dlog	SETAR(7)dlog	SETAR(7)dlog
	AAR(4)dlog	SETAR(7)dlog	SETAR(7)dlog	AAR(5)dlog	AAR(6)dlog	AAR(5)dlog	AAR(6)dlog	AAR(4)log	AAR(3)log	AAR(6)dlog	AAR(6)dlog	AAR(4)dlog
SUBARU	AAR(1)log	AAR(1)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	AAR(1)log	AAR(1)log
	SETAR(2)log	SETAR(2)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	SETAR(1)log	AAR(1)log	AAR(1)log	SETAR(1)log	SETAR(4)log
	AR12	SETAR(9)log	SETAR(3)log	LSTAR(2)log	SETAR(4)log	SETAR(3)log	LSTAR(1)log	AAR(1)log	SETAR(3)log	SETAR(4)log	SETAR(2)log	SETAR(3)log
VOLVO	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	VECongo1112	AR12
	SETAR(3)log	LSTAR(1)log	SETAR(1)log	SETAR(1)log	SETAR(1)log	SETAR(1)log	SETAR(1)log	SETAR(1)log	SETAR(1)log	SETAR(1)log	AR12	AAR(1)log
	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	AAR(1)log	AAR(1)log	AAR(1)log	VECongo1112

Table 30
Top three models in terms of MSPE for each forecasting horizon and each car brand.

Car brand	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**	**Step 7**	**Step 8**	**Step 9**	**Step 10**	**Step 11**	**Step 12**
AUDI	VARongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VARPNOGO	VARPNOGO	AR124	VEPongo1112
	VEPongo1112	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VARPNOGO	VEPongo1112	VEPongo1112	AR12	AR12
	VARPNOGO	VARongo1112	VARP	VARXPNOGO	VARXPNOGO	VADongo1112	VARXPNOGO	VADongo1112	AR12	AR12	VEPongo1112	AR124
FORD	VEPongo1112	VARongo1112	VEPongo1112	VARongo1112	VARongo1112	VARongo1112	AAR(5)log	AAR(5)log	SETAR(1)log	LSTAR(1)log1	AAR(6)log	AAR(6)log
	VADongo1112	VEPongo1112	VARongo1112	VARongo12	LSTAR(2)log	AAR(5)log	VARongo1112	SETAR(2)log	SETAR(2)log	SETAR(4)log	LSTAR(2)log	LSTAR(9)log
	VARongo1112	VARPNOGO	VARongo12	VEPongo1112	AAR(5)log	SETAR(1)log	SETAR(4)log	AAR(4)log	LSTAR(2)log	LSTAR(2)log	AAR(5)log	AAR(4)log
HYUNDAI	VEPongo1112	VEPongo1112	SETAR(11)dlog	AR124	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VARongo12	VADongo1112	VADongo1112
	AR12	AR12	AR12	VEPongo1112	VARongo1112	SETAR(11)dlog	SETAR(11)dlog	VADongo1112	VADongo1112	VADongo12	VEPongo1112	RW
	VADongo1112	VADongo1112	VEPongo1112	VARongo1112	VADongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VEPongo1112	VARongo12	VEPongo1112
MAZDA	VARPNOGO	VARongo1112	VARongo1112	VARongo1112	VECPongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo1112	VARongo1112	VARongo1112
	VARXPNOGO	VARPNOGO	VARPNOGO	VARongo12	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo12	VARongo12	VARongo12
	VARongo1112	VARXPNOGO	VARXPNOGO	VECPongo12	VARongo12	VECPongo12	VADongo12	VADongo12	VADongo1112	VADongo1112	VADongo1112	VADongo12
NISSAN	AR12	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AAR(2)log	AR12	AR12	AR12	AR12	VADongo1112
	VARP	AAR(3)log	AAR(3)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(1)log	AAR(2)log	AAR(2)log	AAR(2)log	BVARP	VEPongo1112
	VARPNOGO	AR12	AAR(1)log	AAR(3)log	AAR(3)log	AR12	AR12	AAR(1)log	AAR(1)log	AAR(1)log	VEPongo1112	BVARP
PEUGEOUT	VADongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VADongo1112	VADongo1112	VADongo1112	SETAR(4)dlog	SETAR(4)dlog	SETAR(10)dlog	SETAR(10)dlog	SETAR(10)dlog
	VEPongo1112	VADongo1112	VADongo1112	VADongo1112	VEPongo1112	VEPongo1112	VEPongo1112	SETAR(10)dlog	SETAR(10)dlog	SETAR(4)dlog	SETAR(4)dlog	AAR(1)dlog
	VARPDNOGO4	VARongo1112	VECMXPNOGO4	VECMXPNOGO4	SETAR(10)dlog	SETAR(10)dlog	SETAR(10)dlog	SETAR(9)dlog	SETAR(5)dlog	SETAR(9)dlog	AAR(1)dlog	SETAR(4)dlog
RENAULT	VEPongo1112	VADongo12	VADongo12	SETAR(2)log	SETAR(3)log	SETAR(1)log	LSTAR(1)log	SETAR(3)log	LSTAR(1)log	LSTAR(1)log	LSTAR(1)log	AAR(4)log
	VARPNOGO	VECPongo12	SETAR(1)log	SETAR(3)log	SETAR(1)log	LSTAR(1)log	SETAR(1)log	AAR(4)log	AAR(4)log	SETAR(3)log	SETAR(5)log	LSTAR(1)log
	VARXPNOGO	VECMXPNOGO	VECMXPNOGO	AAR(2)log	LSTAR(1)log	SETAR(3)log	SETAR(3)log	LSTAR(1)log	SETAR(1)log	SETAR(5)log	SETAR(3)log	SETAR(5)log
HONDA	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO	VARongo1112	VARongo12	VARongo12
	VEPongo1112	VECMXPNOGO	VEPongo1112	VECMXPNOGO	VECMXPNOGO	VECMPNOGO	VECMPNOGO	VECMPNOGO	VECMPNOGO	VARongo12	VARongo1112	VARongo1112

Table 30 (*continued*)

Car brand	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8	Step 9	Step 10	Step 11	Step 12
	VARPNOGO	VECMPNOGO	VECMXPNOGO	VECMPNOGO	VECMPNOGO	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VECMXPNOGO	VECPongo12	VECPongo12
LAND ROVER	AAR(1)dlog	AAR(1)dlog	AAR(1)dlog	AAR(3)dlog	VEPongo1112	VEPongo1112	PECM	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	RW
	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	PECM	AAR(2)dlog	VARPNOGO	VARPNOGO	AAR(3)dlog	AAR(3)dlog	AAR(3)dlog	AAR(2)dlog
	LSTAR(1)dlog	RW	AAR(3)dlog	RW	AAR(2)dlog	PECM	VEPongo1112	AAR(3)dlog	VARPNOGO	VADongo12	VADongo12	AAR(3)dlog
PORSCHE	VARongo1112	VARongo1112	LSTAR(11)dlog	LSTAR(11)dlog	VARongo1112	AR12	AR12	AR12	AR12	AR12	LSTAR(11)dlog	VADongo1112
	LSTAR(11)dlog	VEPongo1112	VARongo1112	VARongo1112	AR12	AR124	AR124	LSTAR(11)dlog	LSTAR(11)dlog	LSTAR(11)dlog	AR12	AR12
	AR12	BVARPNOGO	SETAR(11)dlog	SETAR(11)dlog	AR124	VARongo1112	VARongo1112	AR124	AR124	AR124	SETAR(11)dlog	AR124
SUBARU	VADongo1112	VARP	SETAR(1)log	LSTAR(1)log	SETAR(4)log	SETAR(4)log	SETAR(4)log	SETAR(4)log	SETAR(1)log	SETAR(1)log	AAR(9)log	BVARPDNOGO
	VEPongo1112	VARXP	LSTAR(1)log	SETAR(1)log	LSTAR(1)log	SETAR(3)log	SETAR(8)log	LSTAR(1)log	LSTAR(1)log	LSTAR(1)log	SETAR(4)log	AAR(9)log
	VARongo1112	LSTAR(1)log	AAR(3)log	SETAR(3)log	SETAR(1)log	LSTAR(1)log	LSTAR(1)log	SETAR(8)log	SETAR(4)log	SETAR(8)log	LSTAR(1)log	AAR(8)log
VOLVO	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	AAR(7)log
	VARongo1112	VADongo1112	PECM	PECM	VECMPNOGO	VECMPNOGO	PECM	VECMPNOGO4	VECMP	VECMP	AAR(7)log	BVARPNOGO4
	PECM	VARongo1112	VADongo1112	VARongo1112	PECM	PECM	VECMPNOGO	VECMPNOGO	AAR(6)log	AAR(6)log	BVARPNOGO4	AAR(6)log
	Step 13	**Step 14**	**Step 15**	**Step 16**	**Step 17**	**Step 18**	**Step 19**	**Step 20**	**Step 21**	**Step 22**	**Step 23**	**Step 24**
AUDI	VEPongo1112	VECMPNOGO	VECMPNOGO	VECMPNOGO	VECMPNOGO	VECMPNOGO	VECMPNOGO	VECMPNOGO	VECMPNOGO	VECMPNOGO	BVARPNOGO	BVARPNOGO
	VECMPNOGO	VECMPNOGO4	VEPongo1112	VECMP	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VECMPNOGO	VECMPNOGO
	VECMPNOGO4	VEPongo1112	VECMPNOGO4	VEPongo1112	VARPNOGO	VARPNOGO	VARPNOGO	BVARPNOGO	BVARPNOGO	BVARPNOGO	VEPongo1112	VEPongo1112
FORD	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	AAR(6)log	AAR(6)log	LSTAR(1)log1	LSTAR(1)log1	VARongo12	VEPongo1112	VEPongo1112	LSTAR(1)log1	SETAR(3)log	SETAR(2)log	VARongo12	SETAR(1)log
	AAR(5)log	AAR(5)log	AAR(6)log	VEPongo1112	LSTAR(3)log	SETAR(2)log	LSTAR(1)log1	VEPongo1112	LSTAR(9)log	SETAR(3)log	SETAR(2)log	SETAR(3)log
HYUNDAI	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	VEPongo1112	SETAR(11)dlog	SETAR(11)dlog	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VEPongo1112	VARongo12	VECMXPNOGO	VECMXPNOGO
	VADongo1112	SETAR(10)dlog	VEPongo1112	VADongo1112	VADongo1112	VADongo1112	VADongo1112	VADongo1112	VADongo1112	VEPongo1112	VEPongo1112	VEPongo1112
MAZDA	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112
	VARongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo12
	VADongo12	VADongo12	VADongo1112	VECPongo12	VECPongo12	VECPongo12	VADongo12	VEPongo1112	VEPongo1112	VECPongo12	VECPongo12	VECPongo12
NISSAN	BVARP	BVARP	BVARP	LSTAR(1)log	AAR(2)log	AAR(2)log	AAR(2)log	VARPDNOGO	AAR(2)log	PECM	VARPDNOGO	VARPDNOGO4
	VEPongo1112	AAR(2)log	AAR(2)log	AAR(2)log	VARPDNOGO	AAR(1)log	AAR(1)log	AAR(2)log	AAR(1)log	AAR(2)log	AAR(2)log	VARPDNOGO
	VADongo1112	VEPongo1112	AAR(1)log	AAR(1)log	AAR(1)log	VARongo1112	VARPDNOGO	AAR(1)log	PECM	AAR(1)log	AAR(1)log	PECM
PEUGEOUT	SETAR(10)dlog	SETAR(10)dlog	SETAR(10)dlog	SETAR(10)dlog	SETAR(10)dlog	SETAR(4)dlog	SETAR(4)dlog	SETAR(4)dlog	SETAR(4)dlog	SETAR(4)dlog	SETAR(4)dlog	SETAR(4)dlog
	SETAR(4)dlog	AR124	SETAR(11)dlog	SETAR(4)dlog	SETAR(9)dlog	SETAR(9)dlog	SETAR(9)dlog	SETAR(9)dlog	SETAR(5)dlog	SETAR(5)dlog	SETAR(5)dlog	SETAR(9)dlog
	AAR(1)dlog	SETAR(11)dlog	SETAR(4)dlog	SETAR(9)dlog	SETAR(4)dlog	SETAR(10)dlog	SETAR(5)dlog	SETAR(5)dlog	SETAR(9)dlog	SETAR(9)dlog	SETAR(9)dlog	SETAR(7)dlog
RENAULT	SETAR(5)log	LSTAR(1)log	SETAR(6)log	AAR(4)log	SETAR(5)log	SETAR(4)log	SETAR(4)log	SETAR(5)log	SETAR(6)log	SETAR(9)log	SETAR(8)log	SETAR(5)log
	LSTAR(1)log	SETAR(3)log	AAR(4)log	LSTAR(1)log	SETAR(4)log	SETAR(5)log	SETAR(5)log	SETAR(4)log	SETAR(5)log	SETAR(5)log	SETAR(9)log	SETAR(4)log
	SETAR(4)log	AAR(4)log	LSTAR(1)log	SETAR(4)log	AAR(4)log	AAR(4)log	AAR(4)log	SETAR(7)log	SETAR(4)log	SETAR(4)log	SETAR(6)log	LSTAR(1)log
HONDA	VARongo12	VARongo1112	VARongo12	VARongo12	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo1112	VARongo12	VARongo12
	VARongo1112	VARongo12	VARongo1112	VARongo1112	VARongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo12	VARongo1112	VARongo1112
	VECPongo12	VECMXPNOGO	VECPongo12	VECPongo12	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO	VECPongo12
LAND ROVER	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog	AAR(2)dlog
	AAR(3)dlog	RW	RW	RW	RW	RW	RW	RW	RW	RW	RW	RW
	VADongo1112	AAR(3)dlog	AAR(3)dlog	AAR(3)dlog	AAR(3)dlog	AAR(3)dlog	VARXPNOGO4	AAR(3)dlog	AAR(3)dlog	AAR(3)dlog	AAR(3)dlog	AAR(1)dlog
PORSCHE	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO	VADongo12	VADongo12	VADongo12	VECMXPNOGO	AR124	VARPNOGO	VARPNOGO
	VECMPNOGO	VECMPNOGO	VECMPNOGO	VECMPNOGO	VECMPNOGO	VECMXPNOGO	VECMXPNOGO	VECMXPNOGO	VECMPNOGO	AR12	VARXPNOGO	VARXPNOGO
	PECM	VADongo12	VADongo12	BVARPNOGO	VADongo12	VECMPNOGO	VECMPNOGO	VECMPNOGO	AR124	VADongo12	VADongo12	VADongo12
SUBARU	SETAR(8)log	AAR(5)log	SETAR(8)log	AAR(8)log	AAR(8)log	AR12	SETAR(1)log	AAR(8)log	AR12	SETAR(1)log	SETAR(2)log	SETAR(3)log
	LSTAR(1)log	AR12	AAR(5)log	AAR(5)log	LSTAR(11)dlog	AAR(4)log	AR12	AR12	AR124	LSTAR(1)log	SETAR(1)log	SETAR(1)log
	SETAR(1)log	SETAR(8)log	SETAR(7)log	AAR(6)log	AR12	AAR(6)log	SETAR(2)log	SETAR(1)log	SETAR(2)log	SETAR(4)log	SETAR(4)log	LSTAR(1)log
VOLVO	BVARPNOGO4	VEPongo1112	LSTAR(1)log	LSTAR(1)log	LSTAR(1)log	SETAR(1)log	LSTAR(1)log	LSTAR(1)log	SETAR(1)log	LSTAR(1)log	SETAR(2)log	SETAR(2)log
	VEPongo1112	BVARPNOGO4	LSTAR(2)log	LSTAR(2)log	SETAR(1)log	LSTAR(1)log	SETAR(1)log	SETAR(1)log	LSTAR(1)log	SETAR(1)log	SETAR(1)log	AAR(2)log
	AAR(1)log	LSTAR(2)log	SETAR(1)log	SETAR(1)log	LSTAR(2)log	LSTAR(2)log	AAR(2)log	SETAR(2)log	SETAR(2)log	SETAR(2)log	LSTAR(1)log	AAR(1)log

Appendix H. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.ijpe.2015.09.010.

References

Aparicio, F., Escribano, A., Garcia, A., 2006. Range Unit-Root (RUR) tests: robust against nonlinearities, error distributions, structural breaks and outliers. J. Time Ser. Anal. 27 (4), 545–576.

Askitas, N., Zimmermann, K., 2009. Google econometrics and unemployment forecasting. Appl. Econ. Q. 55, 107–120.

Boivin, J., Ng, S., 2006. Are more data always better for factor analysis? J. Econom. 132 (1), 169–194.

Boswijk, H., 1994. Testing for an unstable root in conditional and structural error correction models. J. Econom. 63, 37–60.

Boswijk, H., Franses, P., 1995. Periodic cointegration-representation and inference. Rev. Econ. Stat. 77, 436–454.

Boswijk, H., Franses, P., 1996. Unit roots in periodic autoregressions. J. Time Ser. Anal. 17, 221–245.

Brühl, B., Hülsmann, M., Borscheid, D., Friedrich, C., Reith, D., 2009. A sales forecast model for the German automobile market based on time series analysis and data mining methods. In: Perner, P. (Ed.), Advances in Data Mining Applications and Theoretical Aspects, Springer, Berlin, pp. 146–160.

Carriere-Swallow, Y., Labbé, F., 2013. Nowcasting with Google trends in an emerging market. J. Forecast. 32 (4), 289–298.

Choi, H., Varian, H., 2012. Predicting the present with Google Trends. Econ. Rec. 88, 2–9.

Da, Z., Engelberg, J., Pengjie, G., 2003. A comparative study of linear and nonlinear models for aggregate retail sales forecasting. Int. J. Prod. Econ. 86, 217–231.

Da, Z., Engelberg, J., Pengjie, G., 2011. In Search of attention. J. Finance 5, 1461–1499.

D'Amuri, F., J. Marcucci, 2013. The predictive power of Google Searches in Forecasting Unemployment, Temi di discussione (Economic Working Papers) 891, Bank of Italy.

Dekimpe, M.G., Hanssens, D.M., Silva-Risso, J.M., 1998. Long-run effects of price promotions in scanner markets. J. Econom. 89 (1), 269–291.

Doan, T., Litterman, R., Sims, C., 1984. Forecasting and conditional projection using realistic prior distributions. Econom. Rev. 3, 1–100.

Fantazzini, D., Fomichev, N., 2014. Forecasting the real price of oil using online search data. Int. J. Comput. Econ. Econom. 4 (1–2), 4–31.

Franses, P., Dijk, V., 2000. Nonlinear Time Series Models in Empirical Finance. Cambridge University Press, Cambridge.

Franses, P., Paap, R., 2004. Periodic Time Series Models. Oxford University Press, Oxford.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014.

Gregory, A., Hansen, B., 1996. Residual-based tests for cointegration in models with regime shifts. J. Econom. 70, 99–126.

GTAI, 2014. http://www.gtai.de.

Hall, S., Henry, S., Greenslade, J., 2002. On the identification of cointegrated systems in small samples: a modelling strategy with an application to UK. J. Econ. Dyn. Control 26, 1517–1537.

Hansen, P., Lunde, A., Nason, J., 2011. The model confidence set. Econometrica 79 (2), 453–497.

Hartigan, J.A., 1975. Clustering Algorithms. 99th ed., John Wiley & Sons, Inc., New York, NY, USA.

Hatemi, J., 2008. Tests for cointegration with two unknown regime shifts with an application to financial market integration. Empir. Econ. 35, 497–505.

Hayashi, F., 2000. Econometrics. Princeton University Press, Princeton.

Hulsmann, M., Borscheid, D., Friedrich, C.M., Reith, D., 2012. General sales forecast models for automobile markets and their analysis. Trans. Mach. Learn. Data Min. 5 (2), 65–86.

International Telecommunications Union, 2014. Internet users' graphs, market information and statistics. http://www.itu.int/ITU-D/ict/statistics.

Internet World States, 2014. http://www.internetworldstats.com/stats.htm.

Johansen, S., 1995. Likelihood-based Inference in Cointegrated Vector Autoregressive Models. Oxford University Press, Oxford.

Johansen, S., 2006. Cointegration: a survey. In: Mills, T., Patterson K. (Eds.), Palgrave Handbook of Econometrics: Econometric Theory, vol. 1, Palgrave MacMillan, Basingstoke, UK and New York, USA, pp. 540–577.

Johansen, S., Mosconi, R., Nielsen, B., 2000. Cointegration analysis in the presence of structural breaks in the deterministic trend. Econom. J. 3, 216–249.

Koop, G., Korobilis, D., 2010. Bayesian multivariate time series methods for empirical macroeconomics. Found. Trends Econom. 3 (4), 267–358.

Kunhui, L., Qiang, L., Changle, Z., Junfeng, Y., 2007. Time series prediction based on linear regression and SVR. In: Third International Conference on Natural Computation (ICNC2007), Haikou, China, pp. 688–691.

Landwehr, J.R., Labroo, A.A., Herrmann, A., 2011. Gut liking for the ordinary: incorporating design fluency improves automobile sales forecasts. Market. Sci. 30, 416–429.

Lee, L., Strazicich, M., 2003. Minimum Lagrange multiplier unit root test with two structural breaks. Rev. Econ. Stat. 85 (4), 1082–1089.

Lin, Y., Zhang, W.J., 2004. Aesthetic design for automobile interiors: critical and conceptual framework. In: 2004 IEEE International Conference on Systems, Man and Cybernetics, The Hague, pp. 6313–6317.

Litterman, R., 1986. Forecasting with Bayesian vector autoregressions: five years of experience. J. Bus. Econ. Stat. 4, 25–38.

Lutkepohl, H., 2005. New Introduction to Multiple Time Series Analysis. Springer, Berlin.

Maddala, GS., Kim, I., 1998. Unit Roots, Cointegration, and Structural Change. Cambridge University Press, York.

Naik, G., Leuthold, R., 1986. A note on qualitative forecast evaluation. Am. J. Agric. Econ. 68, 721–726.

Net Applications, 2014. Market share statistics for internet technologies, www.Marketshare.hitslink.com.

Nijs, V.R., Dekimpe, M.G., Steenkamp, J.-B.E.M., Hanssens, D.M., 2001. The category-demand effects of price promotions. Market. Sci. 20, 1–22.

Pauwels, K., Hanssens, D.M., Siddarth, S., 2002. The long-term effects of price promotions on category incidence, brand choice, and purchase quantity. J. Market. Res. 39, 421–439.

Pauwels, K., Silva-Risso, J., Srinivasan, S., Hanssens, D.M., 2004. New products, sales promotions, and firm value: the case of the automobile industry. J. Market. 68, 142–156.

Pesaran, M., Shin, Y., 1998. Impulse response analysis in linear multivariate models. Econ. Lett. 58, 17–29.

Rossi, E., Fantazzini, D., 2016. Long memory and periodicity in intraday volatility. J. Financ. Econ. forthcoming. < http://dx.doi.org/10.1093/jjfinec/nbu006 > .

Sa-ngasoongsong, A., Bukkapatnam, S.T.S., 2011. Willingness-to-pay prediction based on empirical mode decomposition. In: Doolen, T. Aken, E.V. (Eds.), Proceedings of the 2011 Industrial Engineering Research Conference, Reno.

Sa-ngasoongsong, A., Bukkapatnam, S.T.S., Kim, J., Iyer, P.S., Suresh, R.P., 2012. Multi-step sales forecasting in automotive industry based on structural relationship identification. Int. J. Prod. Econ. 140, 875–887.

Shahabuddin, S., 2009. Forecasting automobile sales. Manag. Res. News 32, 670–682.

Stock, J., Watson, M., 1993. A simple estimator of cointegrating vectors in higher order integrated systems. Econometrica 61, 783–820.

Stock, J.H., Watson, M.W., 2006. Forecasting with many predictors. In: Handbook of Economic Forecasting, vol. 1, pp. 515-554.

Suhoy, T., 2009. Query Indices and a 2008 Downturn, Discussion Paper 6, Bank of Israel.

Terasvirta, T., Tjostheim, D., Granger, C., 2011. Modelling Nonlinear Economic Time Series. Oxford University Press, Amsterdam.

Theil, H., 1961. Economic Forecasts and Policy. North Holland Publishing Company, Amsterdam.

Tong, H., 1990. Non-linear Time Series: A Dynamical System Approach. Oxford University Press, Oxford.

VDA, 2014. 〈http://www.vda.de〉.

Wang, F., Chang, K., Tzeng, C., 2011. Using adaptive network-based fuzzy inference system to forecast automobile sales. Exp. Syst. Appl. 38, 10587–10593.

Wood, S., (2006): Generalized additive models: an introduction with R. Chapman and Hall / CRC, Boca Raton.

Zellner, A., 1978. Seasonal analysis of economic time series, no. zell78-1 in NBER Books. US Department of Commerce, Bureau of the Census.

Appendix 13 "The oil price crash in 2014/15: Was there a (negative) financial bubble?"

Fantazzini, D. (2016). The oil price crash in 2014/15: Was there a (negative) financial bubble? *Energy Policy*, 96, 383-396. DOI: https://doi.org/10.1016/j.enpol.2016.06.020

Permisson to post the Publisher's version/PDF: according to https://www.elsevier.com/about/policies/sharing "*Theses and dissertations which contain embedded Published Journal Article (PJA) as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.*"

Energy Policy 96 (2016) 383–396

Contents lists available at ScienceDirect

Energy Policy

journal homepage: www.elsevier.com/locate/enpol

The oil price crash in 2014/15: Was there a (negative) financial bubble?

Dean Fantazzini

Moscow School of Economics, Moscow State University, Russia

HIGHLIGHTS

- There was a negative bubble in oil prices in 2014/15.
- This bubble decreased oil prices beyond the level justified by economic fundamentals.
- Several bubble detection methods confirm this evidence.

ARTICLE INFO

Article history:
Received 6 November 2015
Received in revised form
6 June 2016
Accepted 11 June 2016

JEL classification:
C15
C22
C51
C53
G17
O13
Q47

Keywords:
Oil
Wti
Brent
Generalized sup ADF test
LPPL
Bubble

ABSTRACT

This paper suggests that there was a negative bubble in oil prices in 2014/15, which decreased them beyond the level justified by economic fundamentals. This proposition is corroborated by two sets of bubble detection strategies: the first set consists of tests for financial bubbles, while the second set consists of the log-periodic power law (LPPL) model for negative financial bubbles. Despite the methodological differences between these detection methods, they provided the same outcome: the oil price experienced a statistically significant negative financial bubble in the last months of 2014 and at the beginning of 2015. These results also hold after several robustness checks which consider the effect of conditional heteroskedasticity, model set-ups with additional restrictions, longer data samples, tests with lower frequency data and with an alternative proxy variable to measure the fundamental value of oil.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

The Brent and WTI prices of crude oil fell by 60% between June 2014 and January 2015, marking one of the quickest and largest declines in oil history. This fall in oil prices is large but it is not an unprecedented event: the oil price fell more than 30% in a seven-month sample already five times in the last three decades (1985–1986, 1990–1991, 1997–1998, 2001, 2008). Of these five episodes, the price slide in 1985–86 has some similarities with the fall in 2014/2015, because it followed a period of strong expansion of oil supply from non-OPEC countries and Saudi-Arabia decided to increase production and to stop defending prices. Several factors have been proposed to explain this latest price crash: Arezki and Blanchard (2014) suggested an important contribution of positive oil supply shocks after June 2014. For example, there was a faster than expected recovery of Libyan oil production due to a lull in the local civil war, as it is visible from the EIA estimated historical unplanned OPEC crude oil production outages:

Moreover, Iraq oil production was not affected by the civil war enraging in the west and in the north of the country, as initially feared. The success of US shale oil production (+0.9 million b/d in 2014) and the OPEC decision in November 2014 to maintain its production level of 30 mb/d, signalling a shift in the cartel's policy from oil price targeting to maintaining market share, put additional pressure on oil prices.

Oil demand seems to have played a minor role compared to

E-mail addresses: fantazzini@mse-msu.ru, dean.fantazzini@gmail.com

http://dx.doi.org/10.1016/j.enpol.2016.06.020
0301-4215/© 2016 Elsevier Ltd. All rights reserved.

supply shocks: Arezki and Blanchard (2014) suggested that unexpected lower demand between June and December 2014 could account for only 20–35% of the price decline, while Hamilton (2014) found that only two-fifths of the fall in oil prices was due to weak global demand. Baumeister and Kilian (2016) used the reduced-form representation of the structural oil market model developed in Kilian and Murphy (2014) and argued that, out of a $49 fall in the Brent oil price, $11 of this decline was due to adverse demand shocks in the first half of 2014, $ 16 to (positive) oil supply shocks that occurred prior to July 2014, while the remaining part was due to a "*shock to oil price expectations in July 2014 that lowered the demand for oil inventories and a shock to the demand for oil associated with an unexpectedly weakening economy in December 2014, which lowered the price of oil by an additional $9 and $13, respectively*".

These and other potential factors which could have influenced the oil price decline are discussed in an extensive World Bank policy research note by Baffes et al. (2015). Similarly to previous works, they also found out that supply shocks roughly accounted for twice as much as demand shocks in explaining the fall in oil prices. An alternative explanation is put forward by Tokic (2015) who suggested that the 2014 oil price collapse was partially an irrational over-reaction to the falling Euro versus the dollar. This seems to be consistent with a Bank of International Settlements report (Domanski et al., 2015), which shows that production and consumption alone are not sufficient for a fully satisfactory explanation of the collapse in oil prices. In this regard, Domanski et al. (2015) advanced the idea that "*if financial constraints keep production levels high and result in increased hedging of future production, the addition to oil sales would magnify price declines. In the extreme, a downward-sloping supply response of increased current and future sales of oil could amplify the initial decline in the oil price and force further deleveraging*".

Given this background, we want to propose a potential explanation for the part of the oil price decline which can not be explained using supply and demand alone, particularly in the last months of 2014, as highlighted by Baumeister and Kilian (2016). More specifically, we suggest that there was a negative financial bubble which decreased oil prices beyond the level justified by economic fundamentals. A negative financial bubble is a situation where the increasing pessimism fuelled by short positions lead investors to run away from the market, which spirals downwards in a self-fulfilling process, see Yan et al. (2012) for a discussion.

We employ two approaches to corroborate this proposition: the first approach consists of tests for financial bubbles proposed by Phillips et al. (2016) (hereafter PSY) and Phillips and Shi (2014) (hereafter PS). These tests are based on recursive and rolling right-tailed Augmented Dickey-Fuller unit root test, wherein the null hypothesis is of a unit root and the alternative is of a mildly explosive process. They can identify periods of statistically significant explosive price behavior and date-stamp their occurrence. The second approach consists of the log-periodic power law (LPPL) model for negative financial bubbles developed by Yan et al. (2012). This model adapts the Johansen-Ledoit-Sornette (JLS) model of rational expectation bubbles developed by Sornette et al. (1999); Johansen et al. (1999) and Johansen et al. (2000) to the case of a price fall occurring during a transient negative bubble, which they interpret as an effective random down payment that rational agents accept to pay in the hope of profiting from the expected occurrence of a possible rally. Despite the methodological differences between these bubble detection methods, they provide the same result: the oil price experienced a statistically significant negative financial bubble in the last months of 2014 and at the beginning of 2015. A set of robustness checks finally show that our results also hold with different tests, different model set-ups and alternative datasets.

The paper is organized as follows: the bubble detection methods are presented in Section 2, while the data employed in the empirical analysis are discussed in Section 3. The main results are described in Section 4, while robustness checks are reported in Section 5. Conclusions and policy implications are presented in Section 6.

2. Methods - testing for financial bubbles

We wanted to verify the presence of a negative financial bubble in oil prices at the end of 2014 using a set of tests for financial bubbles. We first employed the test by Phillips, Shi, and Yu (PSY, 2015) which builds on the previous work by Phillips, Wu, and Yu (2011, hereafter PWY) and it is designed to identify periods of statistically significant explosive price behavior. Strictly related to this, we also employed the test by Phillips and Shi (PS, 2014) for detecting a potential bubble implosion and estimating the date of market recovery. We then used the log-periodic power law (LPPL) model by Yan et al. (2012) which is specifically designed for negative financial bubbles. Differently from the approach by PSY and PS, the LPPL model does not require the formation of a bubble as a pre-requisite for a price crash.

2.1. Econometric tests for explosive behavior

The generalized-supremum ADF test (GSADF) proposed by Phillips et al. (2015) builds upon the work by Phillips and Yu (2011) and Phillips et al. (2011). This is a test procedure based on ADF-type regressions using rolling estimation windows of different size, which is able to consistently identify and date-stamp multiple bubble episodes even in small sample sizes. It was recently used by Caspi et al. (2015) to date stamp historical periods of oil price explosivity using a sample of yearly data ranging between 1876 and 2014.

The first step is to consider an ADF regression for a rolling sample, where the starting point is given by the fraction r_1 of the total number of observations, the ending point by the fraction r_2, while the window size by $r_w = r_2 - r_1$. The ADF regression is given by

$$x_t = \mu + \rho x_{t-1} + \sum_{i=1}^{p} \phi_{r_w}^{i} \Delta x_{t-i} + \varepsilon_t \tag{1}$$

where μ, ρ, and $\phi_{r_w}^{i}$ are estimated by OLS, and the null hypothesis is of a unit root $\rho = 1$ vs an alternative of a mildly explosive autoregressive coefficient $\rho > 1$.[1] Then, PSY (2015) proposed a backward sup ADF test where the endpoint is fixed at r_2 and the window size is expanded from an initial fraction r_0 to r_2. The test statistic is then given by:

$$BSADF_{r_2}(r_0) = \sup_{r_1 \in [0, r_2 - r_0]} ADF_{r_1}^{r_2} \tag{2}$$

We remark that the PWY (2011) procedure for bubble identification is a special case of the backward sup ADF test where $r_1 = 0$, so that the sup operation is superfluous.

The generalized sup ADF (GSADF) test is computed by repeatedly performing the BSADF test for each $r_2 \in [r_0, 1]$:

$$GSADF(r_0) = \sup_{r_2 \in [r_0, 1]} BSADF_{r_2}(r_0) \tag{3}$$

PSY (2015, Theorem 1) provides the limiting distribution of (3) under the null of a random walk with asymptotically negligible

[1] A detailed analysis of model specification sensitivity in right-tailed unit root testing for explosive behavior was performed by Phillips et al. (2014).

drift, while critical values are obtained by numerical simulation.

If the null hypothesis of no bubbles is rejected, it is then possible to date-stamp the starting and ending points of one (or more) bubble(s) in a second step. More specifically, the starting point is given by the date -denoted as T_{r_e}- when the sequence of BSADF test statistics crosses the critical value from below, whereas the ending point -denoted as T_{r_f}- when the BSADF sequence crosses the corresponding critical value from above:

$$\hat{r}_e = \inf_{r_2 \in [r_0, 1]} \left\{ r_2 \colon BSADF_{r_2}(r_0) > cv_{r_2}^{\beta_T} \right\} \tag{4}$$

$$\hat{r}_f = \inf_{r_2 \in [\hat{r}_e + \delta \log(T)/T, 1]} \left\{ r_2 \colon BSADF_{r_2}(r_0) < cv_{r_2}^{\beta_T} \right\} \tag{5}$$

where $cv_{r_2}^{\beta_T}$ is the $100(1-\beta_T)\%$ right-sided critical value of the BSADF statistic based on $\lfloor Tr_2 \rfloor$ observations, and $\lfloor \cdot \rfloor$ is the integer function. β_T was set to 5%. δ is a tuning parameter which determines the minimum duration for a bubble: this value is set to 1 in PWY (2011), PSY (2015) and most of previous applied work, thus implying a minimum bubble-duration condition of $\log(T)$ observations (i.e. a sample fraction of $\log(T)/T$). In this regard, Figuerola-Ferretti et al. (2015) reported results for weekly non-ferrous metals prices with different choices of the tuning parameter $\delta = 1, 2, 4$, and they found that while the imposition of larger minimum length criterion eliminates some cases of mildly exploding periods, the main results did not change.

Homm and Breitung (2012) compared several tests for detecting financial bubbles and found that the PWY strategy has higher power than the other procedures in detecting periodically collapsing bubbles and in real time monitoring. However, Phillips et al. (2016) showed that the PSY strategy outperforms the PWY strategy in the presence of multiple bubbles.

Phillips et al. (2015) and Phillips et al. (2016) examined the power of the previous test under alternative hypotheses where bubbles collapse instantaneously. However, Yiu et al. (2013) and Figuerola-Ferretti et al. (2015) suggest that the PSY procedure might have some efficacy in detecting bubble implosion and market crashes in general. Strictly speaking, the test proposed by PSY (2015) is for explosive behavior, so that a situation of upward explosive behavior can be interpreted as bubbles, while downward explosive behavior can be interpreted as crashes or panic-selling. In this regard, Phillips and Shi (2014) discussed alternative bubble collapse models where the collapse can be "sudden", "disturbing" or "smooth", and they proposed a reverse sample use of the PSY test procedure for detecting crises and estimating the date of market recovery. More specifically, they propose to use the BSDF test to data x_t^* arranged in reverse order to the original series x_t, so that $x_t^* = x_{T+1-t}$, for $t = 1, 2, \ldots T$. The BSDF statistic for detecting a bubble implosion/market crash is then defined as $BSDF_g^*(g_0)$, where the recursion (in reverse direction) initiates with a minimum window size g_0, and the test is repeatedly computed for each fraction $g \in [g_0, 1]$ of X_t^*. The market recovery date (f_r) and the crisis origination date (f_c), both expressed in fractions of the original series sequence, are then computed as follows:

$$\hat{f}_r = 1 - \hat{g}_e, \quad \text{where} \quad \hat{g}_e = \inf_{g \in [g_0, 1]} \left\{ g \colon BSADF_g^*(g_0) > cv_g^{\beta_T} \right\} \tag{6}$$

$$\hat{f}_c = 1 - \hat{g}_f, \quad \text{where} \quad \hat{g}_f = \inf_{g \in [\hat{g}_e, 1]} \left\{ g \colon BSADF_g^*(g_0) < cv_g^{\beta_T} \right\} \tag{7}$$

where $cv_g^{\beta_T}$ is the $100(1-\beta_T)\%$ critical value of the $BSADF_g^*(g_0)$ statistic. Phillips and Shi (2014) pointed out that the slowly varying function $\delta \log(T)/T$ in this case is not needed, given the interest in identifying abrupt market crashes. Similarly to Eq. (3), a generalized sup ADF test ($GSADF^*(g_0)$) can be computed by repeatedly performing the $BSADF_g^*(g_0)$ test for each $g \in [g_0, 1]$. We will also use this second test to verify the presence of a downward market bubble in the oil price in 2014/2015.

2.2. Log-periodic power law (LPPL) models for negative financial bubble detection

PSY (2015) and PS (2014) consider a model where asset prices follow a random walk during normal periods, a mildly explosive process during the bubble period, and then a bubble implosion which can be abrupt -as in PSY (2015)- or modelled by a stationary integrated process -as in PS (2014)-. Even though the PSY procedure is formally to test for explosive behavior, which can be positive or negative, PSY (2015) focus only on upward trending bubbles. Moreover, the model by PS(2014) requires the formation of a bubble as a pre-requisite for the following price crash. Therefore, we employed also the log-periodic power law (LPPL) model by Yan et al. (2012) which is specifically designed for negative financial bubbles and does not require the formation of a bubble as a pre-requisite for a price crash. This model is an extension of the LPPL model proposed by Sornette et al. (1999); Johansen et al. (1999) and Johansen et al. (2000), which posits the presence of two types of agents in the market (traders with rational expectations and irrational "noise" traders with herding behavior), and assumes that they are organized into networks and can have only two states, buy or sell. Moreover, their trading behavior is influenced by the decisions of other traders and by external shocks. A bubble can then emerge when traders form groups with self-similar behavior, which is regarded as a situation of "order", differently from the "disorder" which takes place during normal market conditions, see Geraskin and Fantazzini (2013) for a recent extensive review and Sornette (2003) for a discussion at the textbook level. Several ex-ante forecasts of bubble episodes were discussed by Zhou and Sornette (2003, 2006, 2008 and 2009); Sornette and Zhou (2006); Sornette et al. (2009).

The expected value of the asset log price in a upward trending bubble (before a crash) according to the LPPL equation is given by,

$$E[\ln p(t)] = A + B(t_c - t)^{\beta} + C(t_c - t)^{\beta} \cos[\omega \ln(t_c - t) - \phi] \tag{8}$$

where β quantifies the power law acceleration of prices, ω represents the frequency of the price oscillations during the bubble, t_c is the so-called 'critical time' that corresponds to the end of the bubble, while A, B, C and ϕ are simply units distributions of betas and omegas and do not have any structural information, see Sornette and Johansen (2001); Johansen (2003); Sornette (2003); Geraskin and Fantazzini (2013) and Lin et al. (2014) for more details.

The first major condition for a bubble to occur within the JLS framework is $0 < \beta < 1$, which guarantees that the crash hazard rate accelerates. The second major condition is that the crash rate should be non-negative, as highlighted by Bothmer and Meister (2003), which imposes that

$$b \equiv -B\beta - |C|\sqrt{\beta^2 + \omega^2} \geq 0.$$

Financial bubbles are defined in the LPPL model as transient regimes of faster-than-exponential price growth resulting from positive feedbacks, and these regimes represent "positive bubbles". Positive feedbacks can also occur in a downward price regime with faster-than-exponential downward acceleration: Yan et al. (2012) refer to these regimes as "negative bubbles". In the latter case, the smaller the price, the larger is the decrease of future price. Moreover, the increasing pessimism fuelled by short positions

leads investors to run away from the market which falls down in a self-fulfilling process. The JLS model can be easily modified to accomodate for negative bubbles, requiring only that both the expected excess return and the crash amplitude become negative, see Yan et al. (2012) for details. It is possible to show that Eq. (8) remains the same, with the inequalities $B > 0$, $b < 0$ being the opposite to those corresponding to a positive bubble, while the first major condition $0 < \beta < 1$ does not change.

The estimation of LPPL models can be rather difficult and several algorithms were recently reviewed by Geraskin and Fantazzini (2013). In this regard, Filimonov and Sornette (2013) proposed a stable and robust calibration scheme of the log-periodic power law model by rewriting the formula (8) as follows:

$$E[\ln p(t)] = A + B(t_c - t)^{\beta} + C_1(t_c - t)^{\beta}\cos[\omega \ln(t_c - t)] + C_2(t_c - t)^{\beta}\sin[\omega \ln(t_c - t)] \quad (9)$$

where $C_1 = C\cos\phi$, $C_2 = C\sin\phi$, and which can be derived from (8) by expanding the cosine term. Similarly to Filimonov and Sornette (2013), we estimated (9) with nonlinear least-squares, but differently from them we employed a variant of the multi-stage procedure proposed in Geraskin and Fantazzini (2013) and Fantazzini (2010) to improve the numerical convergence in small-to-medium sized samples, see the Appendix Appendix A for details.

3. Data

3.1. Which oil price to use?

Some studies tried to identify speculative bubbles in the oil market using the standard present-value model for stocks adapted to commodity markets by Pindyck (1992). In this framework, the fundamental value of oil is defined as the sum of discounted oil dividends which are approximated by the convenience yield, see Lammerding et al. (2013); Areal et al. (2013) and Shi and Arora (2012)). Unfortunately, as shown -inter alia- by Figuerola-Ferretti and Gonzalo (2010); Lammerding et al. (2013) and Figuerola-Ferretti et al. (2015), the estimated convenience yield can become negative so that the ratio between the commodity price and the measured convenience yield becomes uninterpretable and cannot be used for testing bubbles, as done for equity prices by PWY (2011). Moreover, in case of daily data, the estimates are rather volatile and has to be smoothed. Given these issues, we preferred to employ the previous tests with nominal oil prices, as done by Gilbert (2010) and Homm and Breitung (2012) and with real oil prices, as done by Caspi et al. (2015) and Phillips and Yu (2011). To compute the daily real oil prices, we built a daily consumer price index (CPI) series using the methodology used by the US and UK governments for the indexation of Treasury Inflation-Protected Securities (TIPS) and of Index-Linked Gilts, respectively.[2] We considered both nominal and real oil prices also due to the current debate about which price series is better suited for analyzing the relationship between the price of oil and the level of economic activity (see the *Macroeconomic Dynamics* special issue on "Oil Price Shocks" published in 2011 for a detailed discussion): we decided to take a neutral stance on this issue and examined both type of prices (Fig. 1).

[2] Both the US and the UK governments calculate the daily CPI using a linear interpolation between the CPI applicable to the first day of the month and the CPI applicable to the first day of the following month. We used a cubic-spline interpolation because of the better mathematical properties. However, the differences with linearly interpolated data were very small and did not change the outcome of the tests.

3.2. Time sample for model estimation

We analyzed the daily nominal and real WTI and Brent oil prices from January 2013 to April 2015. The nominal prices are the spot prices as provided by the US Energy Information Administration (EIA), while the real prices are computed using the US and UK CPIs, using the methodology described in Section 3.1. We chose this time span because we focus on the price crash at the end of 2014. Moreover, Sornette (2003) and Jiang et al. (2010) remarked that a bubble cannot be diagnosed more than 1 year in advance, so that a statistical test for detecting a bubble at the end of 2014 can be computed using data starting from the year 2013 at the latest. Furthermore, a recent literature examined the interaction between market prices and media coverage and suggested that media hype can be a potential source of speculation and financial bubbles, see (among many) Shiller (2000, 2002), Dyck et al. (2003), Case and Shiller (2003), Veldkamp (2006), Bhattacharya et al. (2009). In this regard, Geraskin and Fantazzini (2013) suggested to use the *Search Volume Index* (SVI) by Google Trends to get some insights as to when a potential bubble may have started: this index computes how many searches have been done for a term on Google over time.[3] If a keyword has both a large number of searches and several potential meanings, Google Trends offers the possibility to choose the SVI related to a specific topic, so that unrelated searches are filtered out: we report in Fig. 2 the SVIs for the topics "*West Texas Intermediate*" and "*Brent Crude*".

Fig. 2 shows that a large interest about these oil prices started to build at the beginning of 2014, so that a time sample from January 2013 to April 2015 seems appropriate. We will verify in Section 5.3 whether our results continue to hold with longer samples that start before 2013.

4. Results

4.1. Econometric tests for explosive behavior

Table 1 reports the GSADF and GSADF* statistics with the 95% critical values obtained by Monte Carlo simulation using 1000 replications, with minimum estimation windows $r_0 = g_0 = 0.01 + 1.8/\sqrt{T}$, as suggested by PSY (2015). The start and end dates for weakly explosive behavior as identified using the PSY (2015) procedure, as well as the crisis origination date and market recovery date as identified using the PS (2014) procedure are also reported. The sequences of BSADF and BSADF* statistics (with 95% critical values) for nominal and real oil prices are reported in Figs. 3 and 4, respectively.

The GSADF tests identify a period of explosive behavior in Brent prices between October 2014 and February 2015, whereas between December 2014 and March 2015 in WTI prices. There is also a very short spike of the BSADF statistic in February 2014 associated with a mild increase of the WTI price, but it seems more a computationally anomaly rather than a period of explosive behavior. In this regard, PS (2014) and PSY (2015) warned that the BSADF statistic may exceed its critical value for a small number of observations and give a false signal, so that they suggested to use a minimum length criterion. The imposition of a minimum length requirement of $\log(T) \approx 7$ days does not change the results, but if we consider a tuning parameter $\delta = 4$ (i. e. 1 month), as suggested by Figuerola-Ferretti et al. (2015), the mild exploding period in February 2014 is eliminated. Instead, the GSADF* tests in Table 1 fail to identify significant period of market implosion for all oil

[3] See https://support.google.com/trends for more details. The time span starts from 2004, which is the first year available for this service.

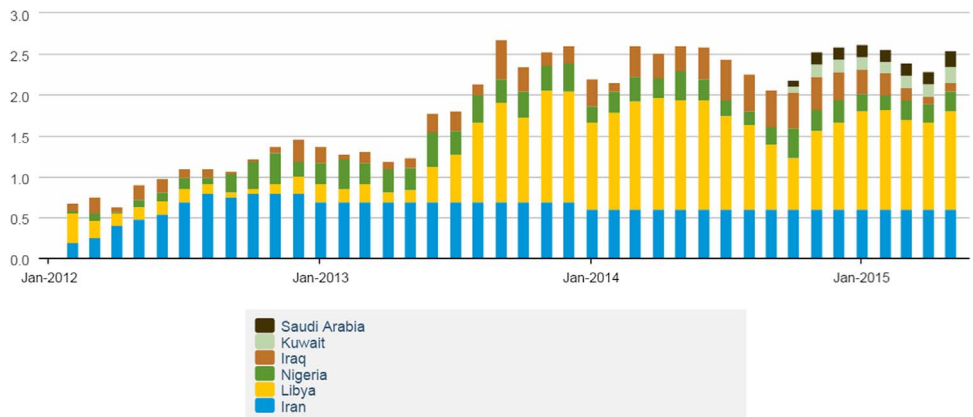


Fig. 1. Estimated Historical Unplanned OPEC Crude Oil Production Outages (million barrels per day).
Source: EIA (2015).

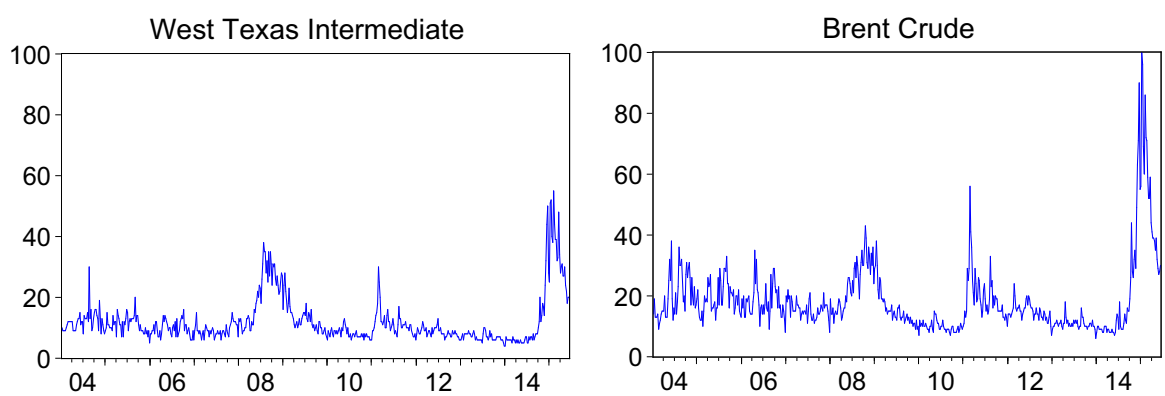


Fig. 2. Google Trends SVIs for "*West Texas Intermediate*" and "*Brent Crude*".

Table 1.
GSADF statistics (upper part) and GSADF* statistics (lower part), 95% critical values, start and end dates for explosive behavior (GSADF test), crisis origination date and market recovery date (GSADF* test). * Significant at the 95% level.

Oil price	GSADF	c.v.	Start (explosive beh.)	End (explosive beh.)
Nominal Brent	4.59*	2.16	02.10.2014	27.02.2015
Nominal WTI	3.22*	2.10	08.12.2014	23.03.2015
Real Brent	4.52*	2.10	02.10.2014	27.02.2015
Real WTI	3.03*	2.16	08.12.2014	23.03.2015
Oil price	**GSADF***	**c.v.**	**Crisis origination date**	**Market recovery date**
Nominal Brent	1.52	2.29	/	/
Nominal WTI	1.92	2.26	/	/
Real Brent	1.51	2.28	/	/
Real WTI	1.97	2.28	/	/

prices considered. In this regard, some insights are given by the BSADF* statistics in the second row of Figs. 3 and 4, which show an erratic behavior and are unable to cross the 95% critical values for sustained periods of time and with high values. These latter results may be due to the relatively short period of time considered for estimation: we will see in Section 5.4 that longer estimation samples make the oil price implosion at the end of 2014 strongly significant.

4.2. LPPL model for negative financial bubble detection

Sornette et al. (2009); Jiang et al. (2010) and Geraskin and Fantazzini (2013) suggested to use estimation samples of varying size to deal with potential parameter instability. Following their example, we fit the logarithm of the examined oil price by using the LPPL Eq. (9) in shrinking windows and in expanding windows. More specifically, for each end date t_2 = 02/01/2014, …, 30/04/2015 the starting date t_1 ranged from $t_2 - 120$ to $t_2 - 250$ in steps of one (trading) day. Following Jiang et al. (2010) and Geraskin and Fantazzini (2013), we then used the set of parameters estimated with all samples $t_2 - j$, $j = 120, \ldots, 250$ to compute the moving 20%/80% and 5%/95% quantile range of the parameters of interest. The 20%/80% and 5%/95% quantile ranges of the LPPL parameters B and β for nominal and real oil prices are reported in Figs. 5 and 6, respectively.

The crash hazard rate b was negative over all time sample (as required by a negative bubble) and therefore was not reported. In general, the LPPL parameters B and β satisfy jointly the conditions for a negative financial bubble between October 2014 and March 2015 for the Brent and between December 2014 and March 2015 for the WTI. However, the evidence for the latter is somewhat weaker. It is interesting to note that despite the methodological differences between the LPPL approach and the econometric tests by PSY (2015), they provide substantially the same result: oil prices experienced a statistically significant negative financial bubble in the last months of 2014 and at the beginning of 2015.

5. Robustness checks

We wanted to verify that our previous results hold also with different tests and alternative datasets. Therefore, we performed the following robustness checks: a) we performed the FTS-GARCH test for financial bubbles by Corsi and Sornette (2014) which takes conditional heteroskedasticity into account; b) we employed the 'volatility-confined' LPPL model by Lin et al. (2014), which is a

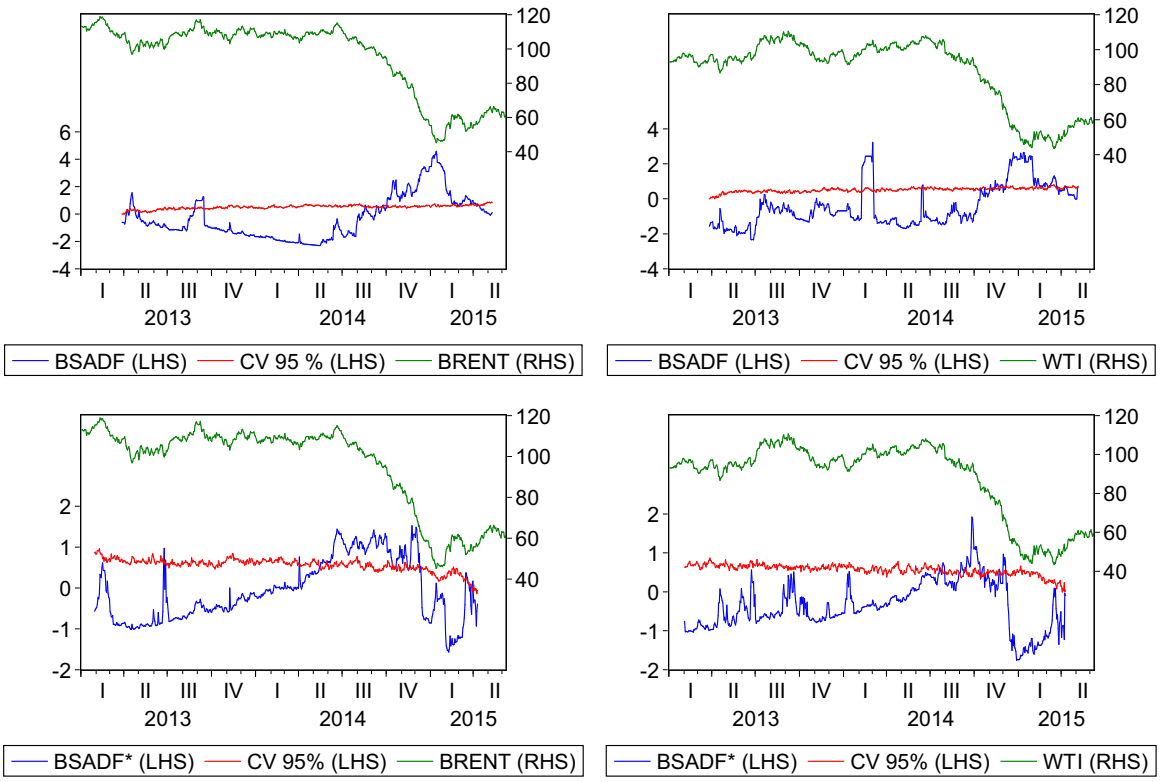


Fig. 3. BSADF statistics (first row) and BSADF* statistics (second row) for date-stamping periods of explosive behavior and market implosions, respectively. *Nominal oil prices*: Brent (first column) and WTI (second column).

generalization of the previous LPPL model; c) we used an alternative longer daily data sample; d) we verified that our results hold also with a weekly dataset. All checks confirmed that the oil price experienced a statistically significant negative financial bubble from the end of 2014 till the beginning of 2015.

5.1. Accounting for heteroskedasticity: The FTS-GARCH test for financial bubbles

Corsi and Sornette (2014) proposed a reduced form model for the joint dynamics of liquidity and asset prices, where the self-reinforcing feedback between credit creation and the market value

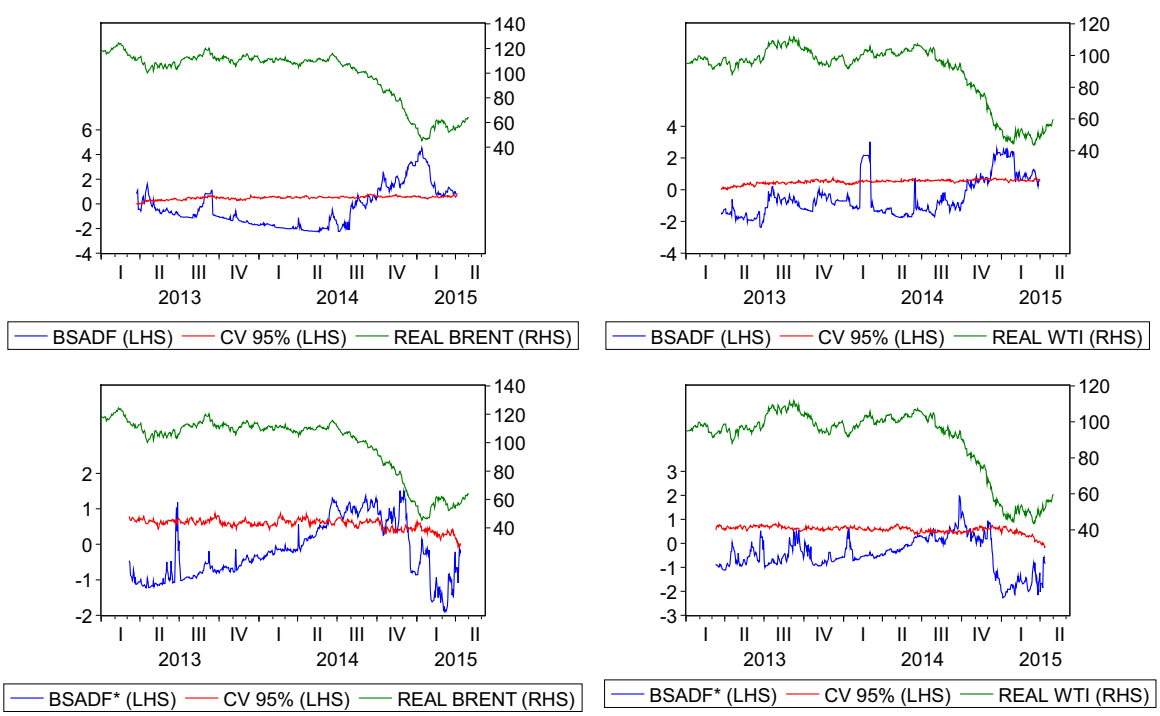


Fig. 4. BSADF statistics (first row) and BSADF* statistics (second row) for date-stamping periods of explosive behavior and market implosions, respectively. *Real oil prices*: Brent (first column) and WTI (second column).

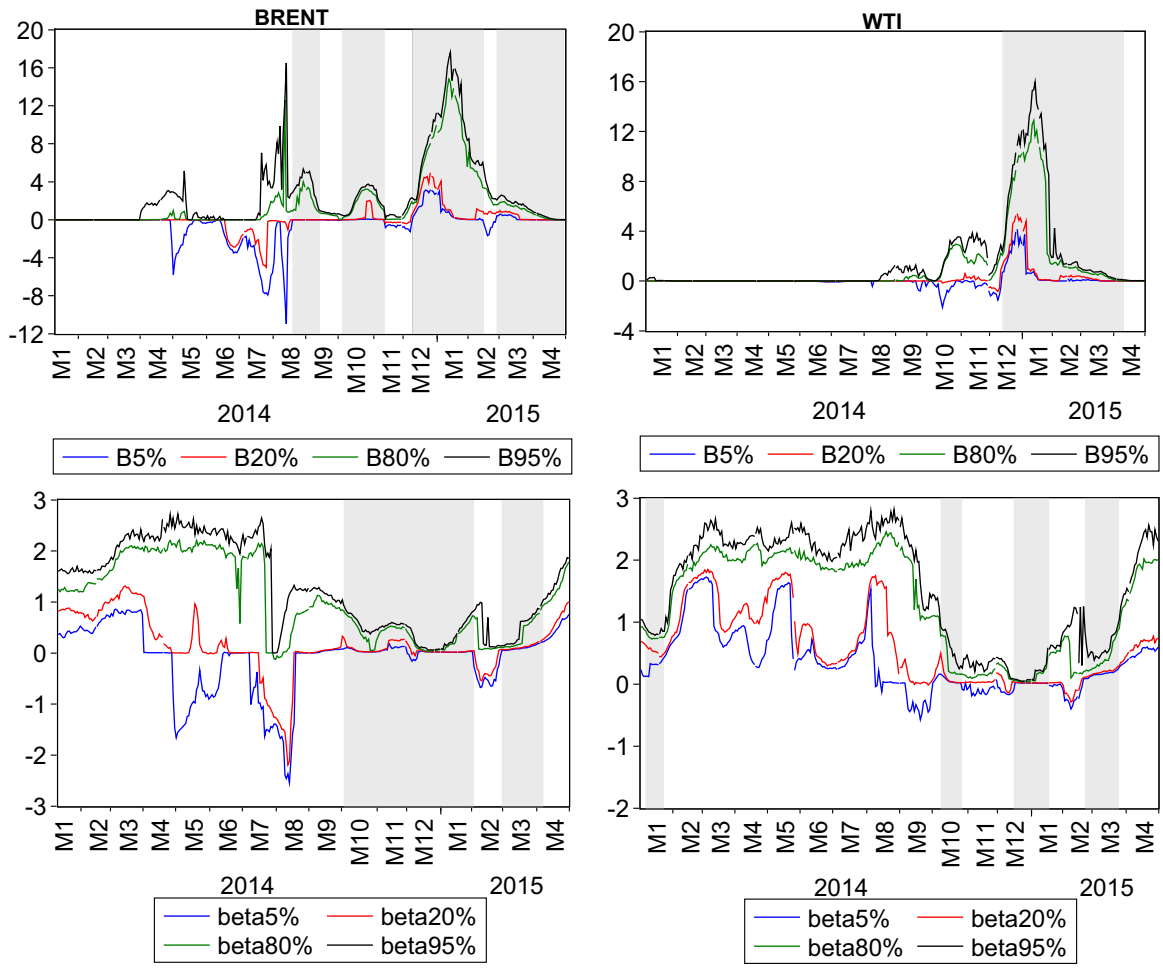


Fig. 5. 20%/80% and 5%/95% quantile range of the LPPL parameters B (first row) and β (second row): shaded areas highlight the time samples when $B > 0$ and $0 < \beta < 1$. *Nominal oil prices*: Brent (first column) and WTI (second column).

of the financial assets employed as collateral in the bank loans (i.e. the financial accelerator) is modelled as a multivariate non-linear stochastic process. They showed that such model can produce explosive dynamics in the financial variables which can lead to a market crash in finite time. Exploiting the implications of their model for asset returns, they proposed an extension of the GARCH process which can provide an early warning identification of financial bubbles. More specifically, they showed that the positive feedbacks of price on money and money on price leads to a finite time singular (FTS) dynamics where these two variables follow a self-reinforcing dynamics of the type $dX/dt \simeq X^{1+\delta}$, with $\delta > 1$, so that the conditional mean of asset log-returns $r_t = \ln(P_t) - \ln(P_{t-1})$ depends on price levels P_t. The resulting FTS-GARCH model proposed by Corsi and Sornette (2014) is given by:

$$r_t = \mu + \gamma P_{t-1} + \varepsilon_t, \quad \varepsilon_t = \sigma_t z_t, \quad z_t \sim N(0, 1) \tag{10}$$

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{11}$$

where ω, α, β are positive parameters and the rejection of the null hypothesis of $\gamma = 0$ is interpreted as evidence of a bubble. This test is a type of right-tailed Dickey-Fuller test with GARCH errors: given the moderate sample size, we employed bootstrap methods to compute the test distribution, following the suggestion by Harvey et al. (2016) who performed a comprehensive analysis of the impact of different volatility structures on the size of the SADF test by PWY (2011). The sequences of t-statistics of the FTS γ parameter (with 95% critical values) for nominal and real oil prices are reported in Fig. 7[4]

The null hypothesis of $\gamma = 0$ is rejected between December 2014 and February 2015 for the Brent, similarly to previous tests, whereas it is not rejected for the WTI. Therefore, the FTS-GARCH approach seems to be more restrictive than the previous tests and the evidence of a potential bubble is confirmed only for Brent oil prices.

5.2. Diagnostic tests based on the LPPL fitting residuals

Lin et al. (2014) proposed a generalization of the LPPL model for financial bubbles where the log-prices fluctuate around the LPPL trajectory and the fitting residuals follow a mean-reverting Ornstein-Uhlenbeck process. The main advantage of the "volatility-confined LPPL model" proposed by Lin et al. (2014) is to guarantee the consistency of direct estimation with prices, which was not possible with the original LPPL model due to the presence of a random walk component with increasing variance.

Lin et al. (2014) used the Phillips-Perron (PP) and the Augmented Dickey-Fuller (ADF) to test the stationarity of the LPPL fitting residuals, whereas we used here the test by Kwiatkowski, Phillips, Schmidt and Shin (Kwiatkowski et al., 1992), where the null hypothesis is a stationary process. We employed the latter test

[4] The sequences start on the 2nd of January 2014, given the need to have a minimum time sample for the estimation of the GARCH models.

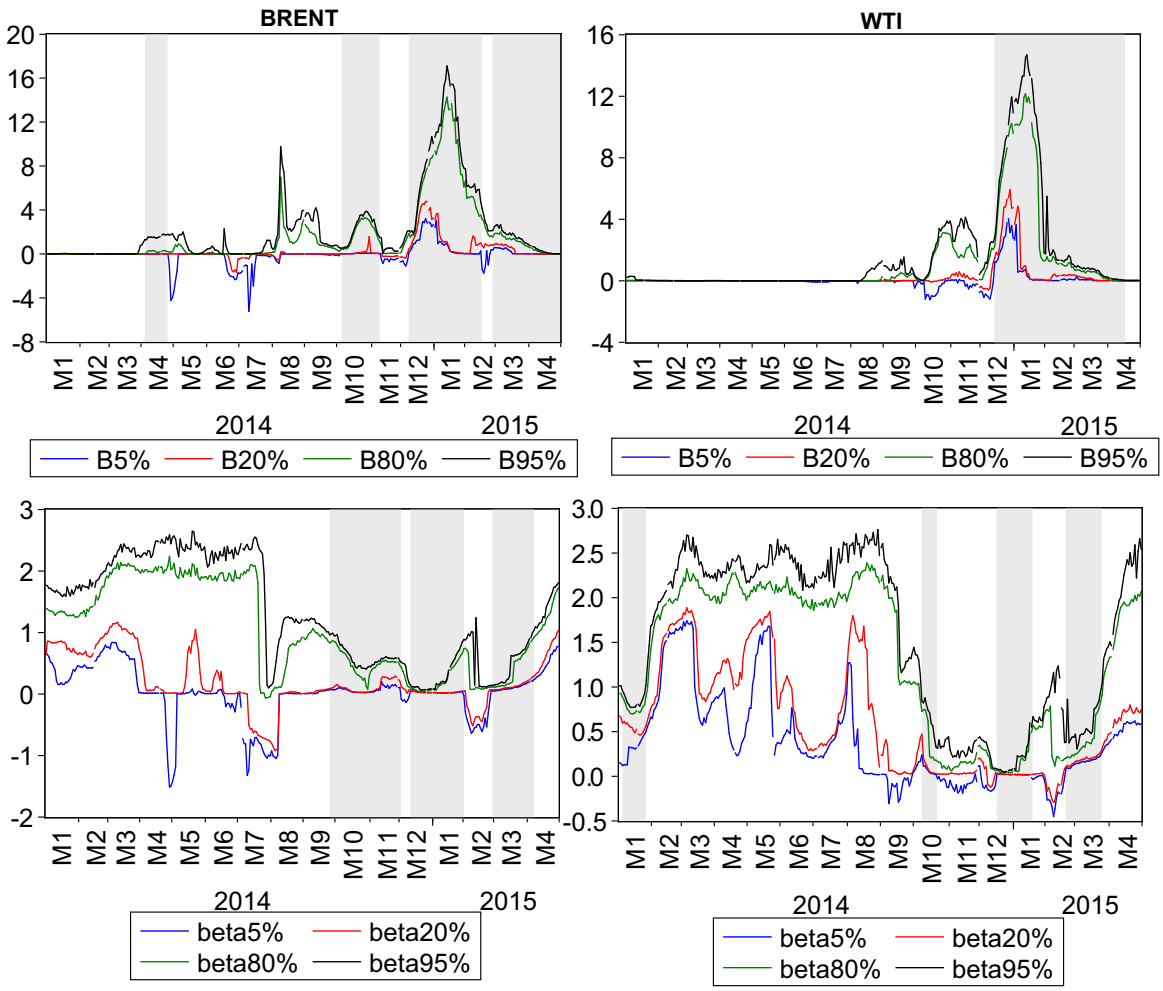


Fig. 6. 20%/80% and 5%/95% quantile range of the LPPL parameters B (first row) and β (second row): shaded areas highlight the time samples when $B > 0$ and $0 < \beta < 1$. *Real oil prices*: Brent (first column) and WTI (second column).

because it has higher power when the underlying data-generating process is an AR(1) process with a coefficient close to one, see Geraskin and Fantazzini (2013). Substantially, the model by Lin et al. (2014) adds an additional restriction to the original LPPL model.

Following Geraskin and Fantazzini (2013) and Lin et al. (2014), we first computed the fraction P_{LPPL} of the previous estimation windows $[t_2 - j;\ t_2]$, $j = 120, \ldots, 250$ that met the LPPL conditions for a negative bubble. Then, we computed the conditional probability $P_{Stat.Res.|LPPL}$ that, out of the fraction P_{LPPL} of windows that satisfied the LPPL conditions, the null hypothesis of stationarity was not rejected for the residuals. The sequences of the probabilities P_{LPPL} and $P_{Stat.Res.|LPPL}$ for nominal and real oil prices are reported in Fig. 8.

The probabilities $P_{Stat.Res.|LPPL}$ are almost always higher than 50% and often close to 100%, thus confirming the previous evidence in the baseline case. These results are similar to those reported by Jiang et al. (2010), Geraskin and Fantazzini (2013) and Lin et al. (2014).

5.3. Longer time sample

The estimation sample used in the baseline case range from January 2013 till May 2015. We wanted to verify that our results continue to hold with a longer sample. In this regard, we used the range January 2005 - June 2015, which is the time span used by Baffes et al. (2015) for their empirical analysis. Table 2 reports the GSADF and GSADF* statistics with the 95% critical values, while the sequences of BSADF and BSADF* statistics (with 95% critical values) for nominal and real oil prices are reported in Figs. 9 and 10, respectively.[5] Similarly to the baseline case, we imposed a minimum length requirement of $4 \cdot \log(T) \approx 32$ days.

The results in Table 2 and in Figs. 9 and 10 not only confirm what we found in the baseline case, but also show that the evidence of explosive behavior in oil prices is stronger for the sample 2014–2015 than for the 2008 oil crash. Moreover, differently from the baseline case, the GSADF* tests are now significant at the 95% level and the identified periods of significant market implosion are June 2008-September 2008 and June 2014-November 2014. The GSADF* test for the Brent real oil price is not significant at the 95% level but only at the 90% level. In this regard, the BSADF* date-stamping procedure seems to anticipate the real market recovery date by a couple of months for both episodes of price declines (2008 and 2014/15). We remark that a large body of the literature examined the oil price crash in 2008 and the underlying factors to the price build-up before this crash: see (among many), Sornette et al. (2009); Khan (2009); Tokic (2010); Lombardi and Van Robays (2011); Areal et al. (2013); Hamilton (2009, 2009, 2011) and Kilian and Murphy (2014). Therefore, we refer the interested reader to these references for more details. In general, this evidence strengthens the case of a negative bubble in oil prices at the end of

[5] The LPPL approach was not considered here because it is not intended for detecting multiple bubbles over a long time span.

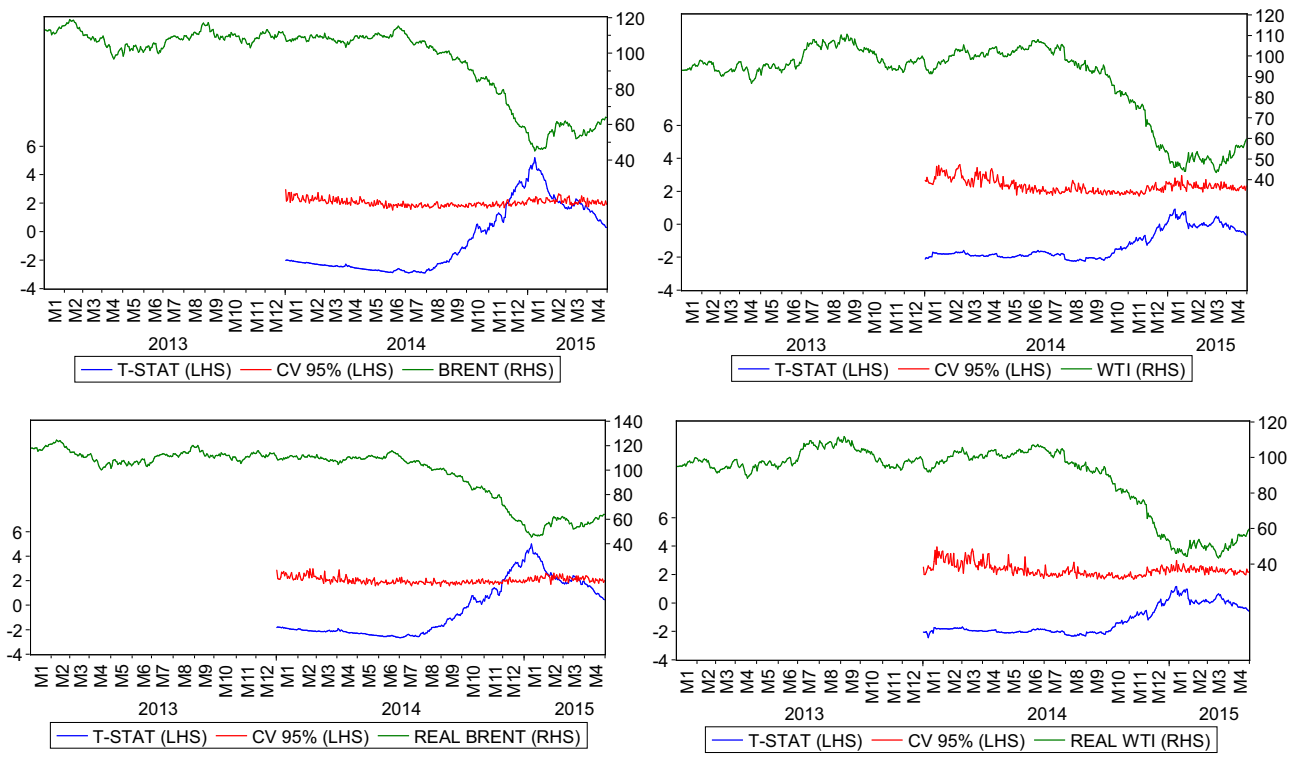


Fig. 7. T-statistics of the FTS parameter γ computed with an expanding window: Brent (first column) and WTI (second column). Nominal oil prices are in the first row, while real price in the second row.

2014 - beginning of 2015, which decreased the prices beyond the level justified by economic fundamentals.

5.4. Tests with lower frequency data

The analysis has so far considered only daily data because we could estimate the competing tests using the most recent observations, an advantage highlighted by the World Bank in the work by Baffes et al. (2015). However, for sake of generality, we considered also a particular weekly dataset which could give additional insights about the 2014/2015 oil price crash.

The US EIA has published weekly the total amount of crude oil stocks in the US since January 1986. We used this data to compute the weekly *supply ratio*, that is the ratio of the WTI nominal price

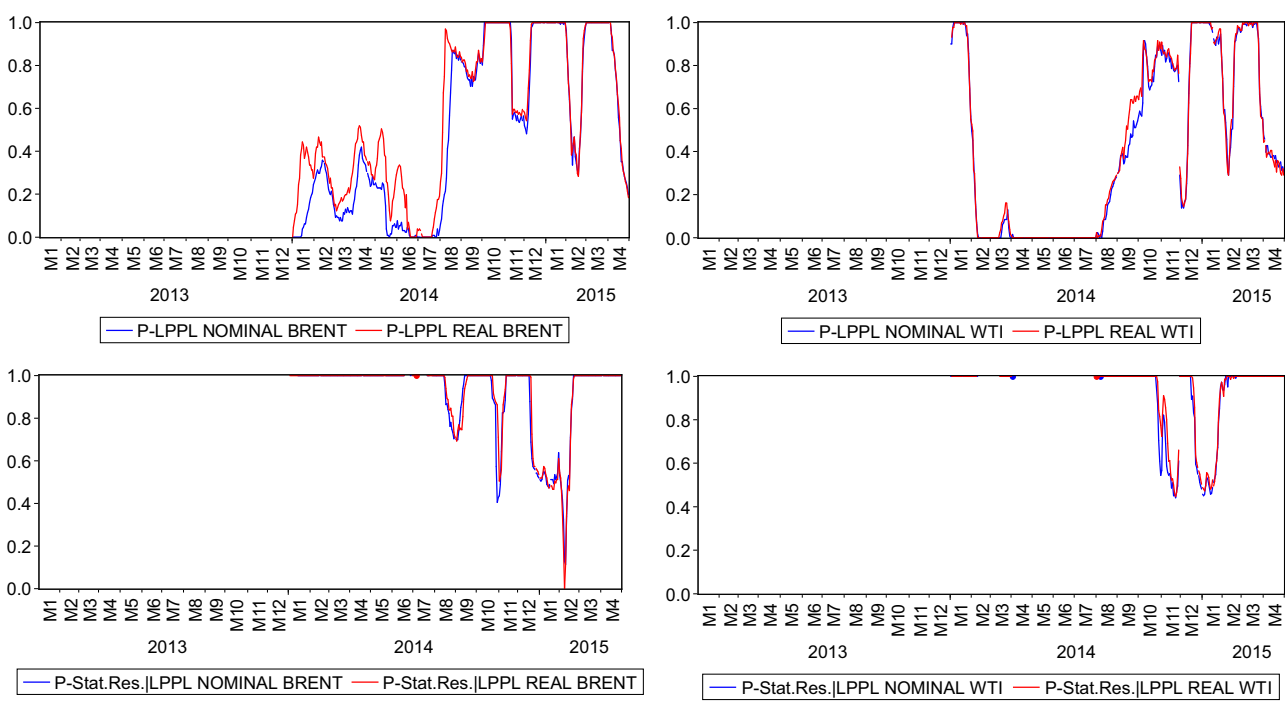


Fig. 8. The fraction P_{LPPL} of estimation windows that met the LPPL conditions (*first row*). The conditional probability $P_{Stat.Res.|LPPL}$ that, out of the fraction P_{LPPL} of windows that satisfied the LPPL conditions, the null hypothesis of stationarity was not rejected for the residuals (*second row*). Brent (first column) and WTI (second column). If $P_{LPPL} = 0$, then $P_{Stat.Res.|LPPL} = NA$.

Table 2.
GSADF statistics (upper part) and GSADF* statistics (lower part), 95% critical values, start and end dates for explosive behavior (GSADF test), crisis origination date and market recovery date (GSADF* test). * Significant at the 95% level.

Oil price	GSADF (PSY-2015)	c.v.	Start (explosive beh.)	End (explosive beh.)
Nominal Brent	4.59*	2.31	05.05.2008	21.07.2008
			24.10.2008	31.12.2008
			02.10.2014	14.04.2014
Nominal WTI	3.34*	2.34	05.05.2008	17.07.2008
			31.10.2014	06.04.2015
Real Brent	4.54*	2.35	05.05.2008	15.07.2008
			15.10.2008	31.12.2008
			02.10.2014	14.04.2014
Real WTI	3.31*	2.38	05.05.2008	15.07.2008
			31.10.2014	06.04.2015
Oil price	**GSADF***(PS-2014)	**c.v.**	**Crisis origination date**	**Market recovery date**
Nominal Brent	2.45*	2.31	06.06.2008	15.09.2008
			16.05.3014	26.11.2014
Nominal WTI	3.74*	2.34	03.06.2014	12.11.2014
Real Brent	2.29	2.40	/	/
Real WTI	3.48*	2.38	09.06.2014	26.11.2014

relative to the US inventory supply stock. This ratio was used by Phillips and Yu (2011) and Caspi et al. (2015) as an alternative proxy variable to measure the fundamental value of oil, using a measure of the oil supply based on the inventory of crude oil in the United States. Table 3 reports the GSADF statistic with the 95% critical values, while the sequence of BSADF statistics (with 95% critical values) for the supply ratio is reported in Fig. 11[6] . The results for the weekly nominal WTI price as well as for the weekly real WTI price are also reported for comparison purposes. The latter price was computed with a methodology similar to that described in section 3.1. Given the use of weekly data, we imposed a minimum length requirement of $\log(T) \approx 7$ weeks.

Table 3 and Fig. 11 provide some evidence of explosive behavior in oil prices from the end of 1999 till March 2000: after reaching a minimum close to 10 $ in December 1998, the WTI rose nearly threefold by March 2000, as world petroleum consumption strongly increased. This was followed by another decline in 2001, following the DotCom bubble and the subsequent recession in the US. However, while the sequences of BSADF statistics for the supply ratio and the nominal WTI price agree on a potential bubble at the beginning of 2000, this is not confirmed by the BSADF statistics for the real WTI price, which is in line with the evidence reported by Caspi et al. (2015) who did not find any price explosivity for this time span using monthly data and the same test procedure. Instead, all three GSADF tests show a period of price explosivity between the end of 2007 and August/September 2008, a range close to those reported by Phillips and Yu (2011) and Caspi et al. (2015) with monthly data. Finally, all three tests identify a period of price explosivity between December 2014 and March 2015, thus confirming our previous evidence. It is interesting to note that the BSADF statistics for nominal and real WTI prices reach a value of 2 or higher in the latter time span, whereas they are much lower (but still significant) for the supply ratio. This may be due to the strong build-up in US oil inventories since January 2015 due to shale oil: the supply ratio is clearly more sensitive to the oil excess supply, which is definitely one of the main factors behind the price crash, as highlighted by Arezki and Blanchard (2014); Baumeister and Kilian (2016) and Baffes et al. (2015).

We also considered two monthly datasets: (1) the US refiners'

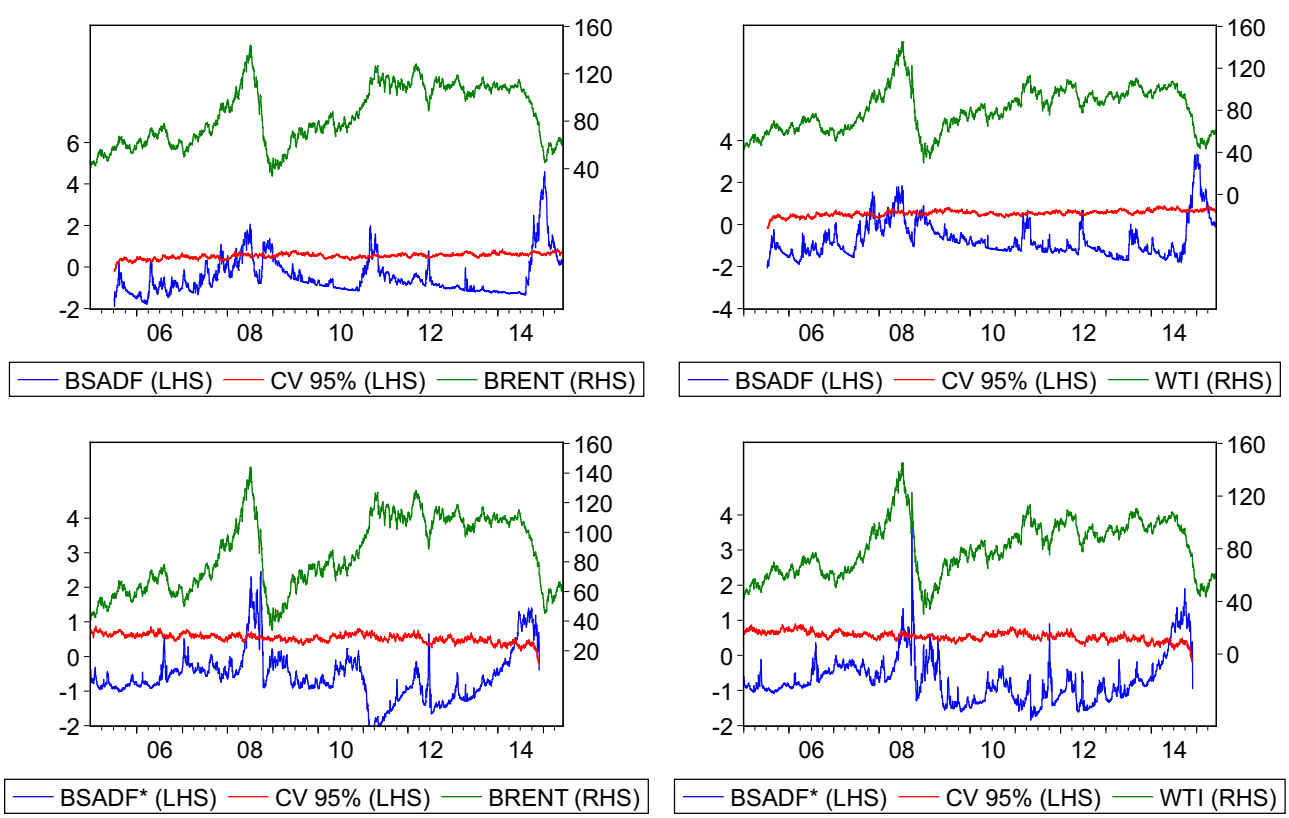


Fig. 9. BSADF statistics (first row) and BSADF* statistics (second row) for date-stamping periods of explosive behavior and market implosions, respectively. *Nominal oil prices*: Brent (first column) and WTI (second column).

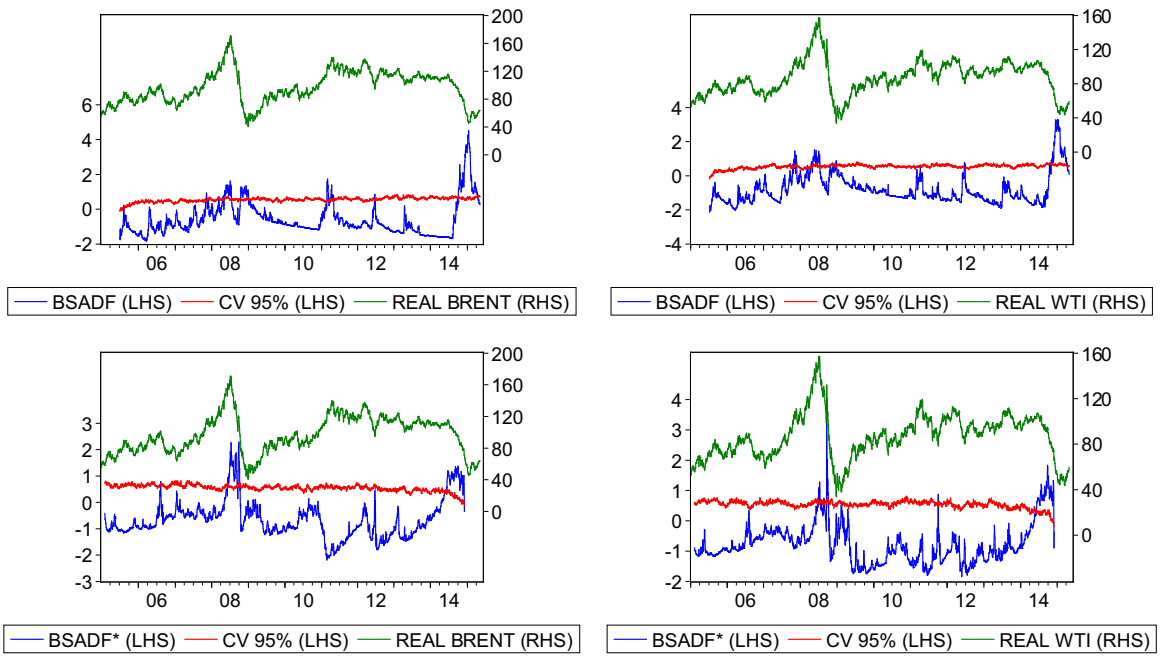


Fig. 10. BSADF statistics (first row) and BSADF* statistics (second row) for date-stamping periods of explosive behavior and market implosions, respectively. *Real oil prices*: Brent (first column) and WTI (second column).

Table 3.
GSADF statistics, 95% critical values, start and end dates for explosive behavior (GSADF test). * Significant at the 95% level.

Oil price	GSADF (PSY-2015)	c.v.	Start (explosive beh.)	End (explosive beh.)
WTI Supply ratio	4.54*	2.31	05.11.1999	17.03.2000
			12.10.2007	19.09.2008
			12.12.2014	13.03.2015
Nominal WTI	4.81*	2.31	14.01.2000	17.03.2000
			07.09.2007	26.09.2008
			28.11.2014	27.03.2015
Real WTI	3.66*	2.36	08.02.2008	22.08.2008
			28.11.2014	27.03.2015

acquisition cost for imported crude oil, as reported by the EIA, extrapolated from 1974M1 back to 1973M1 as in Barsky and Kilian (2002). Kilian and Murphy (2014) suggest to use this oil price since it is a better proxy for the price of oil in global markets than the US price of domestic crude oil, which was regulated during the 1970s and early 1980s; (2) the monthly Brent oil prices as provided by the IMF since January 1980. The GSADF tests were strongly significant in both cases and identified a period of price explosivity from November/December 2014 till February/March 2015, depending on the type of oil price selected and whether nominal or real prices are considered.[7] However, the imposition of a minimum bubble-duration length in this case would eliminate this evidence, even considering the smallest length requirement possible of $1 \cdot \log(T) \approx 6$ months. Given that all previous results point out to a (negative) bubble of 4/5 months, probably the tuning parameter δ discussed above should be smaller than 1. However, this technical issue goes beyond the scope of this paper and we leave it as an avenue of further research. This is why we do not report here the results with monthly data, but they are available from the authors upon request.

5.5. The price fall in 2015/2016: preliminary evidence

At the time of finishing writing this work (May 2016), the oil price experienced a new fall during the winter period in 2015/2016. While a full analysis of this event will be discussed in a separate work -due to the computational efforts needed and the lack of data-, we nevertheless present some preliminary evidence using the GSADF test and the LPPL model with the most recent data of the real WTI oil price till April 2016.[8] The GSADF and GSADF* statistics and the sequences of BSADF and BSADF* statistics (with 95% critical values) for the real WTI price are reported in Fig. 12 (left column), while the 20%/80% and 5%/95% quantile ranges of the LPPL parameters B and β are reported in Fig. 12 (right column).

The results in Fig. 12 not only confirmed again the presence of a negative bubble in oil prices at the end of 2014 - beginning of 2015, but the evidence in this case is even stronger than in the baseline case. Interestingly, both the GSADF test and the LPPL model did not find any significant evidence of a negative bubble during the winter period in 2015/2016. However, the full analysis of this event will be developed in a separate work.

6. Conclusions and policy implications

The aim of this paper is to propose a potential explanation for the part of the oil price decline in 2014/15 which can not be explained using supply and demand alone. More specifically, we suggest that there was a negative financial bubble which decreased oil prices beyond the level justified by economic

[6] The BSADF* statistic to test for significant bubble implosion was not considered because the initial minimum window size g_0 on the reversed time series eliminates the last year and half of data up to the beginning of 2014, so that it is not useful for this analysis. Similarly to Section 5.3, the LPPL approach was also not considered because it is not intended for detecting multiple bubbles over a long time span.

[7] Other periods of price explosivity were also detected, but are not of interest for the current analysis.

[8] The author wants to thank an anonymous referee for pointing out this issue.

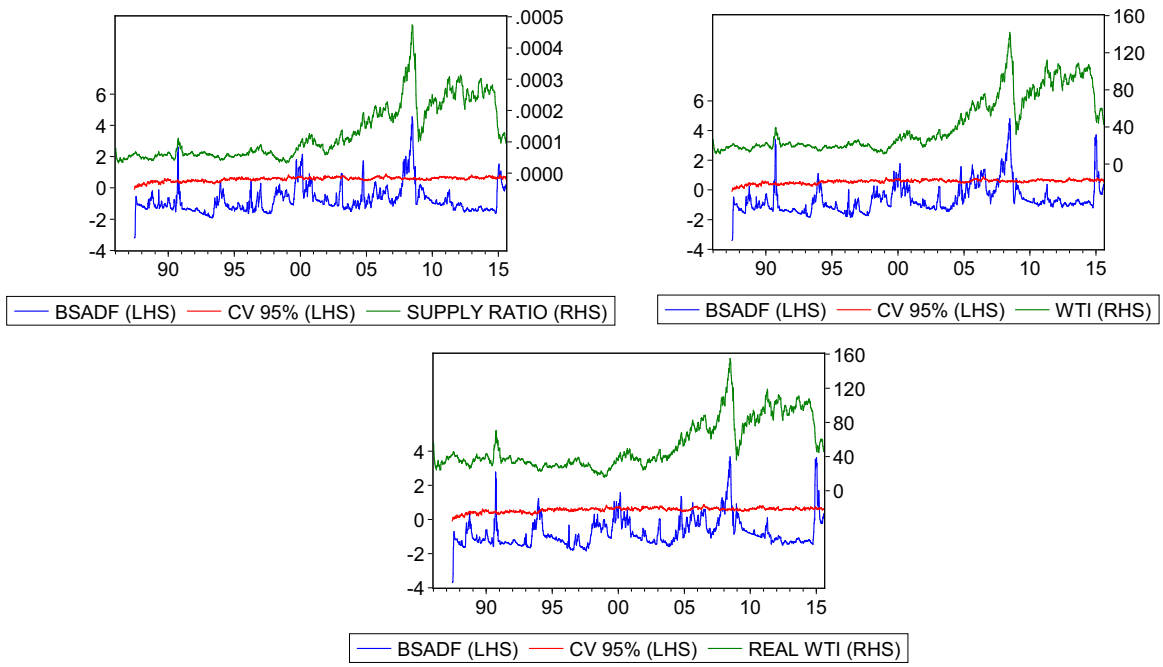


Fig. 11. BSADF statistics for date-stamping periods of explosive behavior. Supply ratio (first plot), nominal WTI prices (second plot) and real WTI prices (third plot).

fundamentals.

We employed two sets of bubble detection strategies to corroborate this proposition: the first set consisted of tests for financial bubbles proposed by Phillips et al. (2016) and Phillips and Shi (2014). These tests are based on recursive and rolling right-tailed Augmented Dickey-Fuller unit root test, wherein the null hypothesis is of a unit root and the alternative is of a mildly explosive process. They can identify periods of statistically significant explosive price behavior and date-stamp their occurrence. The second set consisted of the log-periodic power law (LPPL) model

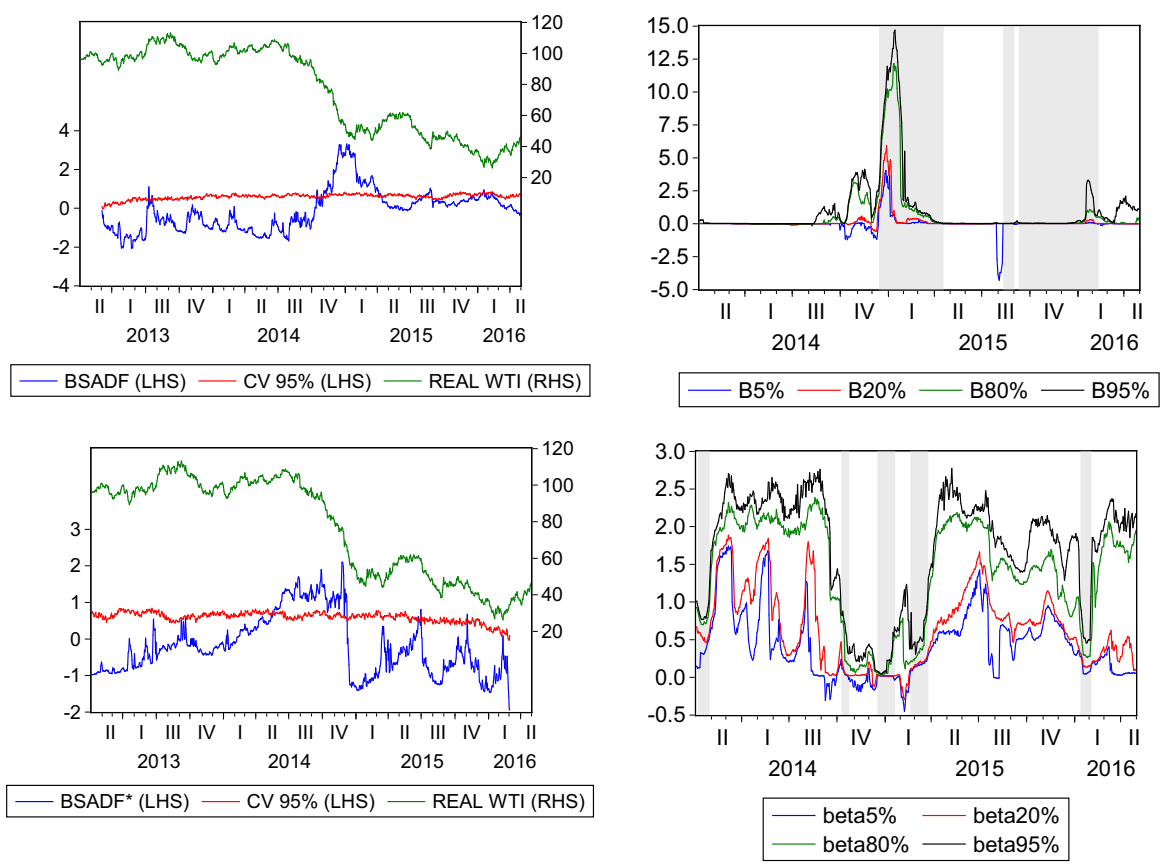


Fig. 12. *Real WTI*: BSADF statistics (left-first row) and BSADF* statistics (left-second row), respectively. The GSADF statistic was equal to 3.31 ($CV_{95} = 2.39$), whereas the GSADF* statistic to 2.10 ($CV_{95} = 2.39$). 20%/80% and 5%/95% quantile ranges of the LPPL parameters B (right-first row) and β (right-second row): shaded areas highlight the time samples when $B > 0$ and $0 < \beta < 1$.

for negative financial bubbles developed by Yan et al. (2012). This model adapts the Johansen-Ledoit-Sornette (JLS) model of rational expectation bubbles developed by Sornette et al. (1999); Johansen et al. (1999) and Johansen et al. (2000) to the case of a price fall occurring during a transient negative bubble. Despite the methodological differences between these bubble detection methods, they provided the same result: the oil price experienced a statistically significant negative financial bubble in the last months of 2014 and at the beginning of 2015.

A set of robustness checks showed that our results also hold with different tests, model set-ups and alternative datasets: all checks confirmed that the oil price experienced a statistically significant negative financial bubble from the end of 2014 till the beginning of 2015, thus supporting the idea put forward by Domanski et al. (2015) and Tokic (2015) that this price collapse cannot be explained by supply and demand alone,

These results can be important for regulatory purposes, since it is clear that the enhanced regulations imposed after the 2008 oil bubble (see Collins (2010) and Cosgrove (2009)) cannot ensure the oil price efficiency. In this regard, Tokic (2015) and Domanski et al. (2015) suggested that the oil price collapse 2014/2015 could have been caused by the increased leverage of oil firms (the debt of oil and gas sector increased from $1 trillion in 2006 to $ 2.5 trillion in 2014): the increasing need to keep high production levels and to hedge future production to satisfy financial constraints could have easily amplified the initial price decline due to economic fundamentals. Therefore, a revised and more effective regulatory framework should include not only oil traders/speculators, but all market participants including oil producers. The design of this revised framework is definitively an important avenue of future research.

Another implication of the evidence found in this work is that market regulators should be concerned not only about positive price bubbles, but also about negative bubbles. In this regard, it is well known that the oil supply shows cyclical boom and bust cycles in prices and production, see Maugeri (2010) for a large historical review. Extremely low prices are not necessarily beneficial, even for countries which are (mainly) oil consumers: for example, Kilian (2008) showed that the large fall in investment in the oil and gas industry following the oil price crash in 1985/1986 was one of the main causes why real consumption in the US did not grow as expected. In general, there is a large literature which tried to find if and why economic activity responds asymmetrically to oil price shocks -i.e. high oil prices decrease economic activity much more than low oil prices stimulate it-, see the *Macroeconomic Dynamics* special issue on "Oil Price Shocks" published in 2011 for more details. Moreover, several authors have recently investigated the linkages between the oil market and other markets, focusing particularly on the volatility transmission across financial markets. Diebold and Yilmaz (2012) found that cross-market volatility spillovers across US stock, bond, foreign exchange and commodities markets were quite limited until 2007, but have increased since then: particularly, they found that the commodity market was a net recipient of small levels of volatility shocks from the other markets till 2007, but it has become a net transmitter after the beginning of the global financial crisis. Similar evidence was found by Ji and Fan (2012) who found that the crude oil market has significant volatility spillover effects on non-energy commodity markets and they have strengthened after the crisis. A similar result was also reported by Creti et al. (2013) who showed increased links between stock and commodity markets, and by Gomes and Chaibi (2014) who highlighted that shock and volatility spillovers tend to go more often from oil to stock markets than viceversa, see also Arouri and Nguyen (2010); Filis et al. (2011); Kumar et al. (2012); Awartani and Maghyereh (2013) and Khalfaoui et al. (2015). Given this increased influence of the oil market on the other markets, regulators should consider a regulatory framework able to mitigate an oil price crash due to panic selling and/or market manipulation: a potential starting point could be the model developed by Dutt and Harris (2005), which can be used to set position limits for cash-settled derivative contracts.

Appendix A. Appendix

We estimated (9) with nonlinear least-squares, using a variant of the 3-step procedure proposed in Geraskin and Fantazzini (2013) and Fantazzini (2010):

1. Set $t_c = t_2 + 0.1\cdot(t_2 - t_1)$, where t_2 and t_1 are the last and the first observation of the estimation sample, respectively. Estimate the remaining LPPL parameters $[A, B, C_1, C_2, \beta, \omega]$ by using the BFGS (Broyden, Fletcher, Goldfarb, Shanno) algorithm.
2. Keeping fixed the LPPL parameters $\hat{\Phi} = [\hat{A}, \hat{B}, \hat{C}_1, \hat{C}_2, \hat{\beta}, \hat{\omega}]$ computed in the first stage, estimate the critical time t_c.
3. Use the estimated parameters in the first and second stages as *starting values* for estimating all the LPPL parameters.

Similarly to Geraskin and Fantazzini (2013), we found that this multi-step procedure improves considerably the numerical convergence and the estimation efficiency in small-to-medium sized samples.

References

Areal, F.J., Balcombe, K., Rapsomanikis, G., 2013. Testing for bubbles in agriculture commodity markets, Discussion paper, ESA Working Paperp. 14.

Arezki, R., Blanchard, O., Seven questions about the recent oil price slump IMFdirect - IMF Blog 2014 22, December 2014.

Arouri, M.E.H., Nguyen, D.K., 2010. Oil prices, stock markets and portfolio investment: evidence from sector analysis in Europe over the last decade. Energy Policy 38 (8), 4528–4539.

Awartani, B., Maghyereh, A.I., 2013. Dynamic spillovers between oil and stock markets in the gulf cooperation council countries. Energy Econ. 36, 28–42.

Baffes, J., Kose, M., Ohnsorge, F., Stocker, M., 2015. The great plunge in oil: causes, consequences, and policy responses. World Bank Group, Dev. Econ.

Barsky, R.B., Kilian, L., 2002. Do we really know that oil caused the great stagflation? A monetary alternative, In NBER Macroeconomics Annual 2001, Volume 16, pp. 137–198. MIT Press.

Baumeister, C., Kilian, L., 2016. Understanding the Decline in the Price of Oil Since June 2014. Journal of the Association of Environmental and Resource Economists 3 (1), 131–158.

Bhattacharya, U., Galpin, N., Ray, R., Yu, X., 2009. The role of the media in the internet IPO bubble. J. Financ. Quant. Anal. 44 (03), 657–682.

Bothmer, H.-C.G.V., Meister, C., 2003. Predicting critical crashes? A new restriction for the free variables. Physica A: Stat. Mech. Appl. 320, 539–547.

Case, K.E., Shiller, R.J., 2003. Is there a bubble in the housing market? Brook. Pap. Econ. Act. 2003 (2), 299–362.

Caspi I., Katzke N. and Gupta R., Date stamping historical periods of oil price explosivity: 1876–2014, Energy Economics, Available online 18 April 2015, ISSN 0140-9883, http://dx.doi.org/10.1016/j.eneco.2015.03.029.

Collins, D., 2010. CFTC steps up on limits: futures news, analysis & strategies for futures. Option Deriv. Traders 39 (2), 14.

Corsi, F., Sornette, D., 2014. Follow the money: the monetary roots of bubbles and crashes. Int. Rev. Financ. Anal. 32, 47–59.

Cosgrove, M., 2009. Banishing energy speculators is not the solution. FOW: Glob. Deriv. Manag. 454, 39.

Creti, A., Joëts, M., Mignon, V., 2013. On the links between stock and commodity markets' volatility. Energy Econ. 37, 16–28.

Diebold, F.X., Yilmaz, K., 2012. Better to give than to Predictive directional measurement of volatility spillovers. Int. J. Forecast. 28 (1), 57–66.

Domanski, D., Kearns, J., Lombardi, M.J., Shin, H.S., 2015. Oil and Debt. BIS Q. Rev. (March), 55–65.

Dutt, H.R., Harris, L.E., 2005. Position limits for cash-settled derivative contracts. J. Futur. Mark. 25 (10), 945–965.

Dyck, A., Zingales, L., 2003. The bubble and the media. In: Cornelius, P., Kogut, B. (Eds.), Corporate Governance and Capital Flows in a Global Economy. Oxford University Press, pp. 83–104.

EIAsteo15 EIA (2015): "STEO," ⟨http://www.eia.gov/forecasts/steo/⟩.

Fantazzini, D., 2010: THE HANDBOOK OF TRADING, chap. Modelling bubbles and anti-bubbles in bear markets: a medium-term trading analysis, McGraw-Hill Finance and Investing pp. 365–388.

Figuerola-Ferretti, I., Gilbert, C.L., McCrorie, J.R., 2015. Testing for Mild Explosivity

and Bubbles in LME Non-Ferrous Metals Prices. Journal of Time Series Analysis 36 (5), 763–782.
Figuerola-Ferretti, I., Gonzalo, J., 2010. Modelling and measuring price discovery in commodity markets. J. Econom. 158 (1), 95–107.
Filimonov, V., Sornette, D., 2013. A stable and robust calibration scheme of the log-periodic power law model. Phys. A: Stat. Mech. its Appl. 392 (17), 3698–3707.
Filis, G., Degiannakis, S., Floros, C., 2011. Dynamic correlation between stock market and oil prices: the case of oil-importing and oil-exporting countries. Int. Rev. Financ. Anal. 20 (3), 152–164.
Geraskin, P., Fantazzini, D., 2013. Everything you always wanted to know about log-periodic power laws for bubble modeling but were afraid to ask. Eur. J. Financ. 19 (5), 366–391.
Gilbert, C.L., 2010. Speculative influence on commodity prices 2006–08," Discussion paper, Discussion Paper 197, United Nations Conference on Trade and Development (UNCTAD), Geneva.
Gomes, M., Chaibi, A., 2014. Volatility spillovers between oil prices and stock returns: a focus on frontier markets. J. Appl. Bus. Res. 30 (2), 509–526.
Hamilton, J.D., 2009a. Causes and Consequences of the Oil Shock of 2007–08, Discussion paper, National Bureau of Economic Research, Working Paper No. 15002.
Hamilton, J.D., 2009. Understanding crude oil prices. Energy J. 30 (2), 179–206.
Hamilton, J.D., 2011. Historical oil shocks, Discussion paper. National Bureau Of Economic Research,.
Hamilton, J.D., Oil prices as an indicator of global economic conditions Econbrowser Blog entry, 14 December, 2014.
Harvey D.I., Leybourne S.J., Sollis R. and Taylor A.R., Tests for explosive financial bubbles in the presence of non-stationary volatility, Journal of Empirical Finance, Available online 25 September 2015, ISSN 0927-5398, http://dx.doi.org/10.1016/j.jempfin.2015.09.002.
Homm, U., Breitung, J., 2012. Testing for speculative bubbles in stock markets: a comparison of alternative methods. J. Financ. Econom. 10 (1), 198–231.
Ji, Q., Fan, Y., 2012. How does oil price volatility affect non-energy commodity markets? Appl. Energy 89 (1), 273–280.
Jiang, Z.-Q., Zhou, W.-X., Sornette, D., Woodard, R., Bastiaensen, K., Cauwels, P., 2010. Bubble diagnosis and prediction of the 2005–2007 and 2008–2009 chinese stock market bubbles. J. Econ. Behav. Organ. 74 (3), 149–162.
Johansen, A., 2003. Characterization of large price variations in financial markets. Phys. A: Stat. Mech. Appl. 324 (1), 157–166.
Johansen, A., Ledoit, O., Sornette, D., 2000. Crashes as critical points. Int. J. Theor. Appl. Financ. 3 (2), 219–255.
Johansen, A., Sornette, D., Ledoit, O., 1999. Predicting financial crashes using discrete scale invariance. J. Risk 1 (4), 5–32.
Khalfaoui, R., Boutahar, M., Boubaker, H., 2015. Analyzing volatility spillovers and hedging between oil and stock markets: evidence from wavelet analysis. Energy Econ. 49, 540–549.
Khan, M.S., 2009. The 2008 Oil Price Bubble, Discussion paper, Peterson Institute for International Economics Policy Brief, Washington, DC.
Kilian, L., 2008. The economic effects of energy price shocks. J. Econ. Lit. 46 (4), 871–909.
Kilian, L., Murphy, D., 2014. The role of inventories and speculative trading in the global market for crude oil. J. Appl. Econom. 29, 454–478.
Kumar, S., Managi, S., Matsuda, A., 2012. Stock prices of clean energy firms, oil and carbon markets: a vector autoregressive analysis. Energy Econ. 34 (1), 215–226.
Kwiatkowski, D., Phillips, P., Schmidt, P., Shin, Y., 1992. Testing the null of stationarity against the alternative of a unit root: how sure are we that the economic time series have a unit root? J. Econom. 54, 159–178.
Lammerding, M., Stephan, P., Trede, M., Wilfling, B., 2013. Speculative bubbles in recent oil price dynamics: evidence from a bayesian markov-switching state-space approach. Energy Econ. 36, 491–502.
Lin, L., Ren, R., Sornette, D., 2014. The volatility-confined LPPL model: a consistent model of 'explosive' financial bubbles with mean-reverting residuals. Int. Rev. Financ. Anal. 33, 210–225.
Lombardi, M.J., Van Robays, I., 2011. Do financial investors destabilize the oil price?, Discussion paper, Working Paper Series No. 1346. European Central Bank, Frankfurt/Main.
Maugeri, L., 2010. Beyond the age of oil: the myths, realities, and future of fossil fuels and their alternatives. Praeger.
Phillips, P., Shi, S., 2014. Financial bubble implosion, Discussion paper.
Phillips, P., Shi, S., Yu, J., 2014. Specification sensitivity in right-tailed unit root testing for explosive behaviour. Oxf. Bull. Econ. Stat. 76 (3), 315–333.
Phillips, P., Shi, S., Yu, J., 2016. Testing for multiple bubbles: historical episodes of exuberance and collapse in the SP500. Int. Econ. Rev. (forthcoming)
Phillips, P., Shi, S., Yu, J., 2016. Testing for multiple bubbles limit theory of real time detectors. Int. Econ. Rev. (forthcoming)
Phillips, P., Wu, Y., Yu, J., 2011. Explosive behavior in the 1990S Nasdaq: when did exuberance escalate asset values? Int. Econ. Rev. 52 (1), 201–226.
Phillips, P., Yu, J., 2011. Dating the timeline of financial bubbles during the subprime crisis. Quant. Econ. 2 (3), 455–491.
Pindyck, R.S., 1992. The present value model of rational commodity pricing, Discussion paper, National Bureau of Economic Research.
Shi, S., Arora, V., 2012. An application of models of speculative behaviour to oil prices. Econ. Lett. 115 (3), 469–472.
Shiller, R.J., 2000. Irrational Exuberance. Princeton University Press.
Shiller, R.J., 2002. Bubbles, human judgment, and expert opinion. Financ. Anal. J. 58 (3), 18–26.
Sornette, D., 2003. Why Stock Markets Crash: Critical Events in Complex Financial Systems. Princeton University Press.
Sornette, D., Johansen, A., et al., 2001. Significance of log-periodic precursors to financial crashes. Quant. Financ. 1 (4), 452–471.
Sornette, D., Ledoit, O., Johansen, A., 1999. Critical crashes. Risk Mag. 12 (1), 91–95.
Sornette, D., Woodard, R., Zhou, W.-X., 2009. The 2006–2008 oil bubble: evidence of speculation, and prediction. Physica A: Stat. Mech. Appl. 388 (8), 1571–1576.
Sornette, D., Zhou, W.-X., 2006. Predictability of large future changes in major financial indices. Int. J. Forecast. 22 (1), 153–168.
Tokic, D., 2010. The 2008 oil bubble causes and consequences. Energy Policy 38 (10), 6009–6015.
Tokic, D., 2015. The 2014 oil bust: causes and consequences. Energy Policy 85, 162–169.
Veldkamp, L.L., 2006. Media frenzies in markets for financial information. Am. Econ. Rev. 96 (3), 577–601.
Yan, W., Woodard, R., Sornette, D., 2012. Diagnosis and prediction of rebounds in financial markets. Physica A: Stat. Mech. Appl. 391 (4), 1361–1380.
Yiu, M., Yu, J., Lu, J., 2013. Detecting bubbles in Hong Kong residential property. J. Asian Econ. 28, 115–124.
Zhou, W.-X., Sornette, D., 2003. 2000–2003 real estate bubble in the UK but not in the USA. Phys. A: Stat. Mech. Appl. 329 (1), 249–263.
Zhou, W.-X., Sornette, D., 2006. Is there a real-estate bubble in the US? Phys. A: Stat. Mech. Appl. 361 (1), 297–308.
Zhou, W.-X., Sornette, D., 2008. Analysis of the real estate market in Las Vegas: bubble, seasonal patterns, and prediction of the CSW indices. Phys. A: Stat. Mech. Appl. 387 (1), 243–260.
Zhou, W.-X., Sornette, D., 2009. A case study of speculative financial bubbles in the South African stock market 2003–2006. Phys. A: Stat. Mech. Appl. 388 (6), 869–880.

Appendix 14 "Everything you always wanted to know about bitcoin modelling but were afraid to ask - Part 1."

Fantazzini, D., Nigmatullin, E., Sukhanovskaya, V., & Ivliev, S. (2016). Everything you always wanted to know about bitcoin modelling but were afraid to ask - Part 1. *Applied Econometrics*, 44, 5-24. http://pe.cemi.rssi.ru/pe_2016_44_005-024.pdf

Permisson to post the Publisher's version/PDF: according to http://appliedeconometrics.cemi.rssi.ru/AppEc.files/contents_en.html, "*full texts are available for downloading with approximately one year lag*". In some cases, this 1-year lag can be shorter.

Прикладная эконометрика, 2016, т. 44, с. 5–24.
Applied Econometrics, 2016, v. 44, pp. 5–24.

Д. Фантаццини, Э. М. Нигматуллин, В. Н. Сухановская, С. В. Ивлиев[1]

Все, что вы хотели знать о моделировании биткойна, но боялись спросить. Часть I

Биткойн — это децентрализованная цифровая валюта и платежная система с открытым кодом. Она привлекла к себе значительное внимание и глобальный интерес, возрастающее число статей посвящено вопросам ее операционной, экономической и финансовой жизнеспособности. Данная статья рассматривает эконометрические и математические инструменты, которые были предложены к настоящему времени для моделирования цены биткойна, и различные другие аспекты, включая преимущества и ограничения. Обсуждаются методы определения основных характеристик биткойн-пользователей, модели оценки фундаментальной стоимости биткойна, эконометрические подходы к моделированию динамики цены биткойна, тесты для выявления финансовых пузырей в ценах и методологии, предложенные для изучения ценообразования на биткойн-биржах.

Ключевые слова: биткойн; криптовалюты; хэшрейт; майнинг; привлекательность инвесторов; социальное взаимодействие; денежное предложение; спрос на деньги; спекуляции; прогнозирование; алгоритмическая торговля; пузыри; ценообразование; LPPL.

JEL classification: C22; C32; C51; C53; E41; E42; E47; E51; G17.

1. Введение

Биткойн — это онлайновая децентрализованная валюта, которая позволяет пользователям покупать товары и услуги и осуществлять транзакции без привлечения третьей стороны. Она была запущена в 2009 году человеком или (что более вероятно) группой людей, действовавшей под псевдонимом Сатоши Накамото. Биткойн заложил основу для развития так называемых «криптовалют», которые базируются на криптографических средствах защиты. Основным свойством криптовалют является их децентрализованная структура: для их функционирования не требуется центральный орган, который выпускает и регулирует валюту, а транзакции осуществляются с использованием пирингового протокола без каких-либо посредников. Обзоры биткойна с точки зрения структуры и операций приведены в (Becker et al., 2013; Segendorf, 2014; Dwyer, 2015; Böhme et al., 2015) и на сайте

[1] **Фантаццини Деан** — Московская школа экономики, МГУ, Москва; fantazzini@mse-msu.ru.
Нигматуллин Эрик Маликович — Bocconi University, Милан, Италия; nigmatullin.erik@gmail.com.
Сухановская Вера Николаевна — Пермский государственный национальный исследовательский университет, Пермь; vera-sukhanovskaya@yandex.ru.
Ивлиев Сергей Владимирович — Пермский государственный национальный исследовательский университет, Пермь; ivliev@gmail.com.

www.bitcoin.org. Несколько центральных банков также изучали биткойн, см. (Velde, 2013; Lo, Wang, 2014; Badev, Chen, 2014; Ali et al., 2014; ECB, 2012, 2015). Обсуждение биткойна как потенциальной альтернативы денежной системы приведено в работах (Rogojanu, Badea, 2014; Weber, 2016), а экономика так называемого майнинга изучается в (Kroll et al., 2013). Анализ правовых вопросов, связанных с биткойном, приведен в работах (Allen, 2016; Murphy et al., 2015).

Данная статья рассматривает эконометрические и математические инструменты, которые были предложены к настоящему времени для моделирования цены биткойна и некоторых других аспектов. Насколько известно авторам, подобный обзор отсутствует в финансовой литературе и может быть интересен как профессионалам рынка, так и исследователям, учитывая раннюю стадию развития эмпирической литературы, посвященной биткойну.

Статья организована следующим образом. Раздел 2 описывает принципы, на которых построено функционирование биткойна, и приводит краткую статистику сети. В разделе 3 приведен обзор исследований, посвященных анализу характеристик, свойственных пользователям биткойна. В разделе 4 обсуждаются основные модели, используемые для оценки фундаментальной стоимости биткойна, от моделей размера рынка до моделей предельных издержек майнинга, основанных на энергопотреблении. Раздел 5 описывает несколько эконометрических подходов, предложенных для моделирования ценовой динамики биткойна, начиная с кросс-секционных регрессионных моделей, включающих в себя большинство торгуемых цифровых валют, и заканчивая одномерными и многомерными моделями временных рядов и моделями частотного диапазона. В разделе 6 приводится обзор тестов, используемых для выявления наличия финансового пузыря в ценах биткойна, которые могут быть условно разделены на два класса в зависимости от того, направлены ли они на определение одного пузыря или (потенциально) множественных пузырей. В разделе 7 изучаются методики, предложенные для оценки информационной доли различных биткойн-бирж с точки зрения информации, генерируемой рынком в целом, что является важным как для краткосрочных трейдеров, так и для долгосрочных инвесторов, которые хотели бы понимать, какая из бирж наиболее быстро реагирует на новую информацию. В разделе 8 приводятся основные выводы и предлагаются несколько возможных направлений дальнейших исследований[2].

2. Определение криптовалют и биткойна

2.1. Как устроен биткойн?

2.1.1. Цифровые подписи и криптографическая хеш-функция

В процессе проведения платежей в сети Биткойн используется криптография для проверки транзакций и создания блоков транзакций. В частности, Биткойн опирается на две криптографические схемы: 1) цифровые подписи и 2) криптографические хэш-функции. Первая схема позволяет обмениваться платежными инструкциями между сторонами транзакции, а вторая используется для поддержания дисциплины внесения записей по транзакциям в публичный реестр (блокчейн). Следует заметить, что ни одна из этих схем не явля-

[2] Разделы 5–8 будут опубликованы в другом выпуске журнала.

ется уникальной для Биткойна — все они широко используются для защиты коммерческих и государственных коммуникаций. Приведем краткое описание ниже.

Цифровые подписи используются для удостоверения подлинности сообщения между отправителем и получателем и обеспечивают:

(i) *удостоверение подлинности*: получатель может проверить, что сообщение пришло от отправителя;

(ii) *невозможность отказа*: отправитель не может отказаться от отправки сообщения;

(iii) *целостность*: сообщение не было подделано.

Применение цифровых подписей включает шифрование с открытым ключом, в котором пара ключей — открытый и закрытый — генерируются с определенными желаемыми свойствами. Цифровая подпись используется для подписания сообщений: транзакция подписывается с помощью закрытого ключа, а затем транслируется в сеть Биткойн. Все участники сети могут проверить, что эта транзакция пришла от владельца открытого ключа, взяв сообщение, подпись, открытый ключ и запустив алгоритм проверки.

Криптографическая хеш-функция принимает на входе строку произвольной длины (сообщение m) и возвращает строку с заранее определенной длиной (хэш h). Функция является детерминированной, это означает, что один и тот же m на входе всегда будет давать один и тот же h на выходе. Помимо этого, функция также должна обладать следующими свойствами.

(i) *Стойкость к восстановлению прообраза*. При данном хэше h сложно найти такое сообщение m, при котором хэш(m) = h.

(ii) *Стойкость к коллизиям*. При заданном сообщении $m1$ сложно найти другое сообщение $m2$, при котором хэш($m1$) = хэш($m2$). Другими словами, изменение сообщения приводит к изменению хэша.

Выход хэш-функции выглядит так, как будто получен случайным образом, хотя и является детерминированным. Биткойн в основном использует алгоритм безопасного хеширования SHA-256 — тип Secure Hash Algorithm (SHA-2), разработанный Агентством национальной безопасности и опубликованный Национальным институтом стандартов и технологий, см. (Dang, 2012).

2.1.2. Владение биткойнами и биткойн-адреса

С технической точки зрения биткойны находятся в Биткойн сети по так называемым биткойн-адресам. Владение определенным количеством биткойнов сводится к возможности отправки платежей по сети Биткойн с биткойн-адресов, к которым эти биткойны привязаны. Возможность отправки платежей с биткойн-адресов контролируется с помощью цифровых подписей, которые включают пары открытого и закрытого ключей. В частности, каждый биткойн-адрес индексируется с помощью уникального открытого ID — буквенно-цифрового идентификатора, который, по сути, соответствует открытому ключу. Закрытый ключ контролирует биткойны, хранящиеся по этому адресу. Любой платеж (сообщение), где этот адрес задействован как адрес отправителя, должен быть подписан соответствующим действительным закрытым ключом. Проще говоря, владение биткойнами по указанному биткойн-адресу означает знание закрытого ключа, который соответствует этому адресу.

В любой момент времени каждый биткойн-адрес привязан к биткойн-балансу, который, по сути, является общественной информацией. Каждую существующую или предлагаемую (вновь транслируемую) транзакцию можно проверить на предмет соответствия предшеству-

ющей истории транзакций, т. е. проверить, что суммы по транзакциям действительно имеются в соответствующих биткойн-адресах.

2.1.3. Транзакция в цепочке блоков (блокчейне)

Субъекты участвуют в транзакциях в сети Биткойн посредством набора биткойн-адресов, называемых кошельком (набор биткойн-адресов, принадлежащих одному лицу). Каждая запись транзакции включает один или несколько адресов отправки (*входов*) и один или несколько адресов получения (*выходов*), а также сведения о том, сколько каждый из этих адресов отправил и получил. Пример транзакции представлен на рис. 1.

Рис. 1. Транзакция в сети Биткойн

В указанном примере Алиса отправляет Бобу транзакцию в размере 8 биткойнов (BTC). У данной транзакции имеется два входа (2 и 7 BTC) и два выхода (8 и 1 BTC), причем выход на 1 BTC является, по существу, сдачей и возвращается на кошелек Алисы. Поскольку у каждой транзакции может быть несколько адресов отправки и получения, часто невозможно привязать конкретный адрес отправки к средствам, отправляемым по конкретному адресу получения. Следствием этого наблюдения является то, что нельзя присвоить серийные номера биткойнам и проследить их пути в сети Биткойн.

Процесс ведения транзакции в сети Биткойн основан на механизмах, которые гарантируют, что:

(а) проверка каждой транзакции распределяется между несколькими участниками сети;

(б) запись каждой транзакции дискретизирована по времени, т. е. транзакции линейно упорядочены с последовательными временными отметками;

(в) участники платежной сети соревнуются и получают вознаграждение за запись транзакции в блок;

(г) несколько узлов ведут перекрестную проверку каждой записанной транзакции.

2.1.4. Инициирование транзакции

Допустим, Алиса хочет отправить Бобу 1 биткойн с использованием сети Биткойн. Для этого у Алисы и Боба должны быть биткойн-адреса. Назовем их ID_Алиса и ID_Боб. Затем Алисе нужно выслать и удостоверить (в цифровом виде) подлинность сообщения типа

«ID_Алиса отправляет ID_Боб 1 биткойн».

После того как Алиса подписывает сообщение транзакции с помощью своего закрытого ключа и транслирует его, любой участник сети Биткойн может проверить, что это именно Алиса отправила сообщение, и сообщение не было подделано. К тому же, как указывалось ранее, цифровые подписи гарантируют, что никто другой не мог подписать это сообщение, т. е. Алиса не может отрицать, что его подписала она.

2.1.5. Проверка транзакции

Перед выполнением транзакции протокол Биткойн должен проверить два аспекта сообщения. Во-первых, именно ли Алиса транслировала сообщение транзакции? Схема цифровой подписи гарантирует, что на самом деле только владелец закрытого ключа для данного адреса мог подписать сообщение. Во-вторых, имеется ли на адресе отправки достаточно средств, чтобы гарантировать завершение транзакции?

Хотя ведение учета и проверка транзакций — основные функции всех электронных платежных систем, эти функции, как правило, осуществляются через частные реестры, поддерживаемые доверенными третьими сторонами. Децентрализованные системы, такие как Биткойн, заменяют сторонних посредников и хранящиеся у них записи общественным реестром, который ведется распределенной информационной системой.

2.1.6. Обновление блокчейна

После первоначальной проверки подписанного сообщения транзакции группы участников сети Биткойн начинают состязаться за возможность записать транзакцию в блокчейн. Сначала в блоке транзакций конкурирующие узлы группируют транзакции, которые транслировались с момента последней записи в блокчейне. Затем блок используется для определения сложной для вычисления задачи. Победителем состязания становится узел, который *первым* решит эту задачу. После определения победителя запись транзакции завершается. Узел-победитель имеет право сделать запись и забрать вознаграждение.

Задача, в решении которой состязаются узлы, опирается на одну из описанных выше криптографических схем — хэш-функции. Сначала блок вновь транслируемых транзакций используется в качестве входных данных в криптографической хэш-функции для получения хэша под названием *дайджест*. Этот дайджест вместе с одноразовым случайным кодом *nonce* (буквенно-цифровой строкой) и хэшем предыдущего блока вводятся в другую хэш-функцию, которая выдает хэш блокчейна нового блока. Задача, которую необходимо решить узлам, включает нахождение такого случайного кода, при котором хэш блокчейна нового блока имеет определенные свойства (в данном случае имеет определенное количество начальных нулей). Первый из соревнующихся узлов, который найдет нужный случайный код, транслирует эту информацию остальным участникам сети, а блокчейн обновляется. В данной схеме реализован Hashcash — система *доказательства правильности работы (proof of work)*, целью которой является обеспечить, чтобы компьютеры использовали определенное количество вычислительных ресурсов для выполнения какой-либо задачи, см. Back (2002).

Узлы, которые осуществляют процесс доказательства правильности работы, в сети Биткойн именуются *майнерами*. Эти майнеры стимулируются на использование вычислительных ресурсов в этом процессе путем получения вознаграждения, встроенного в протокол Биткойн. Обычно вознаграждение — это заранее определенное количество вновь создан-

ных биткойнов. Остальная часть вознаграждения, которая в настоящее время гораздо меньше, — это добровольные операционные сборы (комиссии), уплачиваемые инициаторами транзакций майнерам за обработку их транзакций. Первоначальная идея состояла в том, чтобы эти добровольные взносы заменили вознаграждение в виде создания биткойнов для стимулирования майнеров в случаях, когда эта сумма стремится к нулю (Nakamoto, 2008).

2.2. Статистика сети Биткойн

2.2.1. Капитализация сети

На сегодняшний день в обращении находится более 15 млн биткойнов. Цена одного биткойна за 4 месяца 2016 года колебалась в пределах 400–460 долларов США (см. рис. 2). Таким образом, совокупная стоимость всех выпущенных биткойнов составляет около 7 млрд долларов США.

Рис. 2. Динамика стоимости биткойна к доллару США, долл. США за 1 биткойн

Рыночная капитализация биткойна в 10 раз выше, чем у второй по этому показателю криптовалюты — Ethereum. Таким образом, биткойн, очевидно, является безусловным лидером.

Ценность валюты сильно зависит от числа участников, что, в свою очередь, привлекает еще больше участников, приводя в действие сетевой эффект. У биткойна есть существенное преимущество перед другими криптовалютами, которое имеет три аспекта:

- чем больше пользователей, тем более полезным он становится: появляется больше мест, в которых можно потратить биткойны, и контрагентов, с которыми можно обменять биткойны — это привлекает еще больше пользователей;
- валюты требуют доверия, но оно может быть получено только с течением времени, таким образом, при прочих равных условиях, старейшая валюта получает естественное преимущество перед конкурентами;
- чем больше объем, тем больше комиссии за транзакции, это привлекает больше майнеров и делает сеть более безопасной, что, в свою очередь, снова привлекает дополнительных пользователей и объем средств.

Для валют, служащих средством накопления, есть еще одно препятствие: для перевода накоплений из одной валюты в другую требуются дополнительные усилия. Таким образом, действуют многочисленные эффекты, которые делают смещение биткойна с позиции криптовалюты номер один крайне маловероятным.

2.2.2. Мощность сети

Объем вычислительной мощности, поддерживающий сеть, сейчас приближается к 800 петахэшам/с (см. рис. 3)[3]. Оценка совокупной стоимости используемого оборудования превышает 300 млн долларов.

Ежедневная совокупная выручка майнеров от комиссий за запись транзакций и валидацию новых блоков составляет 1.2–1.5 млн долларов США. Число открытых биткойн-

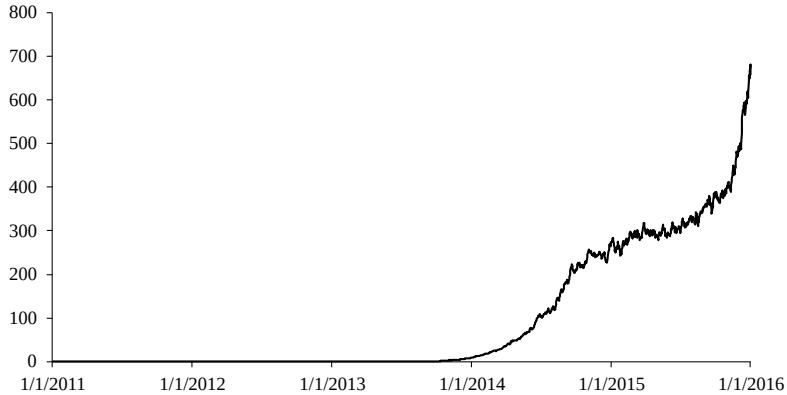


Рис. 3. Динамика вычислительной мощности в сети Биткойн с 01.01.2011 по 31.12.2015, петахэш/с

Рис. 4. Динамика числа транзакций в сети Биткойн с 01.01.2011 по 31.12.2015, тыс. транзакций/день

[3] Петахэш = 1 млн гигахэшей.

кошельков превышает 15 млн. Ежедневно пользователями сети совершается более 200 тысяч транзакций (см. рис. 4).

Динамика транзакций имеет естественный предел, связанный с размером блока (1Мб или около 1500 транзакций). С учетом среднего времени, требуемого на майнинг нового блока (10 минут), теоретический предел Биткойн сети составляет на сегодня 7 транзакций в секунду или около 600 тыс. транзакций в день. Биткойн-сообщество находится в активном поиске оптимального решения по увеличению размера блока. Среди основных инициатив можно назвать программный протокол SegWit, который позволит в 1.5–2 раза поднять пропускную способность, а также LightningNetwork, предполагающий построение сети хабов ликвидности для проведения микроплатежей в защищенных P2P-каналах с возможностью значительного удешевления стоимости.

3. Кто использует биткойн? Эконометрический анализ пользователей биткойна

Биткойн привлек внимание людей по всему миру, и количество статей, посвященных ему, стабильно растет[4]. К сожалению, очень немного исследований посвящено анализу характеристик пользователей Биткойн сети. Такой анализ мог бы обеспечить лучшее понимание этого феномена и его перспектив в будущем. Относительный дефицит академического интереса к данной области не стоит воспринимать как сюрприз, с учетом того, что сбор данных о пользователях биткойна крайне сложен, т. к. они в своем большинстве желают оставаться анонимными. В паре работ авторы попытались преодолеть это препятствие, проведя интервью с десятком пользователей биткойна, см. (Baur et al., 2015; Huhtinen, 2014).

Bohr, Bashir (2014) были первыми, кто проанализировал более крупный структурированный набор данных, полученных на основе опроса, проведенного Lúí Smyth (на тот момент исследователем цифровой антропологии в University College London). Опрос состоит из 1193 ответов, собранных в период с 12 февраля 2013 г. по 4 апреля 2013 г. Bohr и Bashir попытались ответить на три исследовательских вопроса:

1) что предсказывает накопление богатства среди пользователей биткойна;
2) что предсказывает оптимизм относительно краткосрочной и долгосрочной стоимости биткойна;
3) что привлекает людей к биткойну.

Первый вопрос был исследован с помощью простой регрессии заявленного каждым респондентом количества биткойнов в его собственности (логарифмированного по основанию 2 во избежание асимметрии) на набор характеристик пользователей биткойна, извлеченный из опроса:

- возраст пользователя *Age* и возраст, возведенный в квадрат, для учета нелинейности;
- переменная *Installation*, которая отражает период первой загрузки клиента биткойн (программы, которая осуществляет подключение к сети Биткойн), она может принимать значения от 1 (первый квартал 2009 г.) до 17 (первый квартал 2013 г.) и центрирована по среднему значению;

[4] Более подробно см. Böhme et al. (2015) и сайт https://en.bitcoin.it/wiki/Research.

• фиктивная переменная *Miner* для учета того, добывал ли пользователь биткойны самостоятельно;

• переменная взаимодействия *Installation*×*Miner* для проверки наличия преимущества в накоплении биткойнов у ранних майнеров над поздними;

• фиктивная переменная *Bitcoinsins*, которая равна 1, если респондент признал добычу биткойнов через чье-либо оборудование без разрешения владельца (с помощью вредоносных программ) или кражу биткойнов;

• фиктивная переменная *LivesinU.S.,* которая равна 1, если респондент живет в США;

• фиктивная переменная *Illicitgoods*, которая равна 1, если респондент признал покупку с помощью биткойнов наркотиков, услуг по азартным играм или других незаконных товаров;

• фиктивная переменная *Bitcointalk*, которая равна 1, если респондент обозначил, что он использует специализированные для биткойна платформы, чтобы общаться с другими людьми на тему биткойна;

• фиктивная переменная *Investor*, которая равна 1, если респондент описал свою роль в контексте биткойна как инвестора;

• целочисленные переменные *Profit* и *Community*, которые принимают значения от 1 (не мотивирует) до 5 (очень мотивирует) в зависимости от того, считает ли респондент прибыль или сообщество мотивирующими факторами для их изначального присоединения к Биткойну.

Bohr и Bashir (2014) обнаружили, что возраст является статистически значимым фактором для предсказания количества биткойнов, которыми обладал респондент: молодые респонденты обладали меньшим количеством биткойнов, но это количество примерно вдвое увеличивается каждые 10 лет, достигая максимума между 55 и 60 годами, аналогично накоплению других активов. Переменная взаимодействия *Installation*×*Miner* значима, и это подтверждает, что майнинг биткойнов был проще в более ранние периоды действия сети, таким образом, ранние майнеры получили преимущество в накоплении биткойнов. Активные участники онлайн сообществ, посвященных биткойну, обладали вдвое большим количеством биткойнов по сравнению с теми, кто не участвовал в таких сообществах, в то время как определившие себя как *Investor* накопили вчетверо больше биткойнов по сравнению с теми, кто себя так не определил. При прочих равных условиях, превышение накоплений пользователей биткойна, приобретавших запрещенные товары, такие как наркотики, над накоплениями тех, кто покупал только легальные товары, достигало 45%.

Bohr и Bashir (2014) затем представили 2 дополнительные регрессии, в которых краткосрочные (4 месяца от времени опроса) и долгосрочные (6 лет после опроса) ожидаемые стоимости 1 биткойна в долларах США зависели от описанного выше набора данных. Старшие пользователи оказались менее оптимистичными по сравнению с молодыми, а пик оптимизма пришелся на возраст около 35 лет, в то же время, чем выше был уровень социальной вовлеченности в онлайн форумы, тем выше была предсказанная цена. Интересно, что те, кто позже установил биткойн, были более оптимистичны относительно краткосрочной стоимости, а майнеры были более пессимистичны относительно долгосрочной стоимости по сравнению с теми, кто не занимается майнингом биткойнов.

На втором этапе исследования Bohr и Bashir (2014) разделили пользователей на основе их (пользователей) описания биткойна относительно анонимности, свободы и банковской системы, и проанализировали эти три группы с помощью логистической регрессии. Они обнаружили, что политическая идентичность пользователей не являлась фактором, пред-

сказывающим ценность анонимности биткойна. Единственным значимым фактором, определяющим предпочтение биткойна за его анонимность, было осуществление пользователем добычи биткойнов, т. е. являлся пользователь майнером или нет. Пользователи, которые предпочитали биткойн за его потенциал подорвать банковскую систему, оказались старше 40, живущими за пределами США и политически определяющими себя как «зеленые». Наконец, пользователи, которым биткойн нравится за его качества, характеризующие свободу этой сети, политически определяли себя как «либертарианцы», жили за пределами США и были в возрасте от 30 до 39 лет. Интересно, что сами авторы хорошо осведомлены об ограничениях имевшегося набора данных и просят читателя воспринимать их результаты с осторожностью: выборка может не быть репрезентативной с точки зрения всей совокупности пользователей биткойна и принимает во внимание только англоговорящее биткойн-сообщество. Кроме того, опрос является довольно-таки устаревшим, т. к. он проводился до краха биржи Mt. Gox, которая потеряла сотни тысяч биткойнов и сейчас является банкротом. Несмотря на эти ограничения, это исследование, определенно, является перспективным направлением для будущих исследований.

Yelowitz и Wilson (2015) попытались решить проблему маленького набора данных путем использования данных Google Trends для исследования детерминант интереса к биткойну. Google Trends может быть использован для извлечения данных либо по точным условиям поиска, либо по общим темам, в последнем случае учитываются также связанные поисковые запросы. Исследователи построили прокси для четырех возможных классов пользователей биткойна (программисты-энтузиасты, спекулятивные инвесторы, либертарианцы и преступники) и для интереса к биткойну в каждом штате США. В категории «Валюта» они искали данные по общим темам, относящимся к биткойну; в категории «Дисциплина» — запросы на тему «Информатика»; а для оставшихся классов пользователей (преступники, либертарианцы и спекулятивные инвесторы) использовали точные условия поиска «SilkRoad», «Свободный рынок» и «Заработать деньги» соответственно. Отметим, что данные Google Trends представляют, как много веб-поисков было осуществлено по конкретному ключевому слову/словам в заданную неделю и в заданной географической области по отношению к общему числу веб-поисков за ту же неделю и в той же области. Полученный показатель затем масштабируется Google в пределах от 0 до 100 путем его деления на свое максимальное значение и умножения результата на 100. Для каждого штата США Yelowitz и Wilson (2015) изначально вычислили временной ряд с 31 наблюдением (от января 2011 г. до июля 2013 г.) относительной популярности биткойна для каждой группы клиентов. Затем они использовали Google Trends, чтобы измерить относительный уровень популярности в штате для каждого поискового условия для всего периода и масштабировали каждый ряд для штата относительно самого популярного штата. Этот тип анализа имеет два ограничения: Google дополняет выборку в своей базе данных каждый раз, когда выполняется запрос, поэтому точное копирование невозможно, хотя качественные результаты не изменяются (см. также раздел 4.4 в (Fantazzini, Toktamysova, 2015) для обсуждения данной проблемы); Google Trends подставляет значение 0, если число поисков слишком мало[5]. Из 1488 (48 штатов × 31 месяц) потенциальных наблюдений Yelowitz и Wilson (2015) использовали 794 с ненулевыми значениями. Вслед за Stephens-Davidowitz (2014) они нормализовали каждый уровень поиска к его z-баллу и оценили следующую панельную регрессию:

[5] https://support.google.com/trends/answer/4355213?hl=en&ref_topic=4365599.

$$BITCOIN_{jt} = \beta_0 + \beta_1 X_{jt} + \delta_j + \delta_t + \varepsilon_{jt},$$

где $BITCOIN_{jt}$ — интерес к биткойну в штате j в месяц t, X_{jt} — интерес клиентуры, а δ_j и δ_t — фиксированные эффекты штата и времени. Каждый штато-месяц взвешен по населению штата (в июле 2011 года), а стандартные ошибки скорректированы на невложенную двунаправленную кластеризацию по уровням штата и времени, см. (Cameron et al., 2011) для подробностей. Yelowitz и Wilson (2015) применили большой набор спецификаций, постепенно включая дополнительные регулирующие переменные для штата и времени, такие как уровень безработицы и несвязанные «плацебо-переменные», условия взаимодействия исходных переменных с ценами биткойна. Более того, некоторые спецификации были оценены с использованием данных от 2012 года и позже (когда биткойн был более популярен) или для 24 штатов США, для которых было как минимум 20 месячных наблюдений. Во всех случаях они обнаружили положительные связи между интересом к биткойну и двумя группами клиентов — программистами-энтузиастами и теми, кто, возможно, вовлечен в нелегальную деятельность — в то время как никакой значимой связи с теми, кто интересуется либертарианской идеологией или руководствуется инвестиционными мотивами, обнаружено не было.

Хотя работа Yelowitz и Wilson (2015) решила некоторые проблемы анализа в (Bohr, Bashir, 2014), но она относится только к американскому сообществу биткойна, а данные собраны до банкротства биржи Mt. Gox. Тем не менее, в ней предложены некоторые идеи по использованию данных Google Trends, которые впоследствии (в разделе 5) будут включены в более сложные модели, предложенные для моделирования динамики цены биткойна.

4. Сколько стоит биткойн? Обзор финансовых и экономических подходов

Стоимость биткойна была крайне подвержена волатильности в течение последних нескольких лет, в связи с чем был поднят вопрос, не является ли он пузырем. Одним из способов ответа на этот вопрос является проведение тестов на определение финансовых пузырей, которые будут рассмотрены в разделе 6. Другой вариант — попробовать оценить внутреннюю (фундаментальную) стоимость валюты. В этом направлении к настоящему времени были предложены два подхода, базирующиеся на определении емкости рынка и предельных издержек производства, зависящих от потребления электроэнергии.

4.1. Верхняя граница: оценка емкости рынка

Метод, основанный на оценке емкости рынка, широко используется компаниями, намеревающимися запустить новый продукт или услугу. Данный подход может быть использован и для получения ориентира оценки справедливой стоимости биткойна.

Woo et al. (2013) в отчете Банка Америки Мерилл Линч оценили отдельно стоимость биткойна как средства обмена и как средства накопления и затем просуммировали их, чтобы получить грубую оценку справедливой стоимости биткойна. Для того чтобы вычислить стоимость биткойна как средства обмена, они рассмотрели два варианта его использования: *электронную коммерцию* и денежные переводы. В первом варианте авторы оценили скорость

обращения денег в экономике США, поделив расходы на личное потребление (C_{US}) на размер наличных денежных средств и чековых депозитов домашних хозяйств (HD_{US}). Полученная скорость обращения денег, усредненная за последние 10 лет, умножалась далее на совокупный объем продаж в розничной электронной торговле за предыдущий год в предположении, что для онлайн-продаж и расходов домашних хозяйств США скорость обращения в целом одинакова. Woo et al. (2013) предположили, что биткойн станет средством платежа в 10% всех онлайн-покупок ($Bitcoin_{share}$), что дает оценку необходимого объема биткойнов в обращении на сумму 1 млрд долларов США. Наконец, с учетом того, что ВВП США составляет около 20% мирового ВВП, была получена оценка совокупной стоимости биткойнов в 5 млрд долларов, необходимых для онлайн-покупок. Полученная в результате формула имеет вид:

$$V_{e-commerce_t} = \frac{1}{10}\left(\sum_{i=1}^{10} \frac{C_{US_{t-i}}}{HD_{US_{t-i}}}\right) \cdot \mathrm{B2C}_{t\text{-}1} \cdot Bitcoin_{share} \cdot \frac{GDP_{world_{t-1}}}{GDP_{US_{t-1}}}.$$

Woo et al. (2013) подчеркнули, что, помимо роли биткойна как средства платежа в онлайн покупках, возможно его использование для *денежных переводов*. Они рассмотрели трех основных игроков в индустрии денежных переводов: Western Union, MoneyGram и Euronet, обладающих примерно 20% рынка, и предположили, что биткойн мог бы стать одним из трех крупнейших игроков в этой индустрии. Затем, в предположении, что капитализация биткойна является эквивалентом стоимости компании, они добавили среднюю рыночную капитализацию Western Union, MoneyGram и Euronet (приблизительно 4.5 млрд долларов) к максимальной рыночной капитализации биткойнов как средства обмена:

$$V_{money\,transfer_t} = \frac{1}{3}\left(MK_{WU_t} + MK_{MG_t} + MK_{E_t}\right).$$

Woo et al. (2013) предположили, что активами, наиболее близкими к биткойну как к средству накопления, являются, по всей видимости, драгоценные металлы и наличные денежные средства. В частности, у биткойна и золота есть три общих характеристики: 1) они не подразумевают уплату процента, 2) предложение обоих ограничено, и 3) их сложнее отслеживать, чем большинство финансовых активов (кроме наличных денег). Принимая во внимание то, что общая стоимость выпущенных золотых слитков/монет/ETF в 2013 году составляла приблизительно 1.3 трлн долларов, и что биткойн по сравнению с золотом гораздо более волатилен, Woo et al. (2013) предположили, что рыночная капитализация биткойна не может превзойти 300 млрд долларов. Более того, допуская, что биткойн со временем может приобрести репутацию серебра, и цена золота в 2013 году была примерно в 60 раз выше цены на серебро, они предположили, что рыночная капитализация биткойна как средства накопления могла бы достичь 5 млрд долларов. Примечательно, что они обратили внимание на то, что эта величина близка к стоимости всех отчеканенных с 1986 г. серебряных орлов США (около 8 млрд долларов — 12 тысяч тонн). Следовательно, простой грубый способ получить рыночную капитализацию биткойна как средства накопления:

$$V_{store\ of\ value_t} = 0.6 \cdot TSM_t \cdot P_{silver,t},$$

где TSM_t — общий объем отчеканенных с 1986 г. до момента t серебряных орлов США, а $P_{silver,\,t}$ — цена 1 тройской унции серебра в момент времени t.

Наконец, Woo et al. (2013) рассчитали потенциальную справедливую стоимость биткойна как сумму максимальной рыночной капитализации биткойна как средства обмена и накопления, деленную на общее число биткойнов в обращении (TB_t), таким образом, получая максимальную справедливую стоимость биткойна приблизительно равной 1300 долларам:

$$P_{bitcoin_t} = \frac{V_{e-commerce_t} + V_{money\,transfer_t} + V_{store\,of\,value_t}}{TB_t}.$$

Другой подход к определению размера рынка применили Bergstra, de Leeuw (2013), которые, следуя идее, предложенной Yermack (2013), сравнили биткойн с высокотехнологичным стартапом, который либо будет доминировать на рынке, либо провалится. Они предположили, что если биткойн будет успешным и просуществует до 2040 года, то он будет представлять половину всех денег в мире. С учетом технической новизны системы биткойна они присвоили данному событию очень низкую вероятность (p): один шанс из 100 000. Предполагая, что общая денежная масса (MM) в 2040 году будет 10^{14} евро (всего лишь предположение!), они оценивают биткойн в 50 евро:

$$P_{bitcoin_t} = \frac{MM_{2040}}{TB_{2040}} \cdot p_t = \frac{10^{14}}{2 \cdot 10^7} \cdot 10^{-5} = 50 \text{ euro}.$$

Аналогичный подход также исследуется Huhtinen (2014), рассмотревшим текущие денежные агрегаты M2 для доллара США, евро и японской иены и альтернативные сценарии, касающиеся доли денежного предложения, которое могло бы быть замещено биткойном. Он утверждает, что наиболее реалистичным уровнем замещения для трех мировых валют является уровень 0.1%, который мог бы быть достигнут при оценке стоимости биткойна в 1573 евро.

4.2. Нижняя граница: предельная стоимость производства биткойна

Оценка размера рынка может дать представление о потенциале биткойна в долгосрочном периоде, но данный подход определенно неудовлетворителен для объяснения краткосрочной динамики цены биткойна. В связи с этим Garcia et al. (2014) первыми предположили, что фундаментальная стоимость биткойна должна быть как минимум равна стоимости энергии, затраченной на его производство путем *майнинга* (добычи), и что эта стоимость должна использоваться в качестве нижней границы оценки фундаментальной стоимости биткойна. А именно, они разделили накопленную вычислительную мощность добычи в день на число добытых биткойнов, чтобы получить число хэшей (по алгоритму SHA-256), необходимых для майнинга одного биткойна. Затем они использовали приближение энергетических потребностей для добычи в размере 500 ватт за гигахэш/секунду (гх/с, GH/s) — это была средняя эффективность большинства графических процессоров, использованных для майнинга биткойнов в 2010–2013 годах[6], и примерные издержки на электричество в размере 0.15 долларов за квт·час, что являлось средним значением американских и британских цен.

[6] На конец 2015 года потребности гораздо ниже.

Позднее в работах Hayes (2015, 2016) была разработана более изящная модель стоимости производства биткойна, которая детально обсуждается ниже. Hayes обращает внимание на то, что рациональные агенты не стали бы заниматься майнингом биткойнов, если из-за этого они бы понесли реальные убытки, и переменными, влияющими на решение о майнинге, по-существу, являются 5 показателей:

1) стоимость электричества в центах за киловатт·час;

2) потребление энергии на добывающую единицу в ваттах на гигахэш/секунду (1 вт/гигахэш/сек = 1 джоуль/гигахэш), которое является функцией стоимости электричества и энергетической эффективности;

3) рыночная цена биткойна;

4) сложность алгоритма биткойна;

5) награда за блок (на текущий момент — 25 биткойнов), которая уменьшается вдвое примерно каждые 4 года.

На конкурентном товарном рынке агент предпринял бы майнинг, если предельные издержки в день (потребление электричества) были бы меньше или равны предельному продукту (среднее количество добытых в день биткойнов, умноженное на цену биткойна в долларах). Hayes (2015, 2016) утверждает, что спекулятивные свойства биткойна, а также свойства, уподобляющие его деньгам (т. е. как средства обмена и накопления) могут, несомненно, добавлять порцию субъективности в любые объективные попытки оценить внутреннюю стоимость биткойна. Как бы то ни было, предельная стоимость производства, определенная потреблением энергии, может служить нижней границей величины, вблизи которой майнеры будут принимать решение о производстве.

Hayes (2015, 2016) развивает свою модель, предполагая, что дневное производство майнером биткойнов зависит от его собственного уровня доходности, измеренного в ожидаемых биткойнах в день на единицу добывающих мощностей.

Ожидаемое количество биткойнов, которые будут добыты за день, может быть рассчитано следующим образом:

$$\text{BTC/day*} = [(\beta \cdot \rho)/(\delta \cdot 2^{32})] \cdot \text{sec}_{\text{hr}} \cdot \text{hr}_{\text{day}}, \tag{1}$$

где β — награда за блок (на текущий момент — 25 биткойнов/блок), ρ — мощность хеширования, используемая майнером, δ — сложность (которая выражается в единицах гигахэшей/блок). Константа sec_{hr} — количество секунд в часе (3600), hr_{day} — количество часов в дне (24). Константа 2^{32} относится к нормализованной вероятности решения блока отдельно взятым хэшем в секунду и определяется сложностью задачи 256-битного шифрования в ядре алгоритма SHA-256, которую майнеры пытаются решить. Эти константы, нормализующие пространство для дневной периодичности и алгоритма добычи, могут быть записаны в виде одной переменной $\theta = \text{sec}_{\text{hr}} \cdot \text{hr}_{\text{day}} / 2^{32} = 0.00002011656761169 43$. Уравнение (1), таким образом, может быть переписано в более компактном виде:

$$\text{BTC/day*} = \theta \cdot (\beta \cdot \rho)/\delta. \tag{2}$$

Hayes (2015, 2016) устанавливает ρ=1000 гх/с, несмотря на то что реальная мощность хеширования майнера, скорее всего, значительно отклоняется от этой величины. Как бы то ни было, Hayes (2015, 2016) утверждает, что этот уровень обычно является хорошим эта-

лоном измерения при текущих обстоятельствах. Стоимость майнинга в день E_{day} может быть выражена следующим образом:

$$E_{day} = (\text{price_per_kWh}) \cdot 24 \cdot (\text{W_per_GH/s}) \cdot (\rho \,/\, 1000 \text{ GH/s}). \quad (3)$$

В предположении, что рынок биткойнов конкурентен, предельный продукт майнинга должен быть равным предельным издержкам, так чтобы равновесная цена (долларов/биткойн) задавалась коэффициентом «издержки в день / биткойнов в день»:

$$p^* = E_{day} \,/\, (\text{BTC/day}^*). \quad (4)$$

Эта цена может, однако, быть нижней границей для цены, ниже которой майнер будет работать с убытками и, вероятно, прекратит производство. Наоборот, имея рыночную цену биткойна, можно инвертировать уравнение (4), чтобы найти нижние границы (точки безубыточности, как определено Hayes (2015, 2016)) для других переменных, определяющих прибыльность биткойна. Например, при наличии наблюденной рыночной цены (p) и сложности майнинга, точка безубыточности для издержек на электричество (в квт·часах) определяется формулой:

$$\text{price_per_kWh}^* = [p(\text{BTC/day}^*)/24] \,/\, \text{W_per_GH/s}. \quad (5)$$

Аналогично, имея известную стоимость производства и наблюденную рыночную цену, можно решить уравнение для точки безубыточности для сложности майнинга:

$$\delta^* = (\beta \cdot \rho \cdot \text{sec}_{hr} \cdot \text{hr}_{day}) \,/\, [(E_{day}/p) \cdot 2^{32}]. \quad (6)$$

Наконец, имея рыночную цену, издержки на электричество (в квт·часах) и сложность добычи, можно найти точку безубыточности для энергетической эффективности:

$$\text{W_per_GH/s}^* = p(\text{BTC/day}^*)/(\text{price_per_kWh} \cdot 24). \quad (7)$$

Уравнение (4) показывает, что если практическая эффективность майнинга увеличится (что и ожидается многими ввиду более эффективного оборудования), то безубыточная цена для производителей биткойна будет иметь тенденцию к снижению. Например, Garcia et al. (2014) обнаружили, что средняя эффективность майнинга за период с 2010 по 2013 г. составила приблизительно 500 ватт на гх/с. В то же время, используя уравнение (7) или взглянув на лучшее доступное оборудование для майнинга[7], можно увидеть, что на текущий момент средняя эффективность окажется близкой к 0.50–0.90 ватт на гх/с. Более того, уравнения (2) и (4) показывают, что меньшая награда за блок β, при неизменных остальных параметрах, повысит цену биткойна. С учетом того, что в 2016 г. ожидается снижение награды за блок вдвое, т. е. до 12.5 биткойнов, если цена биткойна не вырастет, это будет свидетельствовать о том, что энергетическая эффективность майнинга должна будет скомпенсировать сниженную награду за блок.

[7] https://en.bitcoin.it/wiki/Mining_hardware_comparison.

Приведем небольшой числовой пример. Предположим, что средняя мировая стоимость электричества составляет приблизительно 13.5 центов за киловатт·час (как в (Hayes, 2016)) и средняя энергетическая эффективность оборудования для майнинга составляет 0.75 дж/гх. Тогда, следуя формуле (3), средние издержки в день на оборудование мощностью 1000 гх/с составили бы примерно $(0.135 \cdot 24 \cdot 0.75) \cdot (1000 / 1000) = 2.43$ доллара в день. Количество биткойнов, которое оборудование с мощностью 1000 ГХ/с может добыть в день, с текущей сложностью 60 883 825 480 равно 0.0082602265269544 биткойна в день. Учитывая, что предельные издержки теоретически должны быть равны предельному продукту, получим цену биткойна из уравнения (4):

(2.43 доллара в день) / (0.0082602265269544 биткойна в день) ≈ $ 294.18/BTC,

что незначительно отличается от рыночного курса в 290–300 долларов/биткойн на момент написания настоящей статьи (декабрь 2015 года).

Интересно, что при прочих равных условиях при уменьшении награды за блок вдвое до 12.5 биткойнов[8], справедливая цена биткойна должна быть в районе $ 588.36/BTC. Увидим ли мы дальнейший рост цены в ближайшее время?

Список литературы

Ali R., Barrdear J., Clews R., Southgate J. (2014). The economics of digital currencies. *Bank of England Quarterly Bulletin*, Q3, 276–287.

Allen H. J. (2016). $=€=Bitcoin? *Suffolk University Legal Studies Research Paper Series, Research Paper*, No. 15–33.

Back A. (2002). Hashcash — a denial of service counter-measure. http://www.hashcash.org/papers/hashcash.pdf.

Badev A., Chen M. (2014). Bitcoin: Technical background and data analysis. *Finance and Economics Discussion Series* 2014–104. Divisions of Research and Statistics and Monetary Affairs, Federal Reserve Board, Washington, D.C.

Baur A. W., Bühler J., Bick M., Bonorden C. S. (2015). Cryptocurrencies as a disruption? Empirical findings on user adoption and future potential of Bitcoin and Co. In: *Open and Big Data Management and Innovation*, 63–80. Springer International Publishing.

Becker J., Breuker D., Heide T., Holler J., Rauer H. P., Böhme R. (2013). Can we afford integrity by proof-of-work? Scenarios inspired by the Bitcoin currency. In: *The Economics of Information Security and Privacy*, 135–156. Springer.

Bergstra J. A., de Leeuw K. (2013). Bitcoin and beyond: Exclusively informational monies. *arXiv preprint*, arXiv: 1304.4758.

Böhme R., Christin N., Edelman B., Moore T. (2015). Bitcoin: Economics, technology, and governance. *The Journal of Economic Perspectives*, 29 (2), 213–238.

Bohr J., Bashir M. (2014). Who uses bitcoin? An exploration of the bitcoin community. In: *Privacy, Security and Trust (PST)*, *2014 Twelfth Annual International Conference*, 94–101.

Cameron A. C., Gelbach J. B., Miller D. L. (2011). Robust inference with multiway clustering. *Journal of Business and Economic Statistics*, 29 (2), 238–249.

[8] Ожидаемая дата изменения награды за блок — 23.07.2016, см. http://www.bitcoinblockhalf.com/.

Dang Q. (2012). Recommendation for applications using approved hash algorithms. *Technical Report*, National Institute of Standards and Technology.

Dwyer G. P. (2015). The economics of Bitcoin and similar private digital currencies. *Journal of Financial Stability*, 17, 81–91.

ECB. (2012). Virtual currency schemes. *ECB report*, October 2012. https://www.ecb.europa.eu/pub/pdf/other/virtualcurrencyschemes201210en.pdf.

ECB. (2015). Virtual currency schemes — a further analysis. *ECB report*, February 2015. https://www.ecb.europa.eu/pub/pdf/other/virtualcurrencyschemesen.pdf.

Fantazzini D., Toktamysova Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics*, 170, 97–135.

Garcia D., Tessone C. J., Mavrodiev P., Perony N. (2014). The digital traces of bubbles: Feedback cycles between socio-economic signals in the Bitcoin economy. *Journal of the Royal Society Interface*, 11 (99), 20140623.

Hayes A. (2015). A cost of production model for bitcoin. *Working Paper* 05/2015. Department of Economics, The New School for Social Research.

Hayes A. (2016). Cryptocurrency value formation: An empirical analysis leading to a cost of production model for valuing bitcoin. *Telematics and Informatics*, dx.doi.org/10.1016/j.tele.2016.05.005.

Huhtinen T. P. (2014). Bitcoin as a monetary system: Examining attention and attendance. Master's thesis. Department of Finance, Aalto University School of Business. http://epub.lib.aalto.fi/fi/ethesis/pdf/13626/hse_ethesis_13626.pdf.

Kroll J. A., Davey I. C., Felten E. W. (2013). The economics of Bitcoin mining, or Bitcoin in the presence of adversaries. In: *Proceedings of WEIS 2013*.

Lo S., Wang J. C. (2014). Bitcoin as money? *Current Policy Perspectives*, 14–4, Federal Reserve Bank of Boston.

Murphy E. V., Murphy M. M., Seitzinger M. V. (2015). *Bitcoin: questions, answers, and analysis of legal issues*. US Congressional Research Service 7–5700, R43339. https://www.fas.org/sgp/crs/misc/R43339.pdf.

Nakamoto S. (2008). Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf.

Rogojanu A., Badea L. (2014). The issue of competing currencies. Case study — Bitcoin. *Theoretical and Applied Economics*, 21 (1), 103–114.

Segendorf B. (2014). What is Bitcoin. *Sveriges Riksbank Economic Review*, 2, 71–87.

Stephens-Davidowitz S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics*, 118, 26–40.

Velde F. (2013). *Bitcoin: A primer*. Chicago Fed Letter, 317, December.

Woo D., Gordon I., Iaralov V. (2013). Bitcoin: A first assessment. *FX and Rates,* December 2013, Bank of America Merrill Lynch.

Yelowitz A., Wilson, M. (2015). Characteristics of bitcoin users: An analysis of Google search data. *Applied Economics Letters*, 22 (13), 1030–1036.

Yermack D. (2013). Is bitcoin a real currency? An economic appraisal. *National Bureau of Economic Research working paper* No 19747.

Поступила в редакцию 01.06.2016;
принята в печать 25.09.2016.

Fantazzini D., Nigmatullin E., Sukhanovskaya V., Ivliev S. *Applied Econometrics*, 2016, v. 44, pp. 5–24.

Dean Fantazzini
Moscow School of Economics, Moscow State University, Russian Federation; fantazzini@mse-msu.ru

Erik Nigmatullin
Bocconi University, Milan, Italy; nigmatullin.erik@gmail.com

Vera Sukhanovskaya
Perm State National Research University, Russian Federation; vera-sukhanovskaya@yandex.ru

Sergey Ivliev
Perm State National Research University, Russian Federation; ivliev@gmail.com

Everything you always wanted to know about bitcoin modelling but were afraid to ask. Part I

Bitcoin is an open source decentralized digital currency and a payment system. It has raised a lot of attention and interest worldwide and an increasing number of articles are devoted to its operation, economics and financial viability. This article reviews the econometric and mathematical tools which have been proposed so far to model the bitcoin price and several related issues, highlighting advantages and limits. We discuss the methods employed to determine the main characteristics of bitcoin users, the models proposed to assess the bitcoin fundamental value, the econometric approaches suggested to model bitcoin price dynamics, the tests used for detecting the existence of financial bubbles in bitcoin prices and the methodologies suggested to study the price discovery at bitcoin exchanges.

Keywords: crypto-currencies; hash rate; investors' attractiveness; social interactions; money supply; money demand; speculation; forecasting; algorithmic trading; bubble; price discovery.

JEL classification: C22; C32; C51; C53; E41; E42; E47; E51; G17.

References

Ali R., Barrdear J., Clews R., Southgate J. (2014). The economics of digital currencies. *Bank of England Quarterly Bulletin*, Q3, 276–287.

Allen H. J. (2016). $=€=Bitcoin? *Suffolk University Legal Studies Research Paper Series, Research Paper*, No. 15–33.

Back A. (2002). Hashcash — a denial of service counter-measure. http://www.hashcash.org/papers/hashcash.pdf.

Badev A., Chen M. (2014). Bitcoin: Technical background and data analysis. *Finance and Economics Discussion Series* 2014–104. Divisions of Research and Statistics and Monetary Affairs, Federal Reserve Board, Washington, D.C.

Baur A. W., Bühler J., Bick M., Bonorden C. S. (2015). Cryptocurrencies as a disruption? Empirical findings on user adoption and future potential of Bitcoin and Co. In: *Open and Big Data Management and Innovation*, 63–80. Springer International Publishing.

Becker J., Breuker D., Heide T., Holler J., Rauer H. P., Böhme R. (2013). Can we afford integrity by proof-of-work? Scenarios inspired by the Bitcoin currency. In: *The Economics of Information Security and Privacy*, 135–156. Springer.

Bergstra J. A., de Leeuw K. (2013). Bitcoin and beyond: Exclusively informational monies. *arXiv preprint*, arXiv: 1304.4758.

Böhme R., Christin N., Edelman B., Moore T. (2015). Bitcoin: Economics, technology, and governance. *The Journal of Economic Perspectives*, 29 (2), 213–238.

Bohr J., Bashir M. (2014). Who uses bitcoin? An exploration of the bitcoin community. In: *Privacy, Security and Trust (PST)*, *2014 Twelfth Annual International Conference*, 94–101.

Cameron A. C., Gelbach J. B., Miller D. L. (2011). Robust inference with multiway clustering. *Journal of Business and Economic Statistics*, 29 (2), 238–249.

Dang Q. (2012). Recommendation for applications using approved hash algorithms. *Technical Report*, National Institute of Standards and Technology.

Dwyer G. P. (2015). The economics of Bitcoin and similar private digital currencies. *Journal of Financial Stability*, 17, 81–91.

ECB. (2012). Virtual currency schemes. *ECB report*, October 2012. https://www.ecb.europa.eu/pub/pdf/other/virtualcurrencyschemes201210en.pdf.

ECB. (2015). Virtual currency schemes — a further analysis. *ECB report*, February 2015. https://www.ecb.europa.eu/pub/pdf/other/virtualcurrencyschemesen.pdf.

Fantazzini D., Toktamysova Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics*, 170, 97–135.

Garcia D., Tessone C. J., Mavrodiev P., Perony N. (2014). The digital traces of bubbles: Feedback cycles between socio-economic signals in the Bitcoin economy. *Journal of the Royal Society Interface*, 11 (99), 20140623.

Hayes A. (2015). A cost of production model for bitcoin. *Working Paper* 05/2015. Department of Economics, The New School for Social Research.

Hayes A. (2016). Cryptocurrency value formation: An empirical analysis leading to a cost of production model for valuing bitcoin. *Telematics and Informatics*, dx.doi.org/10.1016/j.tele.2016.05.005.

Huhtinen T. P. (2014). Bitcoin as a monetary system: Examining attention and attendance. Master's thesis. Department of Finance, Aalto University School of Business. http://epub.lib.aalto.fi/fi/ethesis/pdf/13626/hse_ethesis_13626.pdf.

Kroll J. A., Davey I. C., Felten E. W. (2013). The economics of Bitcoin mining, or Bitcoin in the presence of adversaries. In: *Proceedings of WEIS 2013*.

Lo S., Wang J. C. (2014). Bitcoin as money? *Current Policy Perspectives*, 14–4, Federal Reserve Bank of Boston.

Murphy E. V., Murphy M. M., Seitzinger M. V. (2015). *Bitcoin: questions, answers, and analysis of legal issues*. US Congressional Research Service 7–5700, R43339. https://www.fas.org/sgp/crs/misc/R43339.pdf.

Nakamoto S. (2008). Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf.

Rogojanu A., Badea L. (2014). The issue of competing currencies. Case study — Bitcoin. *Theoretical and Applied Economics*, 21 (1), 103–114.

Segendorf B. (2014). What is Bitcoin. *Sveriges Riksbank Economic Review*, 2, 71–87.

Stephens-Davidowitz S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics*, 118, 26–40.

Velde F. (2013). *Bitcoin: A primer*. Chicago Fed Letter, 317, December.

Woo D., Gordon I., Iaralov V. (2013). Bitcoin: A first assessment. *FX and Rates,* December 2013, Bank of America Merrill Lynch.

Yelowitz A., Wilson, M. (2015). Characteristics of bitcoin users: An analysis of Google search data. *Applied Economics Letters*, 22 (13), 1030–1036.

Yermack D. (2013). Is bitcoin a real currency? An economic appraisal. *National Bureau of Economic Research working paper* No 19747.

Received 01.06.2016; accepted 25.09.2016.

Appendix 15 "Everything you always wanted to know about bitcoin modelling but were afraid to ask - Part 2."

Fantazzini, D., Nigmatullin, E., Sukhanovskaya, V., & Ivliev, S. (2017). Everything you always wanted to know about bitcoin modelling but were afraid to ask - Part 2. *Applied Econometrics* 45, 5-28. http://pe.cemi.rssi.ru/pe_2017_45_005-028.pdf

Permisson to post the Publisher's version/PDF: according to http://appliedeconometrics.cemi.rssi.ru/AppEc.files/contents_en.html, "*full texts are available for downloading with approximately one year lag*". In some cases, this 1-year lag can be shorter.

Прикладная эконометрика, 2017, т. 45, с. 5–28.
Applied Econometrics, 2017, v. 45, pp. 5–28.

Д. Фантаццини, Э. М. Нигматуллин, В. Н. Сухановская, С. В. Ивлиев[1]

Все, что вы хотели знать о моделировании биткойна, но боялись спросить. Часть 2[2]

Вторая часть данной статьи завершает серию консультационных публикаций о биткойне. В частности, рассматриваются эконометрические подходы к моделированию динамики цены биткойна, тесты, применяемые для выявления финансовых пузырей в ценах, и методологии, предложенные для изучения ценообразования на биткойн-биржах.

Ключевые слова: биткойн; криптовалюты; хэшрейт; майнинг; привлекательность инвесторов; социальное взаимодействие; денежное предложение; спрос на деньги; спекуляции; прогнозирование; алгоритмическая торговля; пузыри; ценообразование; LPPL.

JEL classification: C22; C32; C51; C53; E41; E42; E47; E51; G17.

5. Моделирование динамики цены биткойна

Практически все исследования, посвященные стоимости биткойна, используют анализ временных рядов. В небольшом числе работ используются простые кросс-секционные регрессии, которые вполне могут быть пригодными: криптовалюты — явление очень недавнее, они высоко спекулятивны и волатильны, из-за чего методы оценки временных рядов могут давать обманчивые и неинформативные результаты с учетом того, что временной интервал исследуемых данных невелик (Hayes, 2015b). Ради общности рассматриваются оба подхода.

[1] **Фантаццини Деан** — Московская школа экономики, МГУ, Москва; fantazzini@mse-msu.ru.
Нигматуллин Эрик Маликович — Bocconi University, Милан, Италия; nigmatullin.erik@gmail.com.
Сухановская Вера Николаевна — Пермский государственный национальный исследовательский университет; Пермь; vera-sukhanovskaya@yandex.ru.
Ивлиев Сергей Владимирович — Пермский государственный национальный исследовательский университет, Пермь; ivliev@gmail.com.

[2] Часть 1 данной статьи была опубликована в журнале *Прикладная эконометрика*, 2016, т. 44, 5–24. Сохранена сквозная нумерация разделов и формул, начатая в первой части.

5.1. Эконометрический анализ на основе кросс-секционных данных

Hayes (2015b) представил регрессию с использованием перекрестного набора данных, состоящего из 66 торгуемых криптовалют (так называемых *альткойнов*), основанную на теоретической модели, развитой в работе (Hayes, 2015c). Была построена регрессия натурального логарифма рыночных цен альткойнов от 18 сентября 2014 года, выраженных в BTC, на набор из 5 переменных:

- натуральный логарифм *вычислительной мощности в Гигахешах в секунду*;
- натуральный логарифм *числа (альт-) койнов, добытых в минуту*, рассчитанный на основе деления вознаграждения за каждый добытый блок на время между блоками;
- *процент монет, добытых к текущему моменту*, относительно общего числа монет, доступных для добычи;
- фиктивная переменная, обозначающая примененный *вычислительный алгоритм* (0 для SHA-256 и 1 для Scrypt);
- *число календарных дней* от создания альткойна до 18 сентября 2014 года.

Hayes (2015b) обнаружил, что чем выше вычислительная мощность, используемая для добычи криптовалюты, тем выше ее (валюты) цена. Это — вполне ожидаемый результат с учетом того, что объем вычислительной мощности является прокси для суммарного применения рассматриваемых альткойнов. Более того, рациональный майнер будет стремиться использовать ресурсы для майнинга только в том случае, если предельная цена добычи превышает предельные издержки. Hayes (2015b) также обнаружил, что число монет, добытых в минуту, отрицательно коррелирует с ценой альткойна, что ожидаемо с учетом того, что дефицит на добытый блок имеет более высокую воспринимаемую ценность. Другой интересный результат заключается в том, что альткойны, основанные на алгоритме Scrypt, имеют более высокую стоимость, чем основанные на SHA-256, при прочих равных условиях. Алгоритм Scrypt был предложен для предотвращения попыток использования специализированного аппаратного обеспечения для того, чтобы обойти других майнеров. Альткойн на основе алгоритма Scrypt предполагает более высокие требования к вычислительным ресурсам на добычу единицы криптовалюты, чем эквивалентный альткойн, использующий алгоритм SHA-256. Hayes (2015b) выяснил, что процент альткойнов, добытых к текущему моменту, относительно альткойнов, оставшихся для добычи, не имеет статистического влияния на цену альткойна. Он утверждает это на основании того факта, что альткойны делимы до 8 знаков после десятичной запятой, и это число знаков потенциально может быть увеличено без ограничения. Однако, по нашему мнению, наиболее вероятной причиной является возможность увеличения объема предложения альткойнов при условии согласия майнеров. Hayes (2015b) также установил, что продолжительность существования криптовалюты не связана с ценой альткойна, что может объясняться коротким рассматриваемым временным периодом (подавляющее большинство альткойнов существует менее 2 лет).

В целом данные результаты могут представлять большой интерес для тех, кто хочет создать успешный альткойн: необходимым условием представляется использование алгоритма Scrypt (или другого, еще более сложного протокола) и удержание количества монет, добываемых в минуту, на относительно низком уровне, что может быть достигнуто увеличением времени, необходимого на добычу одного блока, или снижением награды за успешно добытый новый блок. Напротив, увеличение вычислительной мощности, используемой в добыче альткойнов, является более сложным и, как правило, неконтролируемым фактором

по отношению к создателю альткойна, только если не были вложены очень крупные (и дорогостоящие) инвестиции в ИТ-инфраструктуру альткойна.

5.2. Эконометрический анализ на основе временных рядов

Kristoufek (2013) первым предложил многомерный подход, сконцентрированный на спекулятивной составляющей стоимости биткойна, и показал, что и фазы пузырей, и фазы спадов стоимости биткойна могут быть частично объяснены интересом инвесторов к криптовалюте. В работе (Kristoufek, 2013) в качестве прокси для измерения интереса инвесторов было предложено использовать количество поисковых запросов в GoogleTrends и количество просмотров статьи о биткойне в Википедии.

Kristoufek (2013) применил двумерную векторную авторегрессию для недельных лог-приростов цены биткойна и данных GoogleTrends

$$\Delta Y_{t-1} = \alpha + \Phi_1 \Delta Y_{t-1} + \Phi_2 \Delta Y_{t-2} + \ldots + \Phi_p \Delta Y_{t-p} + \varepsilon_t \tag{8}$$

и двумерную векторную модель коррекции ошибок для дневных лог-цен на биткойн и данных по поискам в Wikipedia

$$\Delta Y_{t-1} = \alpha + B\Gamma Y_{t-1} + \zeta_1 \Delta Y_{t-1} + \zeta_2 \Delta Y_{t-2} + \ldots + \zeta_{p-1} \Delta Y_{t-(p-1)} + \varepsilon_t\,, \tag{9}$$

где B — факторные нагрузки, а Г — коинтеграционный вектор. Более того, Kristoufek (2013) применил трехмерную векторную модель коррекции ошибок для лог-цен на биткойн и двух переменных Q_t^+ и Q_t^-, измеряющих положительный и отрицательный отклик соответственно:

$$Q_t^+ = Q_t \mathbf{1}\left(P_t - \frac{1}{N}\sum_{i=1}^{N} P_{t-i+1} > 0\right), \quad Q_t^- = Q_t \mathbf{1}\left(P_t - \frac{1}{N}\sum_{i=1}^{N} P_{t-i+1} < 0\right), \tag{10}$$

где Q_t — данные по поиску в Google/Wikipedia в момент t, а $\mathbf{1}(\cdot)$ — функция-индикатор, равная 1, если встречается состояние из $(\cdot)$, и 0 в остальных случаях, в то время как N — число периодов, рассматриваемых для скользящего среднего ($N = 4$ для GoogleTrends, $N = 7$ для Wikipedia). Kristoufek (2013) предположил, что эти две переменные могут использоваться как прокси для поисковой активности, связанной с положительным и отрицательным откликом.

Kristoufek (2013) обнаружил значимое двунаправленное отношение, в котором поисковые запросы влияют на цены и наоборот, предполагая, что спекуляция и следование за трендом имеет определяющее влияние на динамику цены биткойна. Он установил, что если цены превышают недавний тренд, это повышает внимание инвесторов и, в свою очередь, влияет на дальнейший рост цены. Аналогично, когда цены находятся ниже их недавнего тренда, растущий интерес инвесторов подталкивает цены еще дальше вниз. Нет нужды говорить, что такой рынок может часто порождать ценовые пузыри, как это будет подробно рассмотрено в разделе 6.

Garcia et al. (2014) расширяют набор переменных, использованный Kristoufek (2013), рассматривая набор, состоящий из данных по ценам, активности в соцсетях, поисковым трендам и росту распространения биткойна. А именно, они рассмотрели следующие переменные:

• *число новых пользователей Биткойна*, использующих валюту в момент t, аппроксимированное числом загрузок программы-клиента Биткойна;

• *цена биткойна*, выраженная в трех мировых валютах (доллар США, евро и юань);

• *информационный поиск*, представленный нормализованными дневными данными по поиску в Google (или дневными просмотрами страницы Биткойна в Wikipedia в качестве проверки надежности);

• *обмен информацией* (*online word-of-mouth communication*), представленный дневным количеством сообщений в Твиттере, относящихся к биткойну, B_t, на миллион сообщений в ленте Твиттера, T_t, рассчитанным как $(B_t / T_t) \times 10^6$. Эти данные были загружены с сайта http://topsy.com, в них рассматривается количество твитов, содержащих хотя бы один из следующих терминов: «BTC», «#BTC», «bitcoin» или «#bitcoin». В качестве проверки робастности Garcia et al. (2014) также рассматривали альтернативную меру обмена информацией в виде числа «репостов» со старейшей, проявляющей регулярную активность, публичной страницы в Facebook, посвященной биткойну. Garcia et al. (2014) оценили четырехмерную векторную авторегрессию первого порядка с первыми разностями по данным за период с января 2009 г. до октября 2013 г. и нашли две положительные обратные связи: 1) между объемом поиска, обменом информацией и ценой, которую они назвали «*социальным циклом*», и 2) между объемом поиска, числом новых пользователей и ценой, названную «*циклом привлечения пользователей*». Первый цикл показывает, что растущая популярность биткойна ведет к увеличению объемов поиска, что приводит к росту активности социальных сетей, а это в свою очередь стимулирует покупку биткойнов новыми пользователями, повышая таким образом цены, и, впоследствии, сказываясь на объемах поиска. Второй цикл показывает, что новые пользователи биткойна загружают программу-клиент после получения информации о технологии биткойна. Рост числа пользователей биткойна впоследствии ведет к повышению цен, с учетом того, что количество биткойнов в обращении не зависит от спроса, а увеличивается определенным заданным образом с течением времени. Garcia et al. (2014) также обнаружили отрицательную зависимость цен от онлайн-поисков, показывая, что трем из четырех самых больших дневных падений цен предшествовали существенные увеличения в объеме поисков в Google днем ранее. Они показали, что онлайновая поисковая активность быстрее реагирует на негативные события, чем цены, так что поисковые пики — это ранние индикаторы падения цен. Проведенные проверки робастности подтвердили данные выводы.

Garcia и Schweitzer (2015) не только расширили предыдущую модель векторной авторегрессии первого порядка дополнительными социальными сигналами, но, что более интересно, впервые применили алгоритмическую стратегию торговли, основанную на этой авторегрессии, показывая возможность получения прибыли, даже с учетом риска и издержек на торговлю. Они использовали следующие переменные со значениями за период с февраля 2011 г. по декабрь 2014 г.:

- дневная *цена закрытия* BTC за каждый день в 23.59 GMT с сайта http://www.coindesk.com/;
- *дневной объем* биткойнов, обмененных на другие валюты на 80 онлайн-рынках, с сайта https://bitcoincharts.com/;
- *дневное количество блокчейн транзакций*, по данным измерений сайта https://blockchain.info/, проводимых каждый день в 18:15:05 UTC, которое они далее апроксимировали до 00:00 GMT следующего дня;

- число загрузок самого популярного биткойн-клиента с сайта http://sourceforge.net/projects/bitcoin/;
- *нормализованный объем поиска* по ключевому слову «bitcoin» в *GoogleTrends*;
- *дневное количество уникальных твитов о Биткойне*, сгруппированные в 24-часовые окна от 00:00 GMT, с сайта http://topsy.com/;
- *средняя дневная валентность твитов, имеющих отношение к Биткойну* (в психологических исследованиях валентность нацелена на количественную оценку степени удовольствия или недовольства при переживании эмоционального опыта[3]). Garcia и Schweitzer (2015) измерили среднюю дневную валентность, используя словарную технику, предложенную Warriner et al. (2013). Эта техника усовершенствована по сравнению с предыдущим ANEW словарным методом Bradley и Lang (1999) более чем тринадцатью тысячами слов, определяющими валентность. Они вычислили среднюю дневную валентность сообщений в твиттере по поводу Биткойна для дня t в два шага: сначала измерив частоту каждого термина в лексиконе в течение этого дня, а затем вычислив среднюю валентность, взвешивая каждое слово по его частоте;
- *дневная поляризация мнений о Биткойне в твиттере*, вычисленная как среднее геометрическое дневных коэффициентов, характеризующих количество положительных и отрицательных слов на один твит, относящийся к Биткойну. Поляризация мнений позволяет определить семантическую направленность слов в оценочных терминах «положительная» и «отрицательная», см. (Osgood, 1964). Garcia и Schweitzer (2015) использовали основанный на лексиконе психолингвистический метод лингвистического исследования и подсчета слов, предложенный Pennebaker et al. (2007), и распространили «лексикон стволов», используемый в этом методе, на слова, ставя их в соответствие наиболее часто употребляемым английским словам из набора данных GoogleBooks, см. подробности в (Lin et al., 2012). В конечном счете, Garcia и Schweitzer (2015) рассмотрели 3463 положительных термина и 4061 отрицательных. Важно отметить, что поляризация может рассматриваться как дополнительное к эмоциональной валентности измерение, поскольку она скорее оценивает одновременное существование положительного и отрицательного субъективного содержания, чем его всеобщую направленность, см. (Osgood, 1964; Tumarkin, Whitelaw, 2001).

Garcia и Schweitzer (2015) обнаружили, что только валентность, поляризация и объем продаж оказывают значимые эффекты на цену биткойна. Эти выбранные переменные затем используются для применения нескольких торговых стратегий, которые далее сравниваются с такими традиционными стратегиями, как «Buy and Hold», «Momentum» (которая предсказывает, что изменение цены в момент времени $t + 1$ будет таким же, как в момент t) и некоторыми другими, см. подробно (Garcia, Schweitzer, 2015). Они обнаружили, что комбинированная стратегия, использующая три указанные выше переменные, является лучшей на периоде бэктестинга даже с учетом риска и торговых издержек. Насколько авторам известно, эта работа на данный момент является единственной, в которой проведен полномасштабный прогнозный анализ на основе бектестинга.

Buchholz et al. (2012) расширили набор переменных, которые могут повлиять на цену биткойна, принимая во внимание не только привлекательность биткойнов, измеренную на основании данных GoogleTrends, но и учитывая воздействия предложения и спроса на биткойн.

[3] См. подробности в работах (Bradley, Lang, 1999; Russell, 2003; Garcia, Schweitzer, 2012).

Для измерения спроса они учитывали сумму предложений существующих биткойнов, общее число транзакций в биткойнах за день, общую стоимость транзакций в биткойнах за день и среднюю стоимость транзакций в биткойнах за день (полученную при делении общей стоимости транзакций на общее число транзакций). К сожалению, Buchholz et al. (2012) работали только с двумерными моделями VAR и VEC, не используя полный набор переменных, что потенциально приводит к смещению пропущенных переменных. Они также вычислили модель «GARCH в среднем», в которой рассматривали компонент волатильности в уравнении средних как показатель спроса на биткойны, однако недостаток управляющих переменных в уравнении средних является, опять же, относительно спорным. Более того, несколько интересных переменных, которые затрагивались в начале их работы (например, данные в статьях исторических новостей и блогах от LexisNexis), не были рассмотрены при эмпирическом анализе данных. Несмотря на эти недостатки, работа Buchholz et al. (2012) может считаться новаторской, поскольку она дала несколько важных подсказок, которые затем были включены в последующие более широкие типы анализа.

Bouoiyour и Selmi (2015), Bouoiyour et al. (2015), Kancs et al. (2015) провели первые исследования, в которых рассматриваются три набора факторов для моделирования динамики цен биткойна: 1) технические драйверы (предложение и спрос на биткойн), 2) показатели привлекательности и 3) макроэкономические переменные.

Использованные Bouoiyour и Selmi (2015) переменные и их описание приведены в табл. 1.

Bouoiyour и Selmi (2015) исследовали долговременные и кратковременные связи между ценами биткойна и предыдущим набором переменных, используя предложенную Pesaran и Shin (1999) процедуру тестирования границ интервала с помощью авторегрессионых распределенных лагов (ARDL). У этого подхода есть несколько преимуществ: во-первых, он является методом коинтеграции одного уравнения, который может быть оценен OLS; во-вторых, он позволяет моделировать динамику на краткосрочном и долгосрочном периодах; в-третьих, эта процедура может использоваться независимо от того, являются ли базовые регрессоры стационарными I(0), нестационарными I(1), или частично интегрированными. Наконец, это относительно более эффективный метод оценки в малых выборках по сравнению с альтернативными методами коинтеграции. Однако эта процедура не будет работать с регрессорами I(2), когда существует более чем одно отношение коинтеграции, или с эндогенными регрессорами. Модель ARDL, использованная Bouoiyour и Selmi (2015), представлена ниже:

$$\begin{aligned}\Delta \ln \mathrm{BPI_t} = {} & a_0 + \sum\nolimits_{i=1}^{n} a_{1i} \Delta \ln \mathrm{BPI}_{t-i} + \sum\nolimits_{i=0}^{m} a_{2i} \Delta \ln \mathrm{TTR}_{t-i} + \\ & + \sum\nolimits_{i=0}^{l} a_{3i} \Delta \ln \mathrm{ETR}_{t-i} + \sum\nolimits_{i=0}^{h} a_{4i} \Delta \ln MBV_{t-i} + \\ & + \sum\nolimits_{i=0}^{v} a_{5i} \Delta \ln \mathrm{EOV}_{t-i} + \sum\nolimits_{i=0}^{r} a_{6i} \Delta \ln \mathrm{HASH}_{t-i} + \\ & + \sum\nolimits_{i=0}^{s} a_{7i} \Delta \ln \mathrm{GP}_{t-i} + \sum\nolimits_{i=0}^{z} a_{8i} \Delta \ln SI_{t-i} + b_1 \ln BPI_{t-1} + \\ & + b_2 \ln TTR_{t-1} + b_3 \ln ETR_{t-1} + b_4 \ln MBV_{t-1} + b_5 \ln EOV_{t-1} + \\ & + b_6 \ln HASH_{t-1} + b_7 \ln GP_{t-1} + b_8 SI_{t-1} + \varepsilon_t .\end{aligned} \tag{11}$$

Таблица 1. Драйверы цены биткойна (BPI), использованные Bouoiyour и Selmi (2015)

Переменная	Пояснение
	Технические драйверы
Коэффициент биржевой торговли (ETR)	Биткойны используются главным образом для двух целей: он-лайн-покупок и биржевой торговли. Веб-сайт https://blockchain.info/ предоставляет общее число транзакций и их объем, за исключением биржевых торгов. Кроме того, предоставляется информация о соотношении между объемом биткойн-транзакций и биржевых транзакций
Скорость денежного обращения в Биткойн (MBV)	Это частота, с которой одна единица биткойна используется для покупок каких-либо активов за указанный период. В системе Биткойн скорость денежного обращения биткойнов заменена на так называемое *количество потраченных биткойн-дней*. Для вычисления этой переменной количество биткойнов в транзакции умножается на число дней, прошедших с момента, когда они были потрачены в предыдущий раз
Предполагаемый объем выходов (EOV)	Аналогичен совокупному объему выходов, но дополнен алгоритмом, который пытается устранить изменения в общей стоимости. Это значение должно точнее отражать реальный объем транзакции. Ожидается отрицательная связь между предполагаемым объемом выходов и ценой биткойна
Хэшрейт	Предполагаемое число гигахэшей (миллиардов хэшей) в секунду, которое выполняет сеть Биткойн. Это — показатель вычислительной мощности сети Биткойн
	Показатели привлекательности
Привлекательность инвесторов (TTR)	Ежедневная статистика просмотров, связанных с Биткойном, в Google, позволяющая оценить спекулятивный настрой пользователей
	Макроэкономические переменные
Цена на золото (GP)	По своим инвестиционным свойствам биткойн близок к золоту. При этом базовая стоимость биткойна не зависит от потребления или производственного процесса
Индекс Шанхайской биржи (SI)	Шанхайский рынок считается одним из крупнейших игроков в экономике Биткойн и рассматривается как потенциальный источник ценовой волатильности биткойна

Используя набор данных за период с 05.12.2010 г. по 14.06.2014 г., Bouoiyour и Selmi (2015) обнаружили, что на краткосрочном периоде привлекательность инвесторов, коэффициент биржевой торговли, предполагаемый объем выходов и индекс Шанхайской биржи оказывают положительное и существенное влияние на цену биткойна, тогда как скорость денежного обращения, хэшрейт и цена на золото не оказывают никакого влияния. Зато в долгосрочной перспективе только коэффициент биржевой торговли и хэшрейт оказывают существенное влияние на динамику цен биткойна. Эти же результаты справедливы и при включении фиктивной переменной для учета банкротства крупной китайской компании по торговле биткойнами в 2013 году, а также при использовании цен на нефть, индекса Доу–Джонса и фиктивной переменной для учета закрытия торговой площадки Silk Road, инициированного ФБР в октябре 2013 года. Аналогичные результаты получаются при декомпозиции дисперсии для цены биткойна и проведении тестов Грейнджера на причинность, рассчитанных на основе модели VEC (однако о значениях коэффициентов для этой последней модели не сообщается).

Kancs et al. (2015) используют полную многомерную модель VEC (см. уравнение (9)), подобно Kristoufek (2013), и ежедневные данные за период 2009–2014 гг. Однако, в отличие от последней работы и в поддержку исследования Bouoiyour и Selmi (2015), они рассматривали три типа драйверов для моделирования динамики цен биткойна: предложение и спрос на биткойн, привлекательность биткойна и глобальные макроэкономические и финансовые факторы. Использованные Kancs et al. (2015) переменные и их описание приведены в табл. 2.

Таблица 2. Драйверы цены биткойна, использованные Kancs et al. (2015)

Переменная	Пояснение
	Предложение и спрос на биткойн
Количество биткойнов	Общее количество всех биткойнов за историю, которое было собрано для учета в качестве суммарного запаса биткойнов в обращении
Количество транзакций	Количество уникальных транзакций в биткойнах за день
Количество адресов	Количество уникальных биткойн-адресов, использованных за день
Количество потраченных биткойн-дней (скорость денежного обращения)	Количество потраченных биткойн-дней для любой данной транзакции Для расчета количество биткойнов в транзакции умножается на число дней, за которые монеты были потрачены
	Показатели привлекательности
Просмотры в Википедии	Объем ежедневных просмотров страницы Биткойн в Википедии. Это — хороший инструмент для измерения интереса потенциальных инвесторов, но он не определяет, для чего используется информация: для принятия решений об инвестировании или же для онлайн-обмена товарами и услугами с помощью Биткойн
Новые участники	Число новых участников на онлайн-форумах Биткойн, полученное с сайта https://bitcointalk.org/. Отражает размер экономики Биткойн и инвестиционное поведение новых участников сети в результате привлечения их внимания
Новые посты	Количество новых постов на онлайн-форумах Биткойн, полученное с сайта https://bitcointalk.org/. Отражает влияние доверия и/или неопределенности на основе интенсивности обсуждений среди участников
	Макроэкономические переменные
Обменный валютный курс	Обменный валютный курс между долларом США и евро. Он выбран, поскольку цена биткойна выражена в долларах
Цена на нефть	Цены на нефть взяты из базы данных Управления по информации в области энергетики при министерстве энергетики США (US Energy Information Administration, EIA)
Доу–Джонс	Биржевой индекс Доу–Джонса

При исследовании краткосрочных воздействий Kancs et al. (2015) обнаружили, что на цену биткойна влияют собственные лаги цены, общее число биткойнов в обращении, скорость денежного обращения биткойна и статистика просмотров в Википедии. При исследовании долговременного воздействия выяснилось, что связанные со спросом на биткойн переменные (например, количество потраченных биткойн-дней, количество адресов) оказывают более сильное влияние на цену биткойна, чем драйверы предложения (например, число биткойнов). Как и предполагалось теоретически, увеличение числа биткойнов в обращении ведет к снижению цены биткойна, тогда как увеличение размера экономики сети Биткойн

(в виде количества адресов) и скорость ее роста ведут к увеличению цены биткойна. Переменные, связанные с привлекательностью биткойнов, оказывают самое сильное и статистически самое существенное влияние на цену биткойна. Число новых участников негативно сказывается на цене биткойна, подразумевая, что доминирует инвестиционное поведение новых инвесторов в результате привлечения их внимания, тогда как количество новых постов оказывает положительное влияние на цену биткойна, отражая при этом рост признания и доверия к системе Биткойн. Как и у Kristoufek (2013), количество просмотров Википедии оказывает статистически существенное и положительное влияние на цены биткойна. Kancs et al. (2015) выяснили, что все рассмотренные макроэкономические переменные не влияют существенно на цену биткойна в долгосрочном периоде, и тем самым подтвердили утверждения Bouoiyour and Selmi (2015) и идею Yermack (2013) о том, что биткойн относительно неэффективен как средство управления рисками по сравнению с другими разработками на рынке, и не может легко хеджироваться по сравнению с другими активами, которые определяются макроэкономическими факторами. В целом, полученные Kancs et al. (2015) результаты подтверждают, что факторы привлекательности биткойна все еще являются основными драйверами цены биткойна, за ними следуют переменные, связанные с традиционным предложением и спросом, а глобальные макрофинансовые переменные не играют никакой роли. Kancs et al. (2015) подчеркнули, что спекулятивное поведение биткойн-инвесторов на краткосрочном периоде не обязательно может являться нежелательной активностью (покрывающей избыточный риск от несклонных к риску участников и обеспечивающей ликвидность на рынке биткойна), но оно может увеличить волатильность цены и создавать ценовые пузыри, а также привести к чрезмерному накоплению биткойнов.

В отличие от предыдущих исследований, Kristoufek (2015) и Bouoiyour et al. (2015) проанализировали цену биткойна с точки зрения частотных интервалов. Kristoufek (2015) использовал подход непрерывных вейвлетов (wavelet coherence) для изучения изменений корреляций между ценой биткойна и широким набором переменных, включая основные факторы спроса и предложения, спекулятивные и технические драйверы, за период 14.09.2011–28.02.2014 и при различной частотности. Он выяснил, что фундаментальные факторы, такие как коэффициент биржевой торговли и предложение биткойна, играют существенную роль в долгосрочной перспективе. Что интересно, китайский фондовый индекс является важным источником эволюции цен биткойна, тогда как вклад динамики цен на золото представляется незначительным. Более того, он считает, что движущей силой цен биткойна является также онлайн-интерес инвесторов, который взвинчивает цены биткойна в моменты резкого скачка цен и обваливает их при быстрых падениях цен, это же заметили Kristoufek (2013) и Garcia et al. (2014). К сожалению, такой анализ страдает некоторыми недостатками, указанными Bouoiyour et al. (2015): зашумленные данные, такие как цены биткойна, могут сильно сместить оценочные связи, и такое смещение может даже быть увеличено в частотно-временных рамках, более подробно это рассмотрено у Ng, Chan (2012). Более того, вейвлет-анализ с использованием только двух переменных, как это делал Kristoufek (2015), имеет проблему, подобную простой регрессии без управляющих переменных, где оценочные параметры могут сильно смещаться. Эти недостатки побудили Bouoiyour et al. (2015) применить условное приближение причинности Грейнджера к частотной области, предложенное Breitung и Candelon (2006). Этот подход позволяет использовать несколько потенциальных управляющих переменных и отличать долгосрочные тренды, бизнес-циклы или краткосрочную динамику. Кроме того, он показывает наличие причинных связей между двумя

переменными даже в случае нелинейной зависимости, к тому же, такой подход устойчив к присутствию кластеризации волатильности, см. (Bodart, Candelon, 2009)[4].

Bouoiyour et al. (2015) использовали предыдущие рамки частотной области, чтобы проверить безусловную и условную причинность Грейнджера в ценах биткойна и наборе объясняющих переменных с целью исследования основных факторов, влияющих на динамику цен биткойна при различной частотности. Во-первых, они показали, что цены на биткойн (BPI) по Грейнджеру вызывают появление коэффициента биржевой торговли (ETR) в краткосрочном и среднесрочном циклическом компоненте, тогда как нулевая гипотеза об отсутствии причинности Грейнджера в диапазоне ETR–BPI не отклоняется при любой частотности. Этот последний результат отличается от полученного Kristoufek (2015). Он зафиксировал значительную причинность от ETR до BPI, которая усиливается в долгосрочной перспективе. Результаты Bouoiyour et al. (2015) не изменились при переходе от анализа безусловной причинности к анализу условной причинности, где в качестве управляющих переменных использовались китайский биржевой индекс и хэшрейт.

Bouoiyour et al. (2015) также обнаружили, что, согласно поисковым запросам Google, привлекательность инвесторов (TTR) становится причиной появления цены биткойна при более высокой частотности, что ожидаемо, потому что интерес к системе Биткойн постепенно растет. Наоборот, обратная причинно-следственная причинность от BPI к TTR значительна при более низкой частотности, что может свидетельствовать о том, что инвесторы покупают биткойны в основном по спекулятивным причинам. Интересно, что эти результаты существенно изменились, когда хэшрейт и китайский биржевой индекс были рассмотрены в качестве управляющих переменных при анализе условной причинности: в данном случае не было никакой существенной причинно-следственной связи в диапазоне BPI–TTR, тогда как обратная причинность TTR–BPI была значительной, как при более низкой, так и более высокой частотности. Это свидетельствует о важности китайского рынка и Биткойн-технологии при объяснении этих причинностей: первый фактор может сильно повлиять на краткосрочную спекулятивную деятельность, а второй может подразумевать, что более высокий интерес инвесторов вызывает увеличение количества аппаратных средств, задействованных в майнинге биткойнов, приводя к большему числу технических трудностей и последующему росту цен на биткойн, чтобы покрыть возросшие расходы на вычисление и энергопотребление. Проверки устойчивости, включающие дополнительные управляющие переменные, такие как скорость обращения биткойна и предполагаемый объем на выходе, не меняют существенно предыдущее доказательство. Полученные результаты вполне аналогичны тем, о которых сообщал Kristoufek (2015).

В целом, анализ, проведенный с использованием методов на основе частотных областей, подтвердил, что основные драйверы динамики цен на биткойн все еще носят спекулятивный характер. Однако есть несколько других значимых задействованных факторов, и не все из них связаны со спекуляцией, а также нельзя исключать возможность увеличения доли Биткойн-технологии в обработке деловых операций в долгосрочной перспективе.

В конце этого раздела хотелось бы отметить, что есть также другие работы, в которых сделана попытка смоделировать динамику цен на биткойн. Но они менее обстоятельны, массивы данных в них меньше, и почти все из них не рецензированы. Для получения дополнительной

[4] Ввиду ограничения объема статьи, технические детали данного подхода здесь не описаны, но они представлены в MPRA Paper No. 71946 по адресу https://mpra.ub.uni-muenchen.de/71946/1/MPRA_paper_71946.pdf.

информации предлагаем заинтересованным читателям посетить веб-страницу Википедии, содержащую справочные материалы по биткойну «Публикации, включающие исследования и анализ системы Биткойн или родственных областей» (https://en.bitcoin.it/wiki/Research).

6. Обнаружение пузырей и взрывоопасного поведения в ценах на биткойн

Устойчивая волатильность в ценах на биткойн вызвала сильные споры о том, может ли «существенный спекулятивный компонент» (Dowd, 2014) быть предвестником большого финансового пузыря. Было разработано несколько статистических критериев для проверки существования финансовых пузырей, некоторые из них недавно были применены к ценам на биткойн. Эти тесты можно разделить на две большие группы: тесты для обнаружения единственного пузыря и тесты для обнаружения (потенциального) множества пузырей.

6.1. Тестирование на наличие единственного пузыря

MacDonell (2014) был первым, кто проверил наличие пузыря в ценах на биткойн, используя подход логопериодических колебаний и степенного роста (Log Periodic Power Law, LPPL), предложенный Johansen et al. (2000) и Sornette (2003a, b). Ниже описывается его основная структура, а заинтересованного читателя можно отослать к указанным трем работам, а также исследованию (Geraskin, Fantazzini, 2013) для получения дополнительной информации.

Модель Johansen–Ledoit–Sornette (JLS) предполагает наличие двух типов агентов на рынке: группы трейдеров с рациональными ожиданиями и группы так называемых «шумных» трейдеров, т. е. иррациональных агентов со стадным поведением. Согласно данной модели, трейдеры организованы в сети и могут иметь только два статуса: покупка (buy) или продажа (sell). Кроме того, их действия по торговле зависят от решений других трейдеров и внешних воздействий. Из-за таких взаимодействий агенты могут образовывать группы с самоподобным поведением, что может привести рынок к появлению пузыря. Еще одна важная особенность, представленная в этой модели — положительные обратные связи, которые появляются при повышающемся риске и взаимодействии агентов, так что пузырь может быть самостоятельным процессом, подробнее об этом написано в (Geraskin, Fantazzini, 2013). Подходы к калибровке моделей финансовых пузырей с помощью LPPL приведены Sornette (2003a), а ожидаемые выводы по нескольким случаям с пузырями представлены в серии проверок в реальных условиях в (Sornette, Zhou, 2006; Sornette et al., 2009; Zhou, Sornette, 2003, 2006, 2008, 2009).

Ожидаемое значение логарифмической цены актива в стремящемся к росту пузыре, в соответствии с уравнением LPPL, представлено в виде

$$\mathrm{E}[\ln p(t)] = A + B(t_c - t)^{\beta} + C(t_c - t)^{\beta} \cdot \cos[\omega \ln(t_c - t) - \phi], \qquad (12)$$

где $A > 0$, $B < 0$, $|C| < 1$, $0 < \beta < 1$ — количественно оценивает ускорение цен по степенному закону, ω — частота колебаний во время существования пузыря, $0 < \phi < 2\pi$ — фазовый параметр.

Выражение (12) описывает временной рост цен до краха. Оно было предложено в различных формах в нескольких работах, см., например, (Sornette, 2003a; Lin et al., 2014) и ссылки в них.

Первое условие для пузыря в рамках JLS-модели — $0 < \beta < 1$, которое гарантирует, что темп риска краха ускоряется. Второе основное условие, предложенное Bothmer и Meister (2003), состоит в том, что показатель краха должен быть неотрицательным:

$$b = -B\beta - |C|\sqrt{\beta^2 + \omega^2} \geq 0. \tag{13}$$

Lin et al. (2014) добавили третье условие, которое требует, чтобы остатки из эмпирического уравнения (12) были постоянными. Они использовали методику Филлипса–Перрона (Phillips–Perron, PP) и расширенный тест Дики–Фуллера (ADF), чтобы сделать проверку на стационарность, а Geraskin и Fantazzini (2013) предложили использовать тест Kwiatkowski et al. (1992), посчитав его более мощным, при условии, что формирование базовых данных представляет собой процесс AR(1) с коэффициентом, близким к единице.

Оценка LPPL-моделей никогда не проходила легко из-за частого наличия множества локальных минимумов функции стоимости, в которой алгоритм минимизации может просто потеряться — см. работы (Fantazzini, 2010; Geraskin, Fantazzini, 2013; Filimonov, Sornette, 2013) для обзора данной проблемы и поиска возможных решений.

MacDonell (2014) использовал модель LPPL, чтобы успешно предсказать падение цен на биткойн, которое произошло 4 декабря 2013 года, показав, что LPPL-модели могут быть ценным инструментом для обнаружения пузыреобразного поведения электронных денег.

Cheah и Fry (2015) проводили проверку на наличие финансовых пузырей в ценах на биткойн, используя тест, предложенный Fry (2014), у которого была та же отправная точка, что и у Johansen et al. (2000). А именно, Cheah и Fry (2014) предположили, что

$$P(t) = P_1(t)(1-k)^{j(t)}, \quad \text{где} \quad dP_1(t) = [\mu(t) + \sigma^2(t)/2]P_1(t)dt + \sigma(t)P_1(t)dW_t,$$

где W_t — винеровский процесс, $j(t)$ — скачкообразный процесс:

$$j(t) = \begin{cases} 0, & \text{до момента краха,} \\ 1, & \text{после момента краха,} \end{cases}$$

тогда как k представляет собой процент потери стоимости активов после краха. До краха имеем $P(t) = P_1(t)$, и с помощью формулы Ито можно показать, что $X_t = \ln P(t)$ удовлетворяет уравнению

$$dX_t = \mu(t)dt + \sigma(t)dW_t - \nu dj(t), \quad \nu = -\ln(1-k) > 0. \tag{14}$$

Далее Fry (2014) и Cheah, Fry (2015) ввели следующие два предположения.

Предположение 1 (коэффициент доходности). Истинный коэффициент доходности постоянен и равен μ:

$$\mathrm{E}[X_{t+\Delta} - X_t \mid X_t] = \mu\Delta + o(\Delta). \tag{15}$$

Предположение 2 (уровень риска). Истинный уровень риска постоянен и равен σ^2:

$$\mathrm{Var}[X_{t+\Delta} - X_t \mid X_t] = \sigma^2\Delta + o(\Delta). \tag{16}$$

Более того, предположив, что краха не произошло к моменту времени t, они получили

$$\mathrm{E}[j(t+\Delta)-j(t)] = \Delta h(t) + o(\Delta), \tag{17}$$

$$\mathrm{Var}[j(t+\Delta)-j(t)] = \Delta h(t) + o(\Delta), \tag{18}$$

где $h(t)$ — уровень риска. Используя уравнение (15) вместе с (14) и (17), получаем

$$\mu(t)-vh(t)=\mu; \quad \mu(t)=\mu+vh(t). \tag{19}$$

Это показывает, что коэффициент доходности должен увеличиться, чтобы компенсировать инвестору риск краха.

Fry (2014), Cheah и Fry (2015) показали, что в пузыре не только цены должны вырасти, но и волатильность должна уменьшиться. Используя уравнения (14), (16) и (18), они получили

$$\sigma^2(t)+v^2h(t)=\sigma^2; \quad \sigma^2(t)=\sigma^2-v^2h(t). \tag{20}$$

Основные уравнения (18) и (19) показывают, что во время существования пузыря инвестор должен получить компенсацию риска краха за счет повышенного коэффициента доходности $\mu(t)>\mu$ (μ — коэффициент доходности в долгосрочной перспективе), тогда как волатильность рынка снижается, обеспечивая чрезмерное доверие на рынке (Fry, 2012, 2014). Более того, возможно провести тестирование на наличие спекулятивного пузыря, проверяя одностороннюю гипотезу

$$\mathbf{H_0}: v = 0, \quad \mathbf{H_1}: v>0. \tag{21}$$

Fry (2014) далее показал, что, с учетом указанных выше предположений, цена базового актива при отсутствии пузыря ($v=0$) представляется как

$$P_F(t)=\mathrm{E}(P(t))=P(0)e^{\tilde{\mu}t}, \tag{22}$$

где $\tilde{\mu}=\mu+\sigma^2/2$.

Во время существования пузыря ($v>0$)

$$X_t=N(X_0+\mu t+vH(t),\sigma^2 t-v^2H(t)), \quad \text{где } H(t)=\int_0^t h(u)du, \tag{23}$$

так что стоимость актива представляется в виде

$$P_B(t)=\mathrm{E}(P(t))=P(0)e^{\tilde{\mu}t+\left(v-v^2/2\right)H(t)}. \tag{24}$$

Уравнения (22) и (24) могут тогда использоваться совместно с надлежащей функцией риска $h(t)$ для расчета составляющей пузыря в цене актива, которая определяется как «среднее расстояние» между базовыми ценами и ценами пузыря. Fry (2014), Cheah и Fry (2015) использовали следующую функцию риска:

$$h(t)=\frac{\beta t^{\beta-1}}{\alpha^\beta+t^\beta}, \tag{25}$$

а составляющая пузыря представлена в виде

$$1-\frac{1}{T}\int_0^T\frac{P_F(t)}{P_B(t)}dt=1-\frac{1}{T}\int_0^T\left(1+t^\beta/\alpha^\beta\right)^{-\left(v-v^2/2\right)}dt, \tag{26}$$

где T является измерением выборки. Из уравнения (22) следует, что если базовая стоимость активов равна нулю, то

$$\lim_{t\to\infty} P_F(t) = 0\,. \tag{27}$$

Вслед за MacDonell (2014) Cheah и Fry (2015) провели проверку на наличие пузыря в ценах на биткойн в период с 1 января 2013 г. по 30 ноября 2013 г., до обвала цен в декабре 2013 г. Они отклонили нулевую гипотезу в (21) и посчитали, что параметр не отличается статистически от нуля, что совпадает с базовым значением нуля в долгосрочной перспективе. Более того, они обнаружили, что составляющая пузыря представлена в приблизительно 48.7% наблюдаемых цен. Эти результаты подтверждены несколькими проверками устойчивости.

6.2. Тестирование на наличие множества пузырей

Предыдущие тесты разработаны для проведения проверок на наличие единственного пузыря и могут использоваться для обнаружения множеств пузырей, только если они повторяются скользящим окном во времени, как это было проведено в (Sornette et al., 2009; Jiang et al., 2010; Geraskin, Fantazzini, 2013; Cheah, Fry, 2015). Тесты, специально разработанные для обнаружения множеств пузырей, недавно были предложены Phillips и Yu (2011), Phillips et al. (2011, 2015), и они все используют последовательные тесты с прокручивающимися оценочными окнами. А именно, эти тесты основаны на последовательных регрессиях ADF-типов, использующих окна различного размера, и могут последовательно определять и датировать случаи с множеством пузырей даже на небольшой выборке. Эти тесты использовали Malhotra и Maloo (2014) для проведения проверки на наличие «взрывной» динамики в ценах на биткойн. Остановимся на ADF-тесте обобщенного супремума (GSADF), предложенного Phillips et al. (2015) (известного как методика PSY) и опирающегося на работы (Phillips, Yu, 2011; Phillips et al., 2011). Данный тест имеет лучшие статистические свойства при обнаружении множеств пузырей, чем два предыдущих.

На первом этапе рассмотрим регрессию ADF для прокручиваемой выборки, где начальная точка задана частью r_1 общего числа наблюдений, конечная точка — частью r_2, а размер окна — $r_w = r_2 - r_1$. Регрессия ADF представлена в виде

$$x_t = \mu + \rho x_{t-1} + \sum_{i=1}^{p} \phi_{r_w}^{i} x_{t-i} + \varepsilon_t\ , \tag{28}$$

где μ, ρ и $\phi_{r_w}^{i}$ оцениваются с помощью OLS, и нулевая гипотеза единичного корня $\rho = 1$ по сравнению с авторегрессионным коэффициентом $\rho > 1$.

Phillips et al. (2015) предложили ADF-тест обратного супремума, при котором конечная точка зафиксирована на r_2, а размер окна расширен от первоначального r_0 до r_2. Тогда статистика критерия представлена как

$$BSADF_{r_2}(r_0) = \sup_{r_1 \in [0, r_2 - r_0]} ADF_{r_1}^{r_2}. \tag{29}$$

Заметим, что тест Phillips et al. (2011) — это частный случай ADF-теста обратного супремума (BSADF), где $r_1 = 0$, так что операция sup излишняя.

Обобщенный ADF-тест супремума (GSADF) вычисляется неоднократным выполнением BSADF-теста для каждого $r_2 \in [r_0, 1]$:

$$GSADF(r_0) = \sup_{r_2 \in [r_0, 1]} BSADF_{r_2}(r_0) \ . \tag{30}$$

Теорема 1 в (Phillips et al., 2015) представляет предельное распределение (30) при нулевом случайном блуждании с асимптотически малым сдвигом, тогда как критические значения достигаются численным моделированием. Если нулевая гипотеза об отсутствии пузырей отклонена, то на втором этапе становится возможным датировать начальные и конечные точки одного пузыря или более. А именно, начальная точка задается датой T_{re}, когда последовательность критериев BSADF-теста пересекает критическое значение снизу, а конечная точка T_{rf} — когда последовательность BSADF пересекает соответствующее критическое значение сверху:

$$\begin{aligned} \hat{r}_e &= \inf_{r_2 \in [r_0, 1]} \left\{ r_2 : BSADF_{r_2}(r_0) > cv_{r_2}^{\beta_T} \right\}, \\ \hat{r}_f &= \inf_{r_2 \in [\hat{r}_e + (\delta \ln T)/T, 1]} \left\{ r_2 : BSADF_{r_2}(r_0) < cv_{r_2}^{\beta_T} \right\}, \end{aligned} \tag{31}$$

где $cv_{r_2}^{\beta_T}$ — критическое значение справа статистического критерия BSADF по наблюдениям $\lfloor T_{r2} \rfloor$, $\lfloor \cdot \rfloor$ — функция, округляющая число до ближайшего целого, δ — выбираемый параметр, который определяет минимальную продолжительность жизни пузыря: это значение обычно берется 1, как в (Phillips et al., 2011, 2015) и в большинстве предыдущих работ, в которых оговаривается условие минимальной продолжительности жизни пузыря при наблюдениях $\ln T$ (т. е. часть выборки $(\ln T)/T$). Однако могут использоваться различные значения в зависимости от частотности данных, см. (Figuerola-Ferretti et al., 2016).

Malhotra и Maloo (2014) проводили проверку на наличие пузырей, используя GSADF-тест с набором данных за период с середины 2011 г. до февраля 2014 г. Они выявили свидетельства взрывоопасного поведения в обменных курсах биткойна по отношению к доллару США в августе — октябре 2012 г. и ноябре 2013 — феврале 2014 г. Они считают, что первый случай пузыреобразного поведения (август — октябрь 2012 г.) может быть приписан внезапному увеличению внимания средств массовой информации к биткойну. А второй случай произошел в силу ряда причин, включая вызванный потолком госдолга кризис в США, закрытие торговой площадки Silk Road по указанию ФБР, рост популярности китайской биржи BTC–China и увеличивающееся число предупреждений регулирующих органов и Центральных банков по всему миру после закрытия японской биржи Mt.Gox.

7. Определение цены

Brandvold et al. (2015) были первыми (и пока единственными), кто проводил изучение процесса определения цен на рынке Биткойн, который состоит из нескольких независимых бирж. Эта тема часто обсуждается в биткойн-сообществе, поскольку знание того, какая биржа быстрее других отреагирует на новую информацию (таким образом, наиболее точно отражая стоимость биткойна), имеет явно первостепенное значение как для краткосрочных

трейдеров, так и для долгосрочных инвесторов. В литературе, посвященной определению цен, описываются в основном три методологии: метод совместного использования информации по Hasbrouck (1995), постоянно-переходящая декомпозиция по Gonzalo, Granger (1995) и структурная многомерная модель временного ряда по De Jong et al. (2001), которая является продолжением класса моделей, первоначально предложенных Harvey (1989). Brandvold et al. (2015) использовали метод (De Jong et al., 2001), потому что он имеет преимущество в том, что совместно используемая информация уникально определена, в отличие от вычисленной с помощью модели (Hasbrouck, 1995), и учитывает изменчивость нововведений, в отличие от методики (Gonzalo, Granger, 1995), тем самым ряд цен с низкой изменчивостью нововведений получает низкий показатель совместно используемой информации. Учитывая важность модели De Jong et al. (2001), приведем ее ниже.

Многомерная модель De Jong et al. (2001) была предложена для оценки совместно используемой информации различных бирж в отношении информации, созданной всем рынком. Цены состоят из двух компонентов: одного простого (ненаблюдаемого) случайного блуждания и характерного специфического шума для каждой биржи. Компонент случайного блуждания в равной степени может означать либо эффективную цену, либо основную составляющую новостей. Из такой структуры моделей непосредственно следует, что цены на биржах коинтегрированы при построении, тогда как специфическая составляющая может иметь место из-за особых условий на бирже, стратегического поведения трейдеров или иных шоков.

Brandvold et al. (2015) предположили наличие n отдельных бирж и соответствующих рынков: рынок для биржи определяется как все остальные биржи вместе взятые. Обозначим P^e — вектор биржевых цен, P^m — вектор рыночных цен, U^e и U^m — векторы специфических шоков для бирж и рынков соответственно, P^* — эффективная цена, $p^e = \ln P^e$, $u^e = \ln U^e$, $p^* = \ln P^*$, так что логарифм вектора n биржевых цен и вектора рыночных цен можно представить как

$$p^e = p^* \iota + u^e, \qquad p^m = p^* \iota + u^m, \tag{32}$$

где $\iota = (1,...,1)'$ — $n \times 1$-вектор, p^* является случайным блужданием.

Это частный случай структурной модели ненаблюдаемых составляющих, см. (Harvey, 1989) для получения дополнительной информации. Если обозначим лог-доходности эффективной цены на интервале (t–1, t) как $r_t = p_t^* - p_{t-1}^*$, то смоделированные предположения можно представить в виде

$$\begin{aligned}
&\mathrm{E}[r_t] = 0, \quad \mathrm{E}[r_t^2] = \sigma^2, \quad \mathrm{E}[r_t\, u_t^e] = \psi^e, \quad \mathrm{E}[r_t\, u_t^m] = \psi^m, \\
&\mathrm{E}[r_t\, u_{t+l}^e] = \gamma_l^e, \quad \mathrm{E}[r_t\, u_{t+l}^m] = \gamma_l^m \ \ (l \ge 1), \quad \mathrm{E}[r_t\, u_{t-k}^e] = 0, \quad \mathrm{E}[r_t\, u_{t-k}^m] = 0 \ \ (k \ge 1), \\
&\mathrm{E}[u_t^e (u_t^e)'] = \Omega^e, \quad \mathrm{E}[u_t^e (u_t^m)'] = \Omega, \quad \mathrm{E}[u_t^e (u_{t-k}^m)'] = 0 \ \ (k \ne 0),
\end{aligned} \tag{33}$$

где $\psi^m, \psi^e, \gamma^m, \gamma^e$ — $n \times 1$-векторы, а $\Omega = \{\omega_{ij}\}$, $\Omega^e = \{\omega_{ij}^e\}$ — $n \times n$-матрицы.

Основная составляющая новостей r_t может коррелировать с действующими и будущими специфическими составляющими, но в иных случаях не коррелирует. И наоборот, специфические составляющие не коррелируют и отражают наличие шума во внутрисуточных данных. Эти ограничения на структуру корреляций необходимы для определения модели,

см. подробнее (Harvey, 1989). Учитывая предыдущую структуру, лог-доходности наблюдаемых цен определяются следующим образом:

$$y_t^e = p_t^e - p_{t-1}^e = p_t^* \iota + u_t^e - p_{t-1}^* \iota - u^e{}_{t-1} = r_t \iota + u_t^e - u_{t-1}^e, \tag{34}$$

$$y_t^m = p_t^m - p_{t-1}^m = p_t^* \iota + u_t^m - p_{t-1}^* \iota - u_{t-1}^m = r_t \iota + u_t^m - u_{t-1}^m. \tag{35}$$

Используя формулы (33), матрицы ковариаций на бирже можно представить как

$$\mathrm{E}[y_t^e (y_t^e)'] = \sigma^2 \iota\iota' + \iota\psi^{e\prime} + \psi^e \iota' + 2\Omega^e,\ \mathrm{E}[y_t^e (y_{t-1}^e)'] = -\psi^e \iota' - \Omega^e + \gamma_1^e \iota',\ \mathrm{E}[y_t^e (y_{t-2}^e)'] = -\gamma_2^e \iota'. \tag{36}$$

Аналогично, матрица ковариаций между биржей и ее соответствующим рынком, т. е. между вектором y^e и вектором y^m, представима в виде

$$\mathrm{E}[y_t^e (y_t^m)'] = \sigma^2 \iota\iota' + \iota(\psi^m)' + \psi^e \iota' + 2\Omega,\ \mathrm{E}[y_t^e (y_{t-1}^m)'] = -\psi^e \iota' - \Omega + \gamma_1^e \iota',\ \mathrm{E}[y_t^e (y_{t-2}^m)'] = -\gamma_2^e \iota', \tag{37}$$

тогда как автокорреляция первого порядка для биржи представима как

$$\rho_{1,i} = \frac{-(\psi_i^e + \omega_{ii}^e - \gamma_i^e)}{\sigma^2 + 2(\psi_i^e + \omega_{ii}^e)}. \tag{38}$$

Параметр ψ_i, который является ковариацией между основной составляющей новостей и специфической составляющей, представляет критическую важность, потому что показывает, как рынок обучается после изменения цен на отдельной бирже: высокое значение ψ_i показывает, что обновление цен, полученное с этой биржи, имеет высокую информационную ценность для всего рынка. Для объяснения данного вопроса рассмотрим ковариацию между основной составляющей новостей и изменением цен на бирже

$$\mathrm{Cov}(y_{it}^e, r_t) = \sigma^2 + \psi_i^e, \tag{39}$$

которая выводится из уравнений (33) и (34). Отсюда следует, что n ковариаций между обновлениями на бирже и основной составляющей новостей определяются $n+1$ параметрами, так что требуются идентифицирующие ограничения. De Jong et al. (2001) предложили идею, согласно которой сгенерированная каждым обновлением цены информация должна в среднем равняться σ^2, т. е. дисперсии r_t. Следовательно, если рассмотреть среднюю ковариацию между изменением цен на выбранной бирже и основными новостями

$$\sum_{i=1}^{n} \pi_i \mathrm{Cov}(y_{it}^e, r_t) = \pi'(\sigma^2 \iota + \psi^e) = \sigma^2 + \pi'\psi^e, \tag{40}$$

где π — вектор весов, составляющих в сумме единицу, то предположение, что безусловная ковариация изменения биржевых цен и новостного компонента равна σ^2, влечет ограничение $\pi'\psi^e = 0$. Этого ограничения достаточно для определения параметров модели и интерпретации π_i как действий на бирже i, определяемых как часть проведенных торговых операций на бирже или просто как вероятность сделки на бирже (Brandvold et al., 2015). Если умножить ковариацию между основной составляющей новостей и изменением цен на бирже i (см. уравнение (39)) на вероятность π_i, то получим, сколько информации сгенерировано изменением цен на этой бирже. Разделив полученную величину на общую информацию, сгенерированную на рынке σ^2, получим величину совместно используемой информации для биржи i:

$$IS_i = \frac{(\sigma^2 + \psi_i^e)\pi_i}{\sigma^2} = \pi_i \left(1 + \frac{\psi_i^e}{\sigma^2}\right). \quad (41)$$

De Jong et al. (2001) и Brandvold et al. (2015) подчеркнули, что такое определение совместно используемой информации имеет некоторые привлекательные свойства. Во-первых, информационные доли в сумме составляют единицу, упрощая, таким образом, интерпретацию и облегчая добавление или удаление бирж из модели. Во-вторых, совместно используемая информация двух бирж просто равняется сумме их индивидуальных долей. В-третьих, биржа с положительной одновременной ковариацией между ее специфической составляющей и основной составляющей новостей ($\psi_i^e > 0$) имеет более высокую долю информации, чем долю в торговой активности.

Процедура оценки параметров моделей состоит из двух этапов: сначала оцениваются ковариации выборки $\mathrm{E}[y_t^e y_{t-k}^e]$, где k=0, 1, 2, и автокорреляции $\rho_{1,i}$, далее вычисляются структурные параметры с помощью уравнений (37)–(38) и программного обеспечения для нелинейной оптимизации. Кроме того, некоторые параметры могут быть найдены напрямую. Учитывая, что σ^2 есть дисперсия r_t, и предположение Brandvold et al. (2015) о том, что семь бирж в их наборе данных представляют весь рынок Биткойн, σ^2 можно вычислить как дисперсию агрегированной доходности семи бирж. Аналогично, γ может быть вычислен непосредственно из ковариации выборки между доходностью рынка и доходностью соответствующей биржи с лагом в два интервала. На втором этапе остается лишь оценить Ω^e, Ω, ψ^e, ψ^m.

(Brandvold et al., 2015) подчеркнули, что в финансовой литературе нет единого мнения о том, как измерить торговую активность конкретной биржи относительно всей торговой активности на рынке (т. е. π_i), и предпочли использовать линейную комбинацию объема торговли и числа сделок. Однако они также отметили, что выбор π_i влияет только лишь на величину совместно используемой информации, а не на отношение между информацией и долей деятельности, т. е. показывает, является ли ψ_i положительным или отрицательным (Brandvold et al., 2015). В этой связи они предложили также рассмотреть простой случай равных π_i для каждой биржи, чтобы проверить устойчивость результатов модели.

Brandvold et al. (2015) использовали данные семи бирж: Bitfinex, Bitstamp, BTC-e (Btce), BTC–China (Btcn), Mt.Gox (Mtgox), Bitcurex и Canadian Virtual Exchange (Virtex). Исходные данные тиковых котировок преобразовывались в 5-минутные интервалы и охватывали период с 1 апреля 2013 г. по 25 февраля 2014 г. до банкротства Mtgox. Они обнаружили, что двумя биржами с положительным ψ в течение всего периода были Btce и Mtgox, таким образом, указывая, что эти биржи были информативнее конкурентов. Об аналогичном свидетельствовали предоставленные сведения о совместно используемой информации, наиболее высокими показатели были у Btce и Mtgox (0.322 и 0.366 соответственно). Однако Brandvold et al. (2015) подчеркнули, что, даже если другие биржи имеют отрицательный ψ и более низкий показатель совместно используемой информации, они все же предоставляют рынку информацию, только менее содержательную. Brandvold et al. (2015) также исследовали, как совместно используемая информация менялась со временем. Показатель совместно используемой информации Btcn сначала увеличился с 0.040 в апреле 2013 г. до 0.325 в декабре 2013 г., потому что некоторые крупные китайские компании (типа Baidu) начали принимать биткойн в качестве платежного средства, но затем показатель совместно используемой информации упал до 0.124 в январе 2014 г. после того, как китайское правительство запретило компаниям проводить расчеты в биткойнах. У Mtgox был самый высокий показатель

совместно используемой информации в начале (0.667), потом он стал постепенно уменьшаться, последние скачки наблюдались в январе и феврале 2014 г. в связи с ростом недоверия из-за возможного банкротства.

Brandvold et al. (2015) рассмотрели, что происходило во время и после ценового шока 2 октября 2013 г., когда владелец торговой площадки Silk Road был арестован американскими властями, а сайт закрыт, см. подробнее (Konrad, 2013). Они увидели, что Btce была единственной биржей с положительным ψ в этот период и имела значительно более высокий показатель совместно используемой информации, чем показатель совместной деятельности. И здесь возникает спор о том, что либо большая доля информированных трейдеров переключилась на Btce в этот период, либо просто трейдеры на Btce вдруг стали знать больше. Более того, они также подчеркнули, что Btce известна в Биткойн-сообществе наличием хорошего прикладного программного интерфейса для трейдеров, позволяющего размещать торговых роботов, и это может помочь объяснить, почему Btce в большей степени способствовала процессу определения цен в этот период, чем ее конкуренты.

8. Заключение

В работе рассмотрены известные на данный момент эконометрические и математические инструменты, предлагаемые для моделирования цены биткойна, а также несколько других аспектов, относящихся к данной тематике. В частности, впервые рассмотрены методы, используемые для определения основных характеристик пользователей системы Биткойн. Обнаружено, что большинство пользователей — это энтузиасты-программисты и люди, вероятно, занимающиеся незаконной деятельностью, и имеется лишь небольшая часть пользователей, которых влекут политические причины или инвестиционные мотивы. Тем не менее, данные методы анализа имеют ряд ограничений, например, возможность того, что исследованные выборки могут быть нерепрезентативными в плане полного контингента пользователей. Еще одно ограничение — скорость, с которой изменяются биткойн-рынки и пользователи с течением времени, ввиду чего проведенные исследования быстро теряют свою актуальность. Проанализированы основные модели, предложенные для оценки базовой стоимости биткойна, начиная от емкости рынка (что больше подходит для среднесрочной и долгосрочной перспективы) и заканчивая расчетом предельного уровня издержек производства на основе потребления электроэнергии, что представляет собой нижнюю границу на краткосрочном периоде. Кроме того, описано несколько эконометрических подходов, предложенных для моделирования динамики цен биткойна, начиная с моделей кросс-секционной регрессии с участием большинства торгуемых цифровых валют, переходя далее к одномерным и многомерным моделям временных рядов и заканчивая частотными моделями. Все эти методы подтвердили, что основные факторы динамики цен биткойна по-прежнему носят в основном спекулятивный характер, далее следуют традиционные переменные, связанные со спросом и предложением, а глобальные макрофинансовые переменные не играют никакой роли. Рассмотрены тесты, используемые для обнаружения наличия финансовых пузырей в ценах на биткойн. Такие тесты могут быть разделены на две большие группы в зависимости от предназначения: для выявления единственного пузыря или обнаружения (потенциального) множества пузырей. В большинстве из этих тестов исследованы несколько месяцев до обвала цен, начавшегося в декабре 2013 г. Один

анализ проводился на обнаружение множества пузырей по выборке 2011–2014 гг., когда были найдены доказательства взрывного поведения в обменных курсах биткойна по отношению к доллару США за период август—октябрь 2012 г. и ноябрь 2013 г. — февраль 2014 г. Наконец, рассмотрено недавнее исследование, посвященное процессу формирования цен на рынке биткойна. Данный процесс имеет большое значение не только для краткосрочных трейдеров, но и для долгосрочных инвесторов, которые хотят знать, какая биржа быстрее других реагирует на новую информацию, отражая, таким образом, стоимость биткойна наиболее точно и эффективно.

Данный обзор указывает несколько возможных направлений для дальнейших исследований. Например, эконометрических методов управления рыночными рисками с ценами на биткойн практически не существует. По сведениям авторов, единственной работой, которая подходит для распределений, зависящих от нескольких параметров, и используется для оценки стоимости, подверженной риску (VaR), и ожидаемых потерь (ES), является исследование Chu et al. (2015). К сожалению, они рассматривали лишь безусловные оценки, которые пренебрегают условной гетероскедастичностью, и поэтому не рекомендуются для крайне неустойчивых временных рядов, таких как цена на биткойн, см. об этом (Fantazzini, 2009; Weiß, 2011, 2013), для проведения крупномасштабного моделирования и эмпирических исследований VaR и ES для линейных портфелей. К тому же, несмотря на изменения в местных нормативных актах, появление новых инвесторов, вмешательство полиции (история с торговой площадкой Silk Road) и существенные улучшения аппаратных средств для майнинга, нет исследовательских работ, посвященных структурным разрывам и долгой памяти в ценах на биткойн. Кроме того, имеется большой перечень эконометрической и статистической литературы по прогнозированию со структурными разрывами, что может представлять интерес для алгоритмической торговли биткойнами, см. недавнее исследование Zhao (2015) нескольких алгоритмов с помощью методов Монте–Карло. Кроме того, все рассмотренные модели являются (логарифмически) линейными, но, принимая во внимание поведение цен на биткойн, нелинейные модели могут оказаться особенно полезными для целей прогнозирования, см. (Tong, 1990; Franses, Dijk, 2000; Wood, 2006; Terasvirta et al., 2011). Пожалуй, самое интересное расхождение, которое авторы нашли при подготовке данного обзора, это то, что документы, относящиеся к сфере ИТ, содержат информацию в основном об издержках на оплату электроэнергии и включают показатели энергоэффективности и вычислительной эффективности, в то время как документы, относящиеся к сфере экономики, редко принимают во внимание эти факторы. Таким образом, еще одним направлением дальнейших исследований является междисциплинарной анализ, позволяющий рассмотреть все данные аспекты вместе.

Список литературы

(см. References ниже).

Поступила в редакцию 01.06.2016;
принята в печать 25.09.2016.

Fantazzini D., Nigmatullin E., Sukhanovskaya V., Ivliev S. Everything you always wanted to know about bitcoin modelling but were afraid to ask. Part 2. *Applied Econometrics*, 2017, v. 45, pp. 5–28.

Dean Fantazzini
Moscow School of Economics, Moscow State University, Russian Federation;
fantazzini@mse-msu.ru

Erik Nigmatullin
Bocconi University, Milan, Italy; nigmatullin.erik@gmail.com

Vera Sukhanovskaya
Perm State National Research University, Russian Federation;
vera-sukhanovskaya@yandex.ru

Sergey Ivliev
Perm State National Research University, Russian Federation;
ivliev@gmail.com

Everything you always wanted to know about bitcoin modelling but were afraid to ask. Part 2

This part completes the consultation series dealing with bitcoin price modelling. Particularly, the analysis focuses on the econometric approaches suggested to model bitcoin price dynamics, the tests used for detecting the existence of financial bubbles in bitcoin prices and the methodologies suggested to study the price discovery at bitcoin exchanges.

Keywords: crypto-currencies; hash rate; investors' attractiveness; social interactions; money supply; money demand; speculation; forecasting; algorithmic trading; bubble; price discovery.

JEL classification: C22; C32; C51; C53; E41; E42; E47; E51; G17.

References

Bouoiyour J., Selmi R. (2015). What does bitcoin look like? *Annals of Economics and Finance*, 16 (2), 449–492.

Bouoiyour J., Selmi R., Tiwari A. K. (2015). Is bitcoin business income or speculative foolery? New ideas through an improved frequency domain analysis. *Annals of Financial Economics*, 10 (1), 1550002.

Bothmer H. C. G. V., Meister C. (2003). Predicting critical crashes? A new restriction for the free variables. *Physica A*, 320, 539–547.

Bodart V., Candelon B. (2009). Evidence of interdependence and contagion using a frequency domain framework. *Emerging Markets Review*, 10 (2), 140–150.

Bradley M. M., Lang P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1. The Center for Research in Psychophysiology, University of Florida.

Brandvold M., Molnár P., Vagstad K., Valstad O. C. A. (2015). Price discovery on Bitcoin exchanges. *Journal of International Financial Markets, Institutions and Money*, 36, 18–35.

Breitung J., Candelon B. (2006) Testing for short and long-run causality: A frequency domain approach. *Journal of Econometrics*, 132, 363–378.

Buchholz M., Delaney J., Warren J., Parker J. (2012). Bits and bets, information, price volatility, and demand for bitcoin. Economics 312. http://www.bitcointrading.com/pdf/bitsandbets.pdf.

Cheah E. T., Fry J. (2015). Speculative bubbles in Bitcoin markets? An empirical investigation into the fundamental value of bitcoin. *Economics Letters*, 130, 32–36.

Chu J., Nadarajah S., Chan S. (2015). Statistical analysis of the exchange rate of bitcoin. *PloS ONE*, 10 (7), e0133678.

De Jong F., Mahieu R., Schotman P., van Leeuwen I. (2001). Price discovery on foreign exchange markets with differentially informed traders. *Tinbergen Institute Discussion Paper Series*, No. TI 99–032/2.

Dowd K. (2014). New private monies: A bit-part player? *Cobden Centre Hobart Paper No. 174.* London: Institute of Economic Affairs.

Fantazzini D. (2009). The effects of misspecified marginals and copulas on computing the Value at Risk: A Monte Carlo study. *Computational Statistics & Data Analysis*, 53 (6), 2168–2188.

Fantazzini D. (2010). Modelling bubbles and anti-bubbles in bear markets: A medium-term trading analysis. In: *Handbook of Trading*, ed. by G. Gregoriou, 365–388. McGraw-Hill.

Figuerola-Ferretti I., Gilbert C., Mccrorie J. (2016). Testing for mild explosivityand bubbles in LME non-ferrous metals prices. *Journal of Time Series Analysis*, 36 (5), 763–782.

Filimonov V., Sornette D. (2013). A stable and robust calibration scheme of the log-periodic power law model. *Physica A*, 392 (17), 3698–3707.

Fry J. (2012). Exogenous and endogenous crashes as phase transitions in complex financial systems. *The European Physical Journal B*, 85 (12), 405.

Fry J. (2014). Multivariate bubbles and antibubbles. *The European Physical Journal B*, 87 (8), 1–7.

Franses P., Dijk V. (2000). *Nonlinear time series models in empirical finance*. Cambridge University Press.

Garcia D., Tessone C. J., Mavrodiev P., Perony N. (2014). The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy. *Journal of the Royal Society Interface*, 11 (99), 20140623.

Garcia D., Schweitzer F. (2012). Modeling online collective emotions. In: *Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media*, 37–38. ACM.

Garcia D., Schweitzer F. (2015). Social signals and algorithmic trading of Bitcoin. *Royal Society open science*, 2 (9), 150288.

Geraskin P., Fantazzini D. (2013). Everything you always wanted to know about log-periodic power laws for bubble modeling but were afraid to ask. *The European Journal of Finance*, 19 (5), 366–391.

Gonzalo J., Granger C. (1995). Estimation of common long-memory components in cointegratedsystems. *Journal of Business and Economic Statistics*, 13 (1), 27–35.

Harvey A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Hasbrouck J. (1995). One security, many markets: determining the contributions to price discovery. *Journal of Finance*, 50, 1175–1199.

Hayes A. (2015b). Cryptocurrency value formation: An empirical analysis leading to a cost of production model for valuing bitcoin. *Telematics and Informatics*, forthcoming. https://papers.ssrn.com/sol3/papers2.cfm?abstract_id=2648366.

Hayes A. (2015c). The decision to produce altcoins: Miners' arbitrage in cryptocurrency markets. *Working Paper 05/2015*. Department of Economics, The New School for Social Research. https://papers.ssrn.com/sol3/papers2.cfm?abstract_id=2579448.

Jiang Z. Q., Zhou W. X., Sornette D., Woodard R., Bastiaensen K., Cauwels P. (2010). Bubble diagnosis and prediction of the 2005–2007 and 2008–2009 Chinese stock market bubbles. *Journal of Economic Behavior and Organization*, 74 (3), 149–162.

Johansen A. (2003). Characterization of large price variations in financial markets. *Physica A*, 324, 157–166.

Johansen A., Ledoit O., Sornette D. (2000). Crashes as critical points. *International Journal of Theoretical and Applied Finance*, 3 (2), 219–255.

Kancs D. A., Ciaian P., Miroslava R. (2015). The digital agenda of virtual currencies. Can bitcoin become a global currency? Joint Research Centre Technical Report JRC97043. Institute for Prospective and Technological Studies.

Konrad A. (2013). Feds say they've arrested 'dread pirate Roberts', shut down his black market 'The Silk Road'. *Forbes*. http://www.forbes.com/sites/alexkonrad/2013/10/02/feds-shut-down-silk-road-owner-known-as-dread-pirate-roberts-arrested.

Kristoufek L. (2013). Bitcoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific reports*, 3, Article number: 3415.

Kristoufek L. (2015). What are the main drivers of the bitcoin price? Evidence from wavelet coherence analysis. *Plos ONE*, 10 (4), e0123923.

Kwiatkowski D., Phillips P., Schmidt P., Shin Y. (1992). Testing the null hypothesis of stationary against the alternative of a unit root. *Journal of Econometrics*, 54, 159–178.

Lin Y., Michel J. B., Aiden E. L., Orwant J., Brockman W., Petrov S. (2012). Syntactic annotations for the Google Books ngram corpus. In: *Proceedings of the ACL 2012 system demonstrations*, 169–174. Association for Computational Linguistics.

Lin L., Ren R. E., Sornette D. (2014). The volatility-confined LPPL model: A consistent model of 'explosive' financial bubbles with mean-reverting residuals. *International Review of Financial Analysis*, 33, 210–225.

MacDonell A. (2014). Popping the bitcoin bubble: An application of log-periodic power law modeling to digital currency. https://economics.nd.edu/assets/134206/mac_donell_popping_the_biticoin_bubble_an_application_of_log_periodic_power_law_modeling_to_digital_currency.pdf.

Malhotra A., Maloo M. (2014). Bitcoin — is it a bubble? Evidence from unit root tests. http://ssrn.com/abstract=2476378.

Ng E. K., Chan J. C. (2012). Geophysical applications of partial wavelet coherence and multiple wavelet coherence. *Journal of Atmospheric and Oceanic Technology*, 29 (12), 1845–1853.

Osgood C. E. (1964). Semantic differential technique in the comparative study of Cultures. *American Anthropologist*, 66 (3), 171–200.

Pennebaker J. W., Chung C. K., Ireland M., Gonzales A., Booth R. J. (2007). The development and psychometric properties of LIWC2007. https://www.researchgate.net/publication/228650445_The_Development_and_Psychometric_Properties_of_LIWC2007.

Pesaran M. H., Shin Y. (1999). An autoregressive distributed-lag modelling approach to cointegration analysis. *Econometric Society Monographs*, 31, 371–413.

Phillips P. C. B., Yu J. (2011). Dating the timeline of financial bubbles during the subprime crisis. *Quantitative Economics*, 2 (3), 455–491.

Phillips P. C. B., Shi S., Yu J. (2015). Testing for multiple bubbles: Historical episodes of exuberance and collapse in the SP500. *International Economic Review*, 56 (4), 1043–1078.

Phillips P. C. B., Wu Y., Yu J. (2011). Explosive behavior in the 1990S Nasdaq: When did exuberance escalate asset values? *International Economic Review*, 52 (1), 201–226.

Russell J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110 (1), 145–172.

Sornette D. (2003a). *Why stock markets crash* (*Critical events in complex financial systems*). Princeton University Press.

Sornette D. (2003b). Critical market crashes. *Physics Reports*, 378 (1), 1–98.

Sornette D., Johansen A. (2001). Significance of log-periodic precursors to financial crashes. *Quantitative Finance*, 1 (4), 452–471.

Sornette D., Zhou W. X. (2006). Predictability of large future changes in major financial indices. *International Journal of Forecasting*, 22, 153–168.

Sornette D., Woodard R., Zhou W. X. (2009). The 2006–2008 oil bubble: Evidence of speculation, and prediction. *Physica A*, 388, 1571–1576.

Terasvirta T., Tjostheim D., Granger C. (2011). *Modelling nonlinear economic time series*. Oxford University Press, Amsterdam.

Tong H. (1990). *Non-linear time series: A dynamical system approach*. Oxford University Press, Oxford.

Tumarkin R., Whitelaw R. F. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal*, 57 (3), 41–51.

Van Bothmer H. C. G., Meister C. (2003). Predicting critical crashes? A new restriction for the free variables, Physica A 320, 539–547.

Warriner A. B., Kuperman V., Brysbaert M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45 (4), 1191–1207.

Weiß G. N. (2011). Are Copula-GoF-tests of any practical use? Empirical evidence for stocks, commodities and FX futures. *The Quarterly Review of Economics and Finance*, 51 (2), 173–188.

Weiß G. N. (2013). Copula-GARCH versus dynamic conditional correlation: An empirical study on VaR and ES forecasting accuracy. *Review of Quantitative Finance and Accounting*, 41 (2), 179–202.

Wood S. (2006). *Generalized additive models: An introduction with R*. Chapman and Hall / CRC, Boca Raton.

Yermack D. (2013). Is bitcoin a real currency? An economic appraisal. *Working paper* No. w19747, National Bureau of Economic Research.

Zhao Y. (2015). Robustness of forecast combination in unstable environment: A Monte Carlo study of advanced algorithms. *Working paper* No. 2015–04, Towson University, Maryland.

Zhou W. X., Sornette D. (2003). 2000–2003 real estate bubble in the UK but not in the USA. *Physica A*, 329, 249–263.

Zhou W. X., Sornette D. (2006). Is there a real-estate bubble in the US? *Physica A*, 361, 297–308.

Zhou W. X., Sornette D. (2008). Analysis of the real estate market in Las Vegas: Bubble, seasonal patterns, and prediction of the CSW indexes. *Physica A*, 387, 243–260.

Zhou W. X., Sornette D. (2009). A case study of speculative financial bubbles in the South African stock market 2003–2006. *Physica A*, 388, 869–880.

Received 01.06.2016; accepted 25.09.2016.

Appendix 16 "Big Data for computing social well-being indices of the Russian population."

Fantazzini, D., Shakleina, M., Yuras, N. (2018). Big Data for computing social well-being indices of the Russian population. *Applied Econometrics*, 50(2), 43-66. http://pe.cemi.rssi.ru/pe_2018_50_043-066.pdf

Permisson to post the Publisher's version/PDF: according to http://appliedeconometrics.cemi.rssi.ru/AppEc.files/contents_en.html, "*full texts are available for downloading with approximately one year lag*". In some cases, this 1-year lag can be shorter.

Прикладная эконометрика, 2018, т. 50, с. 43–66.
Applied Econometrics, 2018, v. 50, pp. 43–66.

Д. Фантаццини, М. В. Шаклеина, Н. А. Юрас[1]

Big Data в определении социального самочувствия населения России

В статье строятся индексы поисковых запросов на основе Google Trends Data с целью возможности предсказания динамики российских индексов социального самочувствия ВЦИОМ. Индексы Google рассчитаны с помощью факторного анализа на основе набора данных Google Trends Data за 2006–2016 гг., который содержит результаты 512 поисковых запросов, касающихся жилищных условий, дохода, образования и др. Модель байесовского усреднения (Bayesian model averaging) была использована для отбора индексов (категорий) поисковых запросов Google, сильно коррелированных с индикаторами социального самочувствия ВЦИОМ. Дополнительные модели множественной регрессии и построенные прогнозы подтвердили результаты модели байесовского усреднения. Построенные на основе данных поисковых запросов индексы субъективного благосостояния Google являются статистически надежными, о чем свидетельствует сильная корреляция между наблюдаемыми и прогнозными значениями индексов ВЦИОМ.

Ключевые слова: индексы социального самочувствия ВЦИОМ; большие данные; факторный анализ; модель байесовского усреднения.

JEL classification: C52; I32.

1. Введение

В последнее время растет интерес к измерению благосостояния населения и социального самочувствия в силу того, что многие экономические макропоказатели не могут измерить нравственные ценности, дружбу, счастье и пр. В этой связи больше внимание уделяется субъективным способам оценки.

С учетом наличия и широкой доступности данных из социальных сетей и онлайн-поисковых систем, растет число исследований, показывающих, что с помощью социальных медиа можно лучше оценить мнения и поведение людей, чем посредством использования традиционных опросов, в которых респонденты искажают ответы (O'Connor et al., 2010; Ceron et al., 2014; Mavragani, Tsagarakis, 2016; Isotalo et al., 2016; Oliveira et al., 2017). Существует обширная литература по социологии, в которой рассматривается проблема искажения респондентом ответа на поставленный вопрос, начиная с известной работы (Hyman, 1944)

[1] **Фантаццини Деан** — Московская школа экономики МГУ имени М. В. Ломоносова; Fantazzini@mse-msu.ru.
Шаклеина Марина Владиславовна — Московская школа экономики МГУ имени М. В. Ломоносова; m.leshaykina@gmail.com.
Юрас Наталья Анатольевна — Московская школа экономики МГУ имени М. В. Ломоносова; yurasnatasha@gmail.com.

до недавних работ (Preisendörfer, Wolter, 2014) и (Whiteley, 2016). Все эти работы подчеркивают, что предсказательная способность данных социальных сетей и онлайн-поиска усиливается по мере того, как число граждан, выражающих свое мнение через интернет, увеличивается с течением времени. Более того, Cody et al. (2016) обнаружили, что «не только данные Twitter могут предсказывать ответы на опросы, но мы можем использовать отдельные слова в твиттах, чтобы определить, почему один период времени более счастливый, чем другой, что невозможно сделать в традиционных опросах». Polykalas et al. (2013) утверждают, что «прогнозы на основе веб-поиска могут вскоре конкурировать с традиционными опросами, несмотря на то что большие данные содержат меньше демографической информации по сравнению с традиционными опросами». Тем не менее, они также подчеркнули, что последствия социальных событий, не связанных с темой исследования, должны быть отфильтрованы как шум, чтобы получить надежную оценку параметров модели. Аналогичным образом, Beauchamp (2017) подчеркнул необходимость фильтрации данных и последующей корректировки модели «…для обеспечения корректности в измерении, экстраполяции, интерполяции репрезентативной вариации опроса как между штатами, так и во времени». Поэтому предлагаемая в настоящей работе методология и выбор эмпирических данных соответствует современному исследовательскому подходу.

В России для оценки благосостояния широко используются индексы социального самочувствия, которые получает Всероссийский центр изучения общественного мнения (ВЦИОМ). Данные индексы отображают ответы респондентов на вопросы касательно личной и общественной жизни и являются характеристикой адаптации населения к условиям жизни в России, что помогает в построении общего вектора настроений, направление которого указывает на поведение населения в различных сферах жизни общества.

Расширить и существенно улучшить понимание благосостояния населения могут данные Google, т. к. обладают рядом важных преимуществ по сравнению со стандартными методами, используемыми для измерения благосостояния. Во-первых, они позволяют исследовать поведение людей, что более информативно по сравнению с ответами, получаемыми из социологического опроса: анализ поисковых запросов в интернете позволяет лучше раскрыть интересы и предпочтения населения. Во-вторых, данные Google очень своевременны, и это представляет особый интерес для органов власти, участвующих в разработке и проведении социально-экономической политики. Кроме того, они доступны на местном уровне и могут охватывать чрезвычайно широкий круг вопросов, что не относится к стандартным показателям благосостояния. И последнее, но не менее важное: они доступны абсолютно для всех.

Цель данной статьи — показать возможности использования данных Google для объяснения и предсказания динамики российских индексов социального самочувствия ВЦИОМ.

2. Информационное обеспечение исследования

2.1. Индексы социального самочувствия ВЦИОМ

В России для характеристики благосостояния населения с субъективной стороны применяют индикаторы общественного мнения — индексы социального самочувствия ВЦИОМ. Социальное самочувствие отслеживается агентством с 2005 г. на ежемесячной основе через проведение экспресс-опросов. Опросы проводятся по репрезентативной общероссийской

выборке: учитываются квоты по полу, возрасту, образованию и населенному пункту. Территориальный охват составляет 42 субъекта РФ и 130 населенных пунктов России. Число респондентов в каждом опросе составляет 1600 человек.

К индексам социального самочувствия ВЦИОМ относят шесть индексов:

- индекс удовлетворенности жизнью;
- индекс социального оптимизма;
- индекс материального положения;
- индекс экономического положения страны;
- индекс политической обстановки;
- индекс общего вектора развития страны.

Рассмотрим на примере индекса удовлетворенности жизнью технику построения индексов социального самочувствия. Для оценки данного индекса респондентам задают один вопрос: «Если говорить в целом, то в какой мере Вас устраивает или не устраивает сейчас жизнь, которую Вы ведете?». Варианты ответов: «Вполне устраивает, по большей части устраивает», «Отчасти устраивает, отчасти нет», «Совершенно не устраивает, по большей части не устраивает», «Затрудняюсь ответить». Далее суммируется число положительных и отрицательных ответов и находится их разность:

$$J_ч = \Sigma_{\text{положительных_и_средних_оценок}} - \Sigma_{\text{отрицательных_оценок}}, \quad (1)$$

где Σ — сумма соответствующих оценок.

Диапазон значений каждого индекса колеблется от –100 до 100 пунктов. При положительном результате позитивные оценки доминируют над негативными. Например, значение индекса, равное 60, показывает, что позитивные оценки доминировали над негативными на 60 пунктов, а значит, в обществе преобладает позитивный настрой касательно исследуемого аспекта социального самочувствия.

К индексам общественных настроений, показывающим восприятие россиянами личной жизни, относятся следующие (см. рис. 1).

1. *Индекс удовлетворенности жизнью («В какой мере Вас устраивает сейчас жизнь, которую Вы ведете?»)* показывает, насколько россияне удовлетворены жизнью, которую ведут. Высокое значение индекса указывает, что респонденты удовлетворены образом жизни.

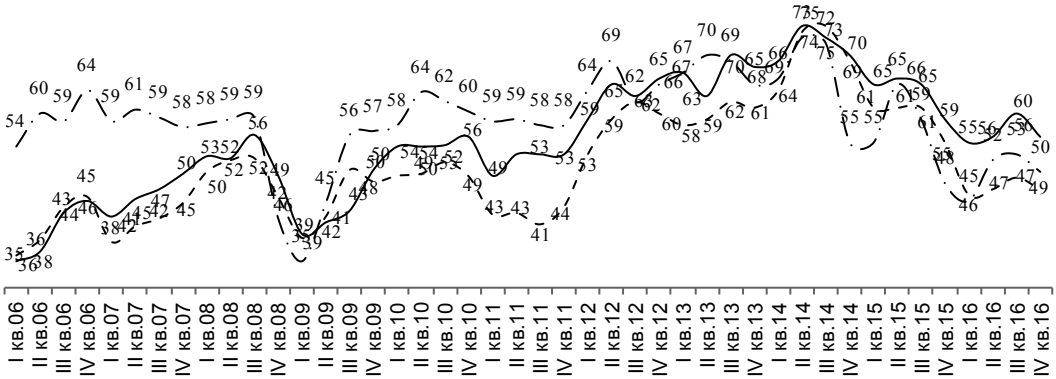
2. *Индекс социального оптимизма («Как Вы считаете, через год Вы (Ваша семья) будете жить лучше или хуже, чем сейчас?»)* показывает, насколько оптимистично россияне смотрят в будущее. Чем выше значение индекса, тем более оптимистично настроены респонденты.

3. *Индекс самооценок материального положения («Как бы Вы оценили в настоящее время материальное положение Вашей семьи?»)* показывает, как россияне оценивают материальное положение своей семьи. Чем выше значение индекса, тем больше респонденты довольны материальным положением семьи.

К индексам положения дел в стране относятся такие (см. рис. 2).

1. *Индекс оценок экономической ситуации («А как бы вы оценили нынешнее экономическое положение России в целом?»)* показывает, как россияне оценивают экономическую ситуацию в стране. Чем выше значение индекса, тем лучше респонденты оценивают экономику.

2. *Индекс оценок политической обстановки («А как бы вы оценили нынешнюю политическую ситуацию в России?»)* показывает, как россияне оценивают политическую обстановку



- - - Индекс «В какой мере Вас устраивает сейчас жизнь, которую Вы ведете?»

— · — Индекс «Как Вы считаете, через год Вы (Ваша семья) будете жить лучше или хуже, чем сейчас?»

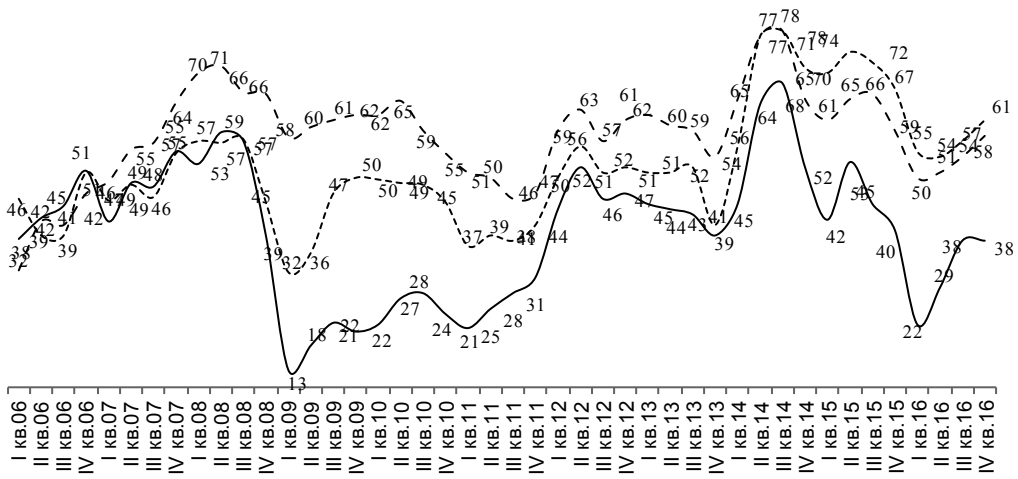
——— Индекс «Как бы Вы оценили в настоящее время материальное положение Вашей семьи?»

Рис. 1. Индексы общественных настроений 2006–2016 годы (в пунктах)

Источник: ВЦИОМ.

в стране. Чем выше значение индекса, тем лучше респонденты оценивают политическую ситуацию.

3. *Индекс оценок общего вектора развития страны («Насколько Вы согласны с тем, что дела в стране идут в правильном направлении?»)* показывает, как россияне оценивают общий курс развития страны. Чем выше значение индекса, тем больше респонденты одобряют направление развития страны.



——— Индекс «А как бы Вы оценили нынешнее экономическое положение России в целом?»

- - - Индекс «Как бы Вы оценили в целом нынешнюю политическую обстановку в России?»

------ Индекс «Насколько Вы согласны с тем, что дела в стране идут в правильном направлении?»

Рис. 2. Индексы положения дел в стране за 2006–2016 гг. (в пунктах)

Источник: ВЦИОМ.

Данные индексы являются отображением ответов респондентов на вопросы касательно личной и общественной жизни, что помогает в построении общего вектора настроений, а его направление указывает на поведение населения в различных сферах жизни общества.

Графики свидетельствуют о том, что социальное самочувствие имеет тренд к снижению в периоды экономического спада (Орлова, 1998; Тощенко, Харченко, 1996). В наибольшей степени отрицательные оценки преобладают в ответах респондентов на вопрос об экономическом положении России. Политическую обстановку респонденты оценивали выше, чем экономическое положение, но после событий 2014 г. началась тенденция к сближению данных индексов. В начале 2016 г. произошло снижение всех индексов, но потом наметился тренд к повышению показателей как для индексов положения дел в стране, так и для индексов социальных настроений. Уровень удовлетворенности жизнью оценивается россиянами положительно, но результаты показывают, что данный уровень намного ниже, чем 2 года назад. В целом это характерно и для оценок материального положения, и для индекса социального оптимизма. По итогам 2016 г. можно сделать вывод о том, что для россиян характерно оптимистическое отношение к жизни, положительная оценка своего самочувствия и положения дел в стране.

2.2. Google Trends Data

Google Trends — сервис, который отображает в виде индекса поисковой активности популярность какого-либо поискового запроса за определенный промежуток времени. Он представляет уникальный взгляд на то, что в данный момент интересует людей, и позволяет отслеживать изменения в предпочтениях и интересах.

Google Trends был создан для отображения фразы или ключевого слова с помощью визуальных и динамических инструментов, которые показывают прошлое, настоящее и потенциально (насколько это можно предсказать) будущее данного запроса (Askitas, 2015).

При выборе слов авторы руководствовались методологией отбора ключевых слов, представленной в работе (Algan et al., 2016). В качестве индексов субъективного благосостояния в ней использовались индексы положительных и отрицательных эмоций из опросов Gallup, которые привязаны к конкретным словам (смех, счастье, уважение, стресс, беспокойство и т. д.). Набор объясняющих индикаторов (ключевых слов) подбирался в поисковой среде Google Trends Data для анализируемых индексов исходя из анализа различных международных опросников: OECD (посвященный нахождению индекса лучшей жизни) и American Time Use Survey (посвященный исследованию использования времени населением). Во внимание принимались ключевые слова, которые потенциально связаны с благосостоянием населения. В результате в (Algan et al., 2016) были выделены следующие аспекты благосостояния: материальные условия, физическое и ментальное здоровье, социальные условия и окружающая среда. В настоящем исследовании в роли индикаторов благосостояния населения выступают индексы социального самочувствия ВЦИОМ. Для исследования авторы отобрали только пять индексов ВЦИОМ, отказавшись от индекса политической обстановки. Прокси-переменными для индексов выступают категории поисковых запросов Google, которые отбирались согласно методологии, предложенной Algan et al. (2016). Ключевые слова для поиска были отобраны из вопросов, которые задавались в рамках исследований, проводимых международными организациями: Gallup World Poll, OECD Better Life Index и NEF. В то же время используемые ключевые слова из опросников международных баз данных

могут быть далеки от повседневной жизни человека, если не характеризуют реальные условия жизни, которые отражаются непосредственно в поисковых запросах, идущих от каждого человека (Algan et al., 2016). В этой связи были добавлены такие ключевые слова, как «поиск работы», «стресс», «цены на жилье», «воспитание ребенка», «репетиторы» и пр. Данный этап отбора слов имеет ряд недостатков, т. к. сопряжен с высокой долей субъективизма.

В результате среди множества поисковых запросов, характеризующих тот или иной аспект социального самочувствия, отбирались только те, которые имеют динамику (видимые изменения) за последние 10 лет. Исходная база из поисковых запросов содержала 512 слов.

Используемый набор данных (поисковых запросов) состоит из квартальных данных, охватывающих промежуток времени 2009–2016 гг. (8 лет), в то время как Algan et al. (2016) использовали набор данных 300 еженедельных наблюдений с 6 января 2008 г. по 4 января 2014 г. (6 лет). Наши оценки могут быть менее эффективными, чем оценки Algan et al. (2016), но для данной задачи продолжительность процесса важнее, чем частота выборки. Например, рассмотрим геометрическое броуновское движение (GBM) и оценку параметра сноса μ с n дискретными данными за временной интервал $[t-h, t]$. Отклонение оценки максимального правдоподобия (ML) сноса μ (т. е. среднего значения выборки дискретных данных) зависит только от длины временного интервала h, а не от частоты дискретизации, в то время как только смещение второго порядка, присутствующее в оценке дисперсии, зависит от n и есть $O(n^{-1})$. Таким образом, широкий диапазон времени с меньшей частотой может быть более эффективным, чем узкий временной интервал с более высокой частотой (Asmussen, Glynn, 2007; Iacus, 2009). Учитывая, что временной диапазон нашего набора данных шире, чем в (Algan et al., 2016), этот факт должен хотя бы частично компенсировать потери эффективности из-за более низкой частоты нашего набора данных.

2.3. Основные проблемы больших данных Google Trends

Данные по поисковому запросу из Google Trends не являются точной характеристикой только данного слова, скорее, это доля от общего числа поисков за определенный период, включающих это ключевое слово, нормированная так, что наибольший объем за период равен 100. В этой связи значения ряда в любой взятый день нельзя сравнивать между поисковыми запросами, т. к. они нормированы к максимальному значению, которое прикреплено к каждому слову. Для решения данной проблемы применяется стандартизованная Z-оценка:

$$Z = \frac{X - \mathrm{E}[X]}{\sigma(X)}, \tag{2}$$

где $\mathrm{E}[X]$ — среднее значение случайной величины X, $\sigma(X)$ — ее стандартное отклонение. В результате Z-оценки имеется распределение с нулевым средним и единичной дисперсией: $\mu = 0$, $\sigma = 1$.

Преимущество стандартизации заключается в том, что Z-оценки адаптируют распределения из различных наборов данных к одному масштабу, что в дальнейшем позволяет напрямую сравнивать ранее несопоставимые переменные.

В данных Google Trends могут быть резкие скачки в популярности поискового запроса (см. Приложение 1). Это создает сложность для оценки, т. к. имеется риск переобучения

(overfitting) при построении. Для того чтобы устранить эту проблему, применяется скользящее среднее. Порядок скользящего среднего определяется числом учитываемых в модели предыдущих значений случайных отклонений.

В настоящем исследовании данные будут сглажены путем трехпериодного скользящего среднего, где период — месяц. Таким образом, каждая точка ряда будет сглажена с учетом среднего значения и с помощью создания составных показателей категорий, чтобы ослабить влияние шока.

Также имеются поисковые запросы, которые в течение продолжительного времени имеют нулевой объем поиска. Такие периоды с большим количеством нулей по своей сути имеют проблемы, схожие с резкими скачками в популярности слова. В результате слова с большим количеством нулей за период исключаются.

Многие поисковые термины имеют сильный временной тренд. Наличие временного тренда в данных является преградой для качественной оценки: во-первых, устойчивый временной тренд может привести к неправильному и нереалистичному прогнозу субъективного благосостояния в будущем, во-вторых, цель данного исследования состоит в составлении категорий слов, которые являются статистически надежными, и поэтому имеется риск в группировке слов, имеющих общий временной тренд, но не отвечающих общей смысловой нагрузке.

Для решения данной проблемы используется скорректированный коэффициент детерминации при построении регрессий с временным трендом. С помощью включения в модель трендовой составляющей можно оценить, насколько поисковый запрос зависит от тренда. В нашем случае все слова, имеющие R^2_{adj} больше, чем 0.6, удаляются. Такой подход использован потому, что невозможно понять, что движет человеком при выполнении запроса: необходимость или влияние внешних факторов, т. к. в лучшем случае важно учитывать первый вариант, но его невозможно отличить от второго. На данном этапе может быть потеряна важная информация, но в действительности путаница в способе использовании интернета является достаточно серьезной.

Многие поисковые запросы демонстрируют экстремальную сезонность. Это может привести к проблемам, поскольку поисковые запросы будут коррелировать только из-за соответствия одной и той же сезонной схеме. Для решения данной проблемы и устранения сезонных эффектов используются месячные фиктивные переменные.

3. Методологические основы исследования

3.1. Построение индексов благосостояния на основе Google Trends Data

Построение индексов (категорий) благосостояния осуществляется на основе группировки поисковых запросов, которая должна соответствовать логической схеме и иметь схожую смысловую нагрузку. Например, невозможно объединить такие слова, как «стресс» и «музеи», в одну категорию, поэтому для упрощения слова были заранее поделены на общие категории, такие как «работа», «семья» и т. д.

Статистическое обоснование полученных категорий поисковых запросов осуществляется посредством проведения факторного анализа, главной целью которого является объединение поисковых запросов в соответствующие категории (индексы) Google, характеризующие

тот или иной аспект социального самочувствия на основе факторных нагрузок. Факторный анализ используется как метод обнаружения взаимосвязей между значениями переменных посредством исследования структуры ковариационных и корреляционных матриц. С помощью выявления факторов (латентных переменных), показывающих наличие линейных связей между переменными, можно построить категории поисковых слов и обнаружить запросы, которые не согласуются ни с одним построенным фактором. Структура факторов достаточно простая и наглядная, т. к. при анализе переменные, сильно коррелирующие между собой, объединяются в один фактор, и затем с помощью перераспределения дисперсии между компонентами получаются факторные нагрузки, на основе которых судят о согласованности поисковых запросов. Использование факторного анализа необходимо для создания составных категорий поисковых слов, что существенно сократит число потенциально-объясняющих переменных.

Для факторного анализа используются слова, которые были нормализованы с помощью Z-оценки и в которых были устранены сезонные колебания.

При проведении факторного анализа поисковые запросы исключаются, если коэффициент факторной нагрузки отрицательный или менее 0.3 (Algan et al., 2016). Многие поисковые запросы не используются, поскольку они не вписываются в какую-либо категорию слов. Неподходящие слова могут разделяться на отдельные категории слов. Так, например, «Здоровье» делится на две составляющие подкатегории: «Состояние здоровья» и «Поддержка тонуса и здоровые привычки». Важно отметить, что если два слова сгруппированы в одну категорию, то это не ведет к тому, что они означают одно и то же, а только то, что им свойственна общая тенденция запросов на данном промежутке времени. В результате после применения скользящего среднего и удаления тренда используются 382 слова. Факторные нагрузки для слов, прошедших проверку, были сгруппированы в категории, которые отображают три аспекта жизни (табл. 1).

Компоненты слов позволяют наглядно оценить, что из себя представляет каждая категория субъективного благосостояния. Так, например, *экономика* страны представлена отдельно от *финансовой безопасности*, хотя они имеют схожие характеристики, но группы поисковых запросов свидетельствуют о том, что экономика страны представляет из себя все то, что является инструментами фискальной и монетарной политики в вопросах регулирования экономики, которые имеют прямое отношение к жизни людей и отражает состояние экономики. Финансовая безопасность выражена в том, какие шоки может испытывать экономика, какие факторы влияют на финансовую безопасность, как это отражается на населении, и чем они обеспокоены при таком положении дел в стране. На графиках в Приложении 1 представлено изменение во времени категорий поисковых запросов Google.

3.2. Модель байесовского усреднения

В настоящем исследовании используется байесовская модель для определения категорий поисковых запросов Google, которые в наибольшей степени коррелируют с индексами социального самочувствия ВЦИОМ и впоследствии могут стать надежными предикторами для объяснения и прогнозирования индексов социального самочувствия ВЦИОМ.

Модель байесовского усреднения, Bayesian model averaging (BMA), направлена на отбор наиболее надежных (робастных) объясняющих переменных (Scott, Varian, 2015).

Таблица 1. Компоненты категорий субъективного благосостояния

Аспекты жизни	Категории	Группы поисковых запросов
Материальные условия	Экономика страны	Цены на жилье, подоходный налог, аренда квартиры, уровень инфляции, экономический рост, налоги, ставка рефинансирования, государственный бюджет, финансовые риски, глобализация, налог на прибыль, свой бизнес, прибыль
	Финансовая безопасность	Экономический кризис, дефолт, банкротство, цены на нефть, курс валюты, девальвация, как заработать деньги, пособие по безработице, служба занятости, увольнение, фриланс
	Рынок труда и поиск работы	Рынок труда, биржа труда, работа, ярмарка вакансий, профессия, производственная практика, условия труда, гибкий график, трудовые отношения, МРОТ, уровень безработицы
Физическое и ментальное здоровье	Состояние здоровья	Стресс, астма, кровяное давление, медицинские услуги, гипертония, ожирение, хроническая усталость, плохое здоровье
	Поддержка тонуса и здоровые привычки	Диета, диета для похудения, здоровый образ жизни, гимнастика, бег, прыжки, фитнес клуб, медитация, база отдыха, отпуск, волейбол
	Образование	Наука, учеба, аспирантура, гуманитарные науки, естественные науки, скорость чтения, грамоты, культура речи, репетиторы, дистанционные курсы, дошкольное образование, стажировки, тайм-менеджмент
	Идеалы и ценности	Этика, религия, расизм, эстетика, свобода слова, человеческий капитал, духовность, свобода, атеизм, материализм, смысл жизни
	Отдых и времяпрепровождение	Рыбалка, активный отдых, экскурсии, отдых на море, спорт, музеи, боулинг, пансионат, квесты, концерт, спектакли, пикник
Социальные условия и окружающая среда	Семья	Детское питание, воспитание ребенка, частные детские сады, уход за ребенком, уход за новорожденным, детские развлечения, творчество, художественная литература, аквапарки, грудное вскармливание
	Семейные проблемы	Как развестись, опека, насилие, поддержка семьи, нападения
	Личная безопасность	Оружие, агрессия, пожарная безопасность, техника безопасности, безопасность жизнедеятельности, кражи, уровень преступности, самозащита, самооборона, самоубийство, социальная защита
	Политическая обстановка и действия государства	Социальная поддержка, демократия, конституция, нормативные акты, транспорт, терроризм, партия, бездомные, иностранцы, взятки, бюрократия, уровень жизни, социальное страхование, социальное обеспечение
	Экология	Охрана природы, природные ресурсы, устойчивое развитие, качество воды, загрязнение окружающей среды, экологические проблемы, загрязнение атмосферы, природа, климат, парниковый эффект, переработка мусора, экологическое воспитание, лесные ресурсы

Пусть X — это множество доступных категорий (регрессоров) и $X_\gamma \in X$ — рассматриваемое подмножество. BMA подход состоит из оценки большого числа моделей M_γ по следующему принципу:

$$y = \alpha_\gamma + \beta_\gamma X_\gamma + \varepsilon, \qquad \varepsilon \sim N\left(0, \sigma^2\right), \tag{3}$$

где y — зависимая переменная (категории социального самочувствия), α_γ — константа, β_γ — коэффициент, X_γ — рассматриваемое подмножество регрессоров.

Когда имеется большое число потенциальных объясняющих переменных в матрице X, сложно оценить, какие переменные $X_\gamma \in X$ должны быть включены в модель и насколько эти переменные важны.

BMA решает проблему с помощью оценки моделей со всеми возможными комбинациями: если X содержит K потенциальных переменных, то это означает оценку 2^K комбинаций переменных, и в итоге мы получаем 2^K моделей.

Оценка модели M_γ с регрессорами X_γ, которые будут выражать категории поисковых запросов, может быть сделана с помощью апостериорной модели вероятностей (posterior model probability, PMP), вытекающей из байесовской теории, где по мере поступления исходных статистических данных распределение уточняется и переходит от априорного распределения к апостериорному:

$$PMP_y = p\left(M_y \middle| y, X\right) = \frac{p\left(y \middle| M_y, X\right) p\left(M_y\right)}{p\left(y|X\right)}, \tag{4}$$

где $p\left(y|X\right)$ не зависит от рассматриваемой модели и является константой, $p\left(M_y\right)$ — априорное распределение, $p\left(y \middle| M_y, X\right)$ — предельное правдоподобие.

Из данного выражения необходимо найти распределение вероятностей параметров модели после того, как приняты во внимание данные — это PMP_y, которое выражается через произведение предельного правдоподобия $p\left(y \middle| M_y, X\right)$ и априорной вероятности модели $p\left(M_y\right)$, а $p\left(y|X\right)$ является константой. Предельное правдоподобие — это вероятность данных при условии фиксированных параметров модели. Априорная вероятность является формализацией нашей интуиции о предмете, т. е. знания, которое имелось еще до проведения анализа, эксперимента. Она определяется исследователем и самый приемлемый вариант — взять равномерную априорную вероятность для каждой модели $p\left(M_y\right) \propto 1$, поскольку имеется недостаток в априорной информации (Zeugner, 2011).

Далее, полученные PMP всех моделей, куда была включена какая-либо переменная (в нашем случае категории поисковых запросов), суммируются и используются для вывода апостериорных включений вероятностей (posterior inclusion probability, PIP), анализ которых необходим, чтобы оценить, насколько переменные важны для описания данных, и отразить корреляционную связь.

В работе (Algan, 2016) наиболее надежные категории были отобраны, если апостериорные включения вероятностей (PIP) больше, чем 0.7. В настоящем исследовании будет использоваться такой же критерий. Категории поисковых запросов Google будут являться детерминантами социального самочувствия, если PIP входит в промежуток (0.7, 1).

Результаты, полученные после применения байесовского метода, представлены в Приложении 3.

Стоит отметить, что BMA не подсчитывает предельную вероятность и апостериорное распределение для каждой возможной модели, напротив, она идентифицирует подмножество моделей, где сосредоточен наибольший вес апостериорной вероятности. Это достигается с помощью алгоритма Метрополиса–Гастингса, который, просматривая участки модели с помощью случайного блуждания, находит наиболее важные части. Этот алгоритм был предложен одновременно с методом Монте-Карло (Metropolis, Ulam, 1949), а затем развит в работах (Metropolis et. al., 1953; Hastings, 1970).

На практике был использован «birth-death» процесс, который случайно выбирает одну из *K* представленных переменных на каждом шагу алгоритма и на следующем шаге отбрасывает переменные из модели, если они уже были включены.

Тем не менее, анализ BMA может быть неэффективен, когда имеется большое количество переменных, сильно коррелирующих друг с другом, в том смысле, что в результате моделирования учитываются надежные регрессоры, но отводится в сторону рассмотрение переменных, не сильно коррелирующих с зависимой переменной. Эта известная проблема связана с проблемой «независимости нерелевантных альтернатив» (independence of irrelevant alternatives) в моделях дискретного выбора и может быть устранена с помощью предварительного факторного анализа для уменьшения числа потенциальных объясняющих переменных.

3.3. Тестирование модели

Модель байесовского усреднения строится на промежутке времени, охватывающем 2009–2016 гг., т. к. по поведению категорий поисковых запросов Google можно было заметить, что некоторые из них становились наиболее важными в кризисные периоды, а скачки в траектории могут привести к риску переобучения в ситуациях, если выбранные категории будут применяться на других выборках. В результате модель строится на данных режима обучения (training set), которые чаще всего используются для нахождения потенциальных связей (т. е. 2009–2016 гг.).

На заключительном этапе моделирования результаты байесовской модели уточняются с помощью регрессионного анализа. В качестве зависимой переменной выступает какой-либо индекс социального самочувствия, а независимых переменных — отобранные с помощью байесовской модели усреднения категории поисковых запросов Google (Приложение 4). Однако следует отметить, что использование байесовской модели усреднения (BMA) в нашем случае является более надежным инструментом моделирования по сравнению с множественной регрессией, учитывая малый размер набора данных (Fernandez et al., 2001; Liang et al., 2008; Claeskens, Hjort, 2008).

Статистическая проверка надежности построенной модели осуществляется путем разбиения заданного набора данных на внутривыборочный период, который используется для первоначальной оценки параметров модели, и вневыборочный период, используемый для оценки качества прогнозирования. Algan et al. (2016) используют центральную часть своего набора данных для оценки модели, а первый и последний год — в качестве вневыборочного периода для проверки качества модели. В нашем исследовании внутривыборочным периодом для построения и оценки модели являются данные с 2010 по 2015 г., в то время как данные за 2009 и 2016 гг. используются как два отдельных вневыборочных периода, на которых тестируется построенная модель.

Обоснованием для определения двух вневыборочных периодов является проверка того, как работают наши индексы Google в период кризиса (например в 2009 г., когда российский ВВП упал почти на 10%), а также в период относительной стабильности (например в 2016 г., когда ВВП упал незначительно — на 0.2%). Поэтому сначала выбираются самые важные категории Google с использованием байесовской модели усреднения, затем строятся регрессии индексов ВЦИОМ по выбранным категориям, а далее удаляется любая категория, которая является статистически незначимой на уровне 5%. Для прогнозирования индексов ВЦИОМ используется только модель со значимыми коэффициентами за период с 2009 по 2016 г.

4. Эмпирические результаты исследования

4.1. Детерминанты индексов социального самочувствия ВЦИОМ, выявленные из корреляций с категориями поисковых запросов Google

На основе байесовской модели усреднения для каждого индекса социального самочувствия ВЦИОМ были определены следующие детерминанты. Курсивом выделены категории поисковых запросов Google, включенные в модель регрессии (PIP больше 0.7).

Индексы положения дел в стране

1. Для оценки экономической ситуации в России важными являются: *рынок труда и поиск работы*, *отдых и времяпрепровождение*, *личная безопасность*, образование, политическая обстановка и действия государства, поддержка тонуса и здоровые привычки, семья (Приложение 3).

2. Для оценки общего вектора развития страны: *образование*, *состояние здоровья*, *семья*, *личная безопасность*, отдых и времяпрепровождение, поддержка тонуса и здоровые привычки, политическая обстановка и действия государства.

Индексы социальных настроений

1. Для оценки материального положения россиянам важны: *рынок труда и поиск работы*, *семья*, *отдых и времяпрепровождение*, *личная безопасность*, образование, поддержка тонуса и здоровые привычки, идеалы и ценности.

2. Для оценки удовлетворенности жизнью: *семья*, *рынок труда и поиск работы*, *образование*, личная безопасность, политическая обстановка и действия государства, отдых и времяпрепровождение, поддержка тонуса и здоровые привычки.

3. Для оценки социального оптимизма: *рынок труда и поиск работы*, *семья*, *поддержка тонуса и здоровые привычки*, семейные проблемы, финансовая безопасность.

Интересным остается тот факт, что ни одна из таких анализируемых категорий, как «*Экология*» и «*Экономика страны*», не оказывают никакого влияния на индексы социального самочувствия.

Результаты байесовской модели проверялись с помощью множественной регрессии. В итоге было построено шесть уравнений регрессии, в каждом из них в роли зависимой переменной выступал один из индексов социального самочувствия ВЦИОМ, а в качестве независимых — категории поисковых запросов Google, которые были отобраны на предыдущем шаге с помощью байесовской модели усреднения (они представлены в Приложении 3). Регрессионная модель построена на квартальных данных с 2009 по 2016 г. и включает три фиктивные переменные, соответствующие первым трем кварталам. Результаты регрессионного моделирования представлены в Приложении 4. Перейдем к интерпретации полученных результатов.

1. ***Индекс самооценок материального положения.*** Большее число поисковых запросов, связанных с *рынком труда,* сигнализирует об ухудшении материального положения (что вполне понятно). Увеличение запросов о *семье* и *личной безопасности* положительно связано с материальным положением. Следует отметить неоднозначную интерпретацию индекса поисковых запросов «личная безопасность». Данный индекс составлен из запросов, касающихся проблем безопасности и имеющих негативную коннотацию, однако в него вошли и запросы (социальная защита, безопасность жизнедеятельности), рост интереса к которым не означает ухудшение индекса личной безопасности Google. В то же время исследуемый индекс самооценок материального положения имеет тенденцию быть немного выше

в течение первых двух кварталов года, что, вероятно, является следствием большого числа праздников в этот период. Затруднительна интерпретация у коэффициента при переменной «*отдых и времяпрепровождение*», который имеет отрицательное значение. Данная категория постоянно снижается с 2009 до 2014 г., а затем стабилизируется в течение последних трех лет. Возможно, проблемы чартерных авиакомпаний России в последние годы, а также теракты в некоторых крупных курортах (Турции и Египта) могут объяснить отрицательную корреляцию между этой категорией и индексом самооценок материального положения.

2. ***Индекс удовлетворенности жизнью.*** Увеличение запросов, связанных с *семьей*, отражает повышение удовлетворенности жизнью, в то время как большее число запросов, связанных с *рынком труда*, отражает ее снижение. Сезонных эффектов не наблюдается.

3. ***Индекс социального оптимизма.*** Рост запросов, связанных с *семьей*, положительно коррелирует с данным индексом. Однако увеличение времени на запросы, посвященные *рынку труда и поиску работы*, отражает снижение социального оптимизма. Аналогично динамике индекса самооценок материального положения, в первые два квартала значения индекса социального оптимизма выше, что объясняется большим числом праздников в этот период.

4. ***Индекс оценок экономической ситуации.*** Высокая *личная безопасность* предполагает улучшение текущей экономической ситуации, в то время как озабоченность *поиском работы* сигнализирует об ухудшении экономической обстановки. Переменная «*отдых и времяпрепровождение*» имеет отрицательный коэффициент, который, вероятно, подтверждается предыдущей интерпретацией индекса самооценок материального положения.

5. ***Индекс оценок общего вектора развития страны.*** Отсутствие проблем со *здоровьем* и гармония в *семье* сигнализируют о большей уверенности в направлении развития страны. Интересно, что с увеличением запросов, связанных с *образованием*, уменьшается уверенность в направлении развития страны. Это интересная, но достаточно спорная проблема (см., например, (Karlin, 2017; Kirkegaard, Grigoriev, 2017)).

4.2. Прогнозирование индексов социального самочувствия ВЦИОМ и надежность построенной модели

В таблице 2 представлены корреляции между фактическими значениями индексов ВЦИОМ и прогнозируемыми значениями на основе данных Google, как для внутривыборочного, так и для вневыборочного (тестового) периодов. На рисунке 3 показаны прогнозируемые и наблюдаемые значения индексов ВЦИОМ. Корреляции относительно внутривыборочного периода очень высоки, как и ожидалось, и всегда выше 0.9. Эти корреляции остаются довольно высокими также в течение первого вневыборочного периода в 2009 г. со значениями в диапазоне от 0.11 до 0.97, тогда как индексы ВЦИОМ и индексы на основе Google показывают близкие и аналогичные изменения (рис. 3). Другие результаты имеют место для второго периода вне выборки в 2016 г.: корреляции между прогнозируемыми и наблюдаемыми индексами ВЦИОМ всегда отрицательны в диапазоне от –0.29 до –0.93, но есть единственное исключение с положительной корреляцией 0.65. В целом, все индексы благосостояния, основанные на Google, выше индексов ВЦИОМ в 2016 г. В этой связи хочется отметить, что использование месячных данных может улучшить результаты прогнозирования. ВЦИОМ с июля 2017 г. начал публиковать индексы социального самочувствия по месяцам, что может расширить рамки будущих исследований.

Таблица 2. Корреляции между фактическими и прогнозными значениями индексов социального самочувствия ВЦИОМ

Период	Индекс самооценок материального положения	Индекс удовлетворенности жизнью
Внутривыборочный	0.93	0.92
2009	0.66	0.11
2016	–0.29	–0.42
	Индекс оценок экономической ситуации	Индекс оценок общего вектора развития страны
Внутривыборочный	0.95	0.96
2009	0.97	0.43
2016	–0.58	–0.93
	Индекс социального оптимизма	
Внутривыборочный	0.91	
2009	0.89	
2016	0.65	

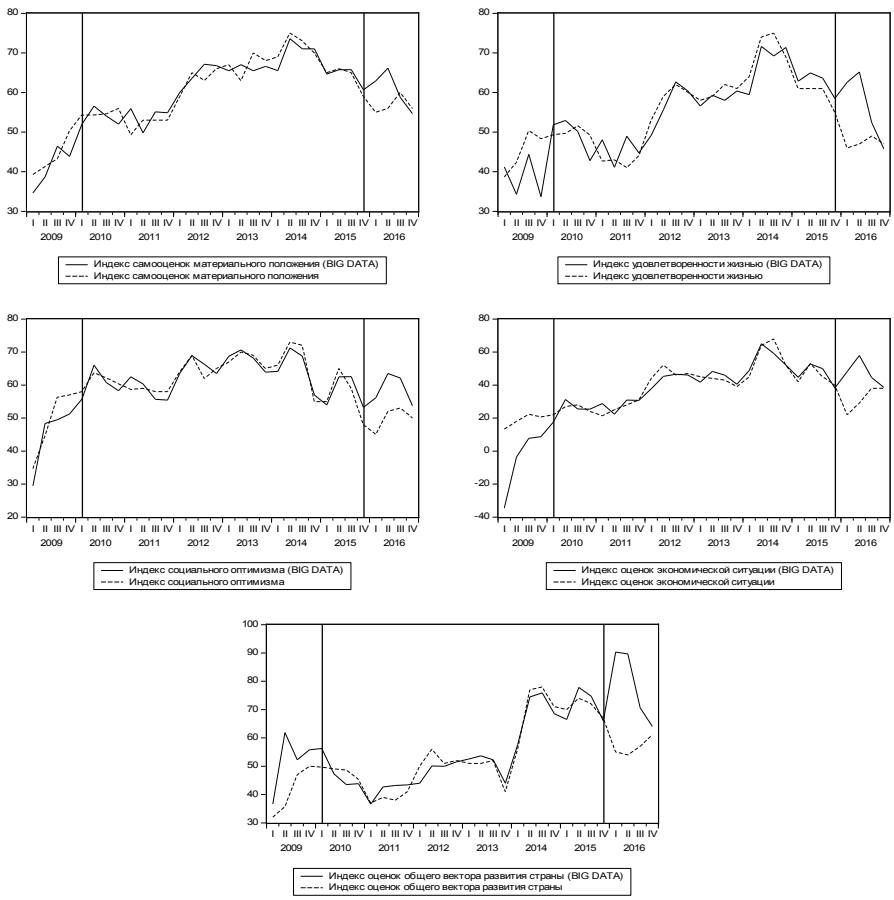


Рис. 3. Прогнозные и фактические значения индексов социального самочувствия ВЦИОМ (внутривыборочный период находится в пределах вертикальных границ)

Плохая точность прогноза для двух последних индексов может быть связана с влиянием экономического кризиса и небольшой выборкой данных, которая была доступна. Вполне возможно, что в действительности социальное самочувствие опрашиваемых было намного выше, чем следует из их ответов интервьюерам ВЦИОМ. Как ранее обсуждалось, расширяются исследования, в которых демонстрируется, что социальные медиа инструменты могут лучше представлять интересы и мнения людей, чем традиционные опросы, которые искажаются из-за ложных отвеов респондентов. Это подчеркивает в своей работе Hyman (1944). Однако для проверки данного предположения в исследовании необходимы гораздо более длительные периоды наблюдения. В этом смысле новые публикации ВЦИОМ, представляющие ежемесячные индексы социального самочувствия, могут, бесспорно, улучшить качество построенных моделей.

5. Заключение

Категории поисковых запросов Google, являющиеся отражением различных аспектов жизни, были использованы как детерминанты социального самочувствия россиян. Для построения категорий этих запросов была реализована следующая последовательность шагов. Во-первых, проанализированы поисковые запросы на наличие скачков в объеме поиска, трендов и нулевого объема поиска; во-вторых, у всех слов устранены сезонные эффекты, после чего они были приведены к Z-оценке; в-третьих, на основе полученных данных проведен факторный анализ для проверки согласованности слов.

На основе модели байесовского усреднения были отобраны категории, которые в наибольшей степени коррелируют с индексами социального самочувствия, а модель множественной регрессии подтвердила полученные результаты.

Индексы Google представляют собой интересный дополнительный инструмент, который может использоваться наряду со стандартными индексами. Найденные детерминанты индексов социального самочувствия ВЦИОМ — категории (индексы) поисковых запросов Google могут быть использованы в определении ключевых направлений в социально-экономической политике, а обеспечение существенно более высокого уровня не только материальных, но и социальных благ позволит улучшить социальное самочувствие и, как следствие, качество жизни.

Построенные на основе данных поисковых запросов категории субъективного благосостояния являются статистически надежными и могут использоваться в дальнейшем анализе, например, для понимания того, что в наибольшей степени определяет социальное самочувствие человека в период экономического кризиса или после реализации каких-либо мер социально-экономической политики.

Список литературы

Орлова Л. А. (1998). О социальном самочувствии учителей Московской области (по результатам социологических опросов). *Социологические исследования*, 8, 89–94.

Тощенко Ж. Т., Харченко С. В. (1996). *Социальное настроение*. М.: Academia.

Algan Y., Beasley E., Guyot F., Higa K., Murtin F., Senik C. (2016). Big data measures of well-being: Evidence from a Google well-being index in the United States. *OECD Statistics Working Papers 2016/3*, 1–38.

Askitas N. (2015). Google search activity data and breaking trends. IZA World of Labor. https://wol.iza.org/articles/google-search-activity-data-and-breaking-trends/long.

Asmussen S., Glynn P. (2007). *Stochastic simulation: Algorithms and analysis* (No. 57). Springer Science and Business Media.

Beauchamp N. (2017). Predicting and interpolating state-level polls using twitter textual data. *American Journal of Political Science*, 61 (2), 490–503.

Claeskens G., Hjort N. (2008). Model selection and model averaging. Cambridge: Cambridge University Press.

Ceron A., Curini L., Iacus S. M., Porro G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media and Society*, 16 (2), 340–358.

Cody E. M., Reagan A. J., Dodds P. S., Danforth C. M. (2016). Public opinion polling with twitter. *arXiv preprint.* arXiv:1608.02024.

Fernandez C., Ley E., Steel M. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100 (2), 381–427.

Hastings W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 (1), 97–109.

Hyman H. (1944). Do they tell the truth? *The Public Opinion Quarterly*, 8 (4), 557–559.

Iacus S. (2009). *Simulation and inference for stochastic differential equations: With R examples*. Springer Science and Business Media.

Isotalo V., Saari P., Paasivaara M., Steineker (2016). Predicting 2016 US presidential election polls with online and media variables. *Designing Networks for Innovation and Improvisation.* Springer International Publishing, 45–53.

Karlin A. (2017). The Moscow municipal elections 2017. http://www.unz.com/akarlin/the-moscow-municipal-elections-2017/.

Kirkegaard O., Grigoriev A. (2017). Moscow districts analysis. http://rpubs.com/EmilOWK/Moscow_districts.

Liang F., Paulo R., Molina G., Clyde M., Berger J. (2008). Mixtures of *g* priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103 (481), 410–423.

Mavragani A., Tsagarakis K. (2016). YES or NO: Predicting the 2015 Greek referendum results using Google Trends. *Technological Forecasting and Social Change*, 109, 1–5.

Metropolis N., Rosenbluth A., Rosenbluth M. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21 (6), 1087–1092.

Metropolis N., Ulam S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44 (247), 335–341.

O'Connor B., Balasubramanyan R., Routledge B. R., Smith N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 122–129.

Oliveira D., Bermejo S., Santos P. (2017). Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology and Politics*, 14 (1), 34–45.

Polykalas S. E., Prezerakos G. N., Konidaris A. (2013). An algorithm based on Google Trends' data for future prediction. Case study: German Elections. In: *Signal Processing and Information Technology*, 69–73, IEEE International Symposium.

Preisendörfer P., Wolter F. (2014). Who is telling the truth? A validation study on determinants of response behavior in surveys. *Public Opinion Quarterly*, 78 (1), 126–146.

Scott S., Varian H. (2015). Bayesian variable selection for nowcasting economic time series. In: *Economic analysis of the digital economy*. Chicago, University of Chicago Press.

Whiteley P. (2016). Why do voters lie to the pollsters? *Political Insight*, 7 (1), 16–19.

Zeugner S. (2011). *Bayesian model averaging with BMS*. Tutorial to the R-package BMS 1e30. https://cran.r-project.org/web/packages/BMS/vignettes/bms.pdf.

Поступила в редакцию 14.10.2017;
принята в печать 08.06.2018.

Приложение 1
Результаты анализа больших данных и первичной обработки

Для построения надежных категорий субъективного благосостояния необходимо отобрать данные, которые являются релевантными и не будут приводить к риску переобучения.

Первым пунктом анализа данных будет рассмотрение поисковых запросов, имеющих резкие скачки в объеме поиска.

Например, скачок в поиске слова «рак» в январе 2014 г. был связан с известием об обнаружении рака головного мозга у певицы Жанны Фриске, что очень сильно взволновало российскую общественность и привело к увеличению запросов по данному слову (рис. 1). Это показывает, что люди были заинтересованы данным известием, но не отражает, насколько данная категория важна для формирования субъективного благосостояния россиян.

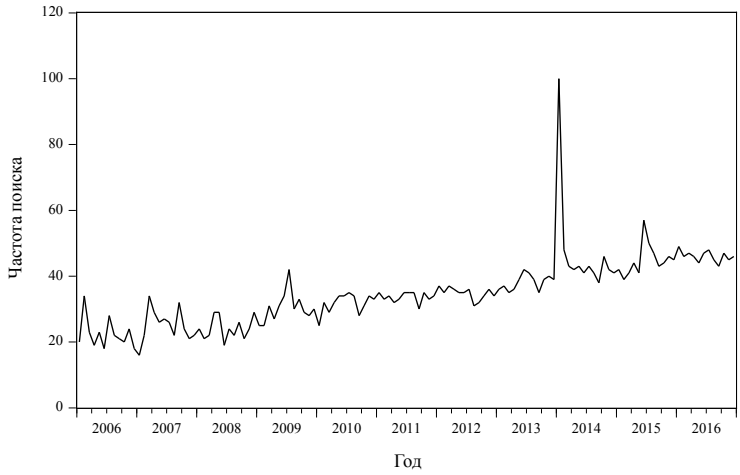


Рис. 1. Частота поиска слова «рак» за 2006–2016 гг. (в пунктах)

Как было отмечено ранее, такие поисковые запросы приводят к риску переобучения, поэтому для решения этой проблемы применяется скользящее среднее третьего порядка — MA(3) (рис. 2).

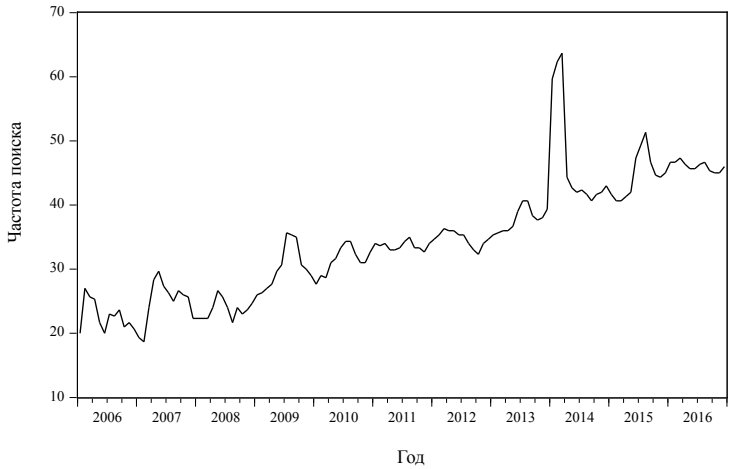


Рис. 2. Частота поиска слова «рак» за 2006–2016 гг. (в пунктах) после применения скользящего среднего

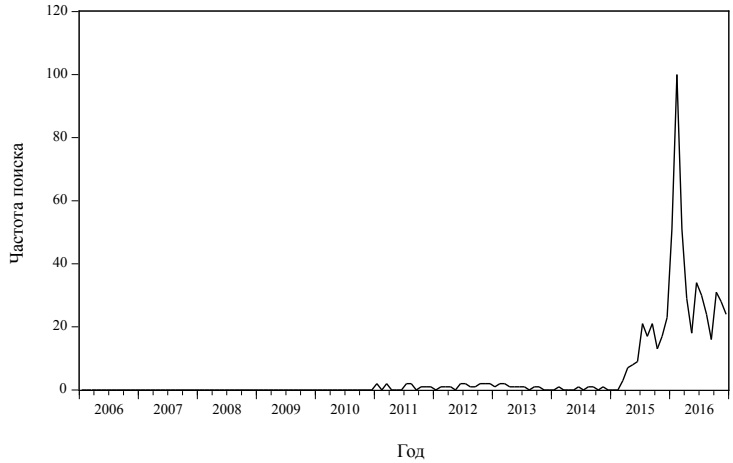


Рис. 3. Частота поиска фразы «ипотека с господдержкой» за 2006–2016 гг. (в пунктах)

В результате после применения скользящего среднего получаем сглаженный временной ряд для поискового запроса. Стоит отметить, что данный запрос имеет временной тренд, который тоже является преградой для построения надежных категорий субъективного благосостояния, что будет отмечено ниже.

Следующая категория слов, которую необходимо рассмотреть — слова, продолжительное время остающиеся без запроса. Такие запросы хорошо работают на одном массиве данных, но плохо объясняют общие закономерности на другом. Пример такого поискового запроса представлен на рис. 3. Данный график иллюстрирует, что в течение 5 лет поисковый запрос «ипотека с господдержкой» имел низкую популярность, в то же время нельзя исключать тот факт, что данное слово играет важную роль в жизни россиян, но в связи с принятой методологией такие запросы в данной статье были удалены.

Далее, в данных были найдены поисковые запросы, которые имеют сильный временной тренд. Рассмотрим пример такого запроса — «купить квартиру» (рис. 4).

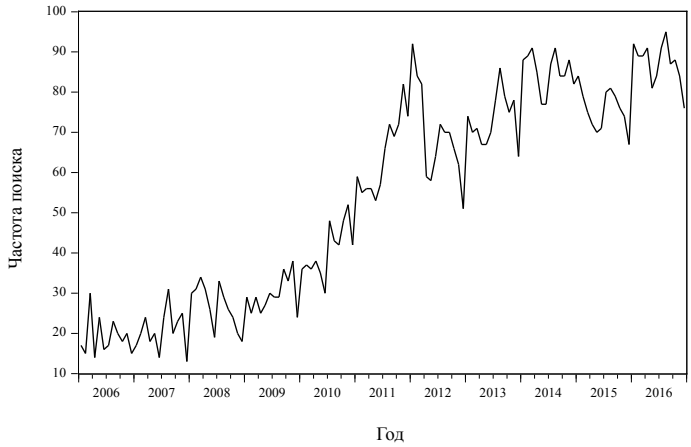


Рис. 4. Частота поиска фразы «купить квартиру» за 2006–2016 гг. (в пунктах)

В первую очередь, может быть так, что покупка квартиры с каждым годом становится более важным и необходимым условием в жизни россиян, вследствие этого наблюдается рост объема поиска в течение 6 лет. Возможен и такой вариант, что люди не чувствуют необходимости покупать квартиру, но обращаются к интернету за советом потому, что произошел какой-либо культурный сдвиг или влияние внешних факторов, как например, рост цен на квартиры в связи с экономическим кризисом, или люди стали пользоваться интернетом намного чаще.

Для устранения этой проблемы проверяем слова на временной тренд с помощью построения регрессии методом наименьших квадратов. Регрессия слова с временным трендом, где $R^2_{adj} > 0.6$, исключаются из дальнейшего анализа.

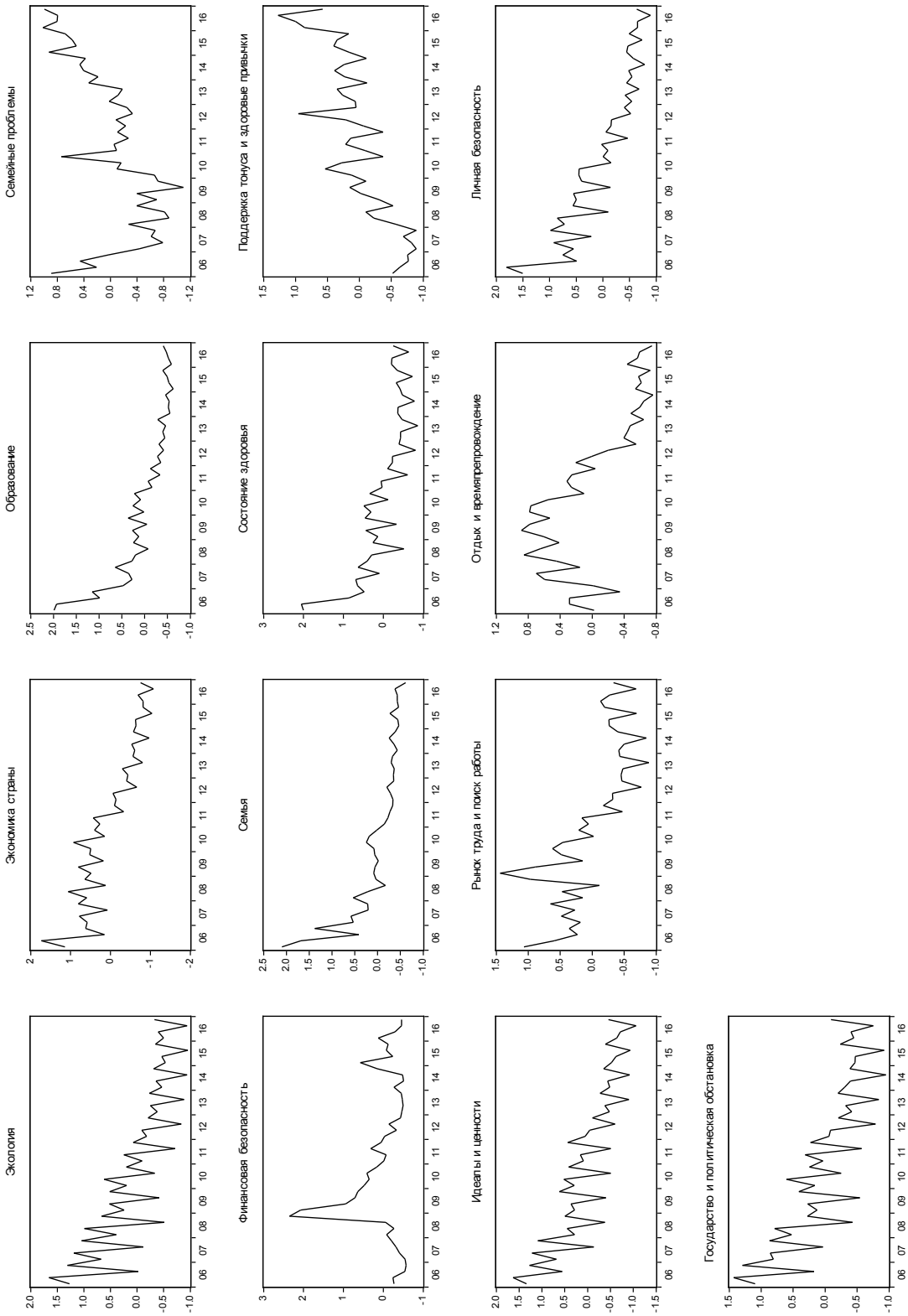
Прежде чем провести факторный анализ, необходимо выполнить два этапа действий: построение Z-оценки и устранение сезонности.

Z-оценка для каждого поискового запроса необходима, чтобы отследить, как изменялся объем поиска именно по данному запросу, а не для оценки того, какие запросы были популярны среди остальных поисковых запросов. Таким образом, все слова были приведены к стандартному нормальному распределению.

Далее, многие поисковые слова имеют экстремальную сезонность, и для того чтобы переменные не коррелировали только потому, что соответствуют одной сезонной схеме, были созданы месячные фиктивные переменные, и с помощью использования месячных остатков регрессий, которые образуют очищенный ряд, устранялась сезонность в данных.

Приложение 2

Графики, иллюстрирующие изменение категорий субъективного благосостояния за 2006–2016 гг.



Приложение 3
Модель байесовского усреднения

Индекс самооценок материального положения Как бы Вы оценили в настоящее время материальное положение Вашей семьи?		Индекс удовлетворенности жизнью В какой мере Вас устраивает сейчас жизнь, которую Вы ведете?		Индекс социального оптимизма Как Вы считаете, через год Вы (Ваша семья) будете жить лучше или хуже, чем сейчас?	
Категория	PIP	Категория	PIP	Категория	PIP
Рынок труда и поиск работы	**0.98**[2]	Семья	**0.99**	Рынок труда и поиск работы	**1.00**
Семья	**0.95**	Рынок труда и поиск работы	**0.80**	Семья	**0.98**
Отдых и времяпрепровождение	**0.90**	Образование	0.68	Поддержка тонуса и здоровые привычки	**0.88**
Личная безопасность	**0.81**	Личная безопасность	0.58	Семейные проблемы	0.54
Образование	0.40	Политическая обстановка и действия государства	0.55	Финансовая безопасность	0.40
Поддержка тонуса и здоровые привычки	0.40	Отдых и времяпрепровождение	0.49	Политическая обстановка и действия государства	0.27
Идеалы и ценности	0.30	Поддержка тонуса и здоровые привычки	0.41	Личная безопасность	0.25
Политическая обстановка и действия государства	0.20	Семейные проблемы	0.28	Экология	0.19
Экономика страны	0.20	Идеалы и ценности	0.26	Состояние здоровья	0.19
Состояние здоровья	0.19	Состояние здоровья	0.23	Образование	0.18
Семейные проблемы	0.17	Финансовая безопасность	0.23	Экономика страны	0.16
Финансовая безопасность	0.16	Экономика страны	0.18	Идеалы и ценности	0.15
Экология	0.14	Экология	0.14	Отдых и времяпрепровождение	0.14

Индекс оценок экономической ситуации А как бы вы оценили нынешнее экономическое положение России в целом?		Индекс оценок общего вектора развития страны Насколько Вы согласны с тем, что дела в стране идут в правильном направлении?	
Категория	PIP	Категория	PIP
Рынок труда и поиск работы	**0.91**	Образование	**0.92**
Отдых и времяпрепровождение	**0.85**	Состояние здоровья	**0.90**
Личная безопасность	**0.83**	Семья	**0.84**
Образование	0.41	Личная безопасность	0.54
Политическая обстановка и действия государства	0.40	Отдых и времяпрепровождение	0.46
Поддержка тонуса и здоровые привычки	0.39	Поддержка тонуса и здоровые привычки	0.36
Семья	0.30	Политическая обстановка и действия государства	0.32
Экология	0.22	Экономика страны	0.23
Идеалы и ценности	0.21	Семейные проблемы	0.20
Финансовая безопасность	0.20	Финансовая безопасность	0.20
Экономика страны	0.20	Рынок труда и поиск работы	0.20
Семейные проблемы	0.20	Идеалы и ценности	0.19
Состояние здоровья	0.17	Экология	0.11

[2] Жирным выделены категории поисковых запросов Google, для которых апостериорные включения вероятностей (PIP) больше, чем 0.7.

Приложение 4
Регрессии категорий субъективного благосостояния ВЦИОМ

		Индекс самооценок материального положения	Индекс удовлетворенности жизнью	Индекс социального оптимизма	Индекс оценок экономической ситуации	Индекс оценок общего вектора развития страны
Материальные условия	Экономика страны					
	Финансовая безопасность					
	Рынок труда и поиск работы	–19.62****	–20.43****	–27.35****	–25.46****	
Физическое и ментальное здоровье	Состояние здоровья					36.30**
	Поддержка тонуса и здоровые привычки			–7.91****		
	Образование					–118.98****
	Идеалы и ценности					
	Отдых и времяпрепровождение	–21.40****			–32.89****	
Социальные условия и окружающая среда	Семья	24.93****	19.27	37.09****		72.45****
	Семейные проблемы					
	Личная безопасность	20.25****			48.60****	
	Политическая обстановка и действия государства					
	Экология					
Константа		62.10****	56.63****	63.65****	34.16****	60.11****
I квартал		6.42**	1.13	4.45**	8.37*	–24.79****
II квартал		4.21**	1.26	8.60****	9.90**	–13.57**
III квартал		1.11	–7.29*	–5.99**	15.04**	–8.59
R^2		0.90	0.50	0.87	0.78	0.68
R^2_{adj}		0.87	0.40	0.84	0.73	0.60
Число наблюдений		32	32	32	32	32

Примечание. p-значения: **** — $p < 0.001$, *** — $p < 0.01$, **— $p < 0.05$, * — $p < 0.1$.

Fantazzini D., Shakleina M., Yuras N. Big Data for computing social well-being indices of the Russian population. *Applied Econometrics*, 2018, v. 50, pp. 43–66.

Dean Fantazzini
Moscow School of Economics — Moscow State University;
Fantazzini@mse-msu.ru

Marina Shakleina
Moscow School of Economics — Moscow State University;
m.leshaykina@gmail.com

Natalia Yuras
Moscow School of Economics — Moscow State University;
yurasnatasha@gmail.com

Big Data for computing social well-being indices of the Russian population

The article builds indices of social well-being based on Google Trends Data for predicting VCIOM indices. The Google indices were computed using a Google Trends dataset for 2006–2016 containing 512 search queries relative to housing conditions, income, education, etc., and applying factor analysis. Bayesian Model Averaging was then used to select the indexes of individual social well-being mostly associated with the VCIOM indices which measure the social well-being of the Russian population. Additional regression models and forecasting exercises confirmed the previous results. Based on the Google Trends Data, the indices of the subjective social well-being are statistically reliable, as evidenced by a strong correlation between the observed and predicted values of the VCIOM indices.

Keywords: social well-being indices; Google Trends Data; Factor analysis; Bayesian Model Averaging

JEL classification: C52; I32.

References

Orlova L. A. (1998). O social'nom samochuvstvii uchitelej Moskovskoj oblasti (po rezul'tatam sociologicheskih oprosov). *Sociological Studies*, 8, 89–94 (in Russian).

Toshchenko Zh. T., Harchenko S. V. (1996). *Social'noe nastroenie*. M.: Academia (in Russian).

Algan Y., Beasley E., Guyot F., Higa K., Murtin F., Senik C. (2016). Big data measures of well-being: Evidence from a Google well-being index in the United States. *OECD Statistics Working Papers 2016/3*, 1–38.

Askitas N. (2015). Google search activity data and breaking trends. IZA World of Labor. https://wol.iza.org/articles/google-search-activity-data-and-breaking-trends/long.

Asmussen S., Glynn P. (2007). *Stochastic simulation: Algorithms and analysis* (No. 57). Springer Science and Business Media.

Beauchamp N. (2017). Predicting and interpolating state-level polls using twitter textual data. *American Journal of Political Science*, 61 (2), 490–503.

Claeskens G., Hjort N. (2008). Model selection and model averaging. Cambridge: Cambridge University Press.

Ceron A., Curini L., Iacus S. M., Porro G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media and Society*, 16 (2), 340–358.

Cody E. M., Reagan A. J., Dodds P. S., Danforth C. M. (2016). Public opinion polling with twitter. *arXiv preprint.* arXiv:1608.02024.

Fernandez C., Ley E., Steel M. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100 (2), 381–427.

Hastings W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 (1), 97–109.

Hyman H. (1944). Do they tell the truth? *The Public Opinion Quarterly*, 8 (4), 557–559.

Iacus S. (2009). *Simulation and inference for stochastic differential equations: With R examples*. Springer Science and Business Media.

Isotalo V., Saari P., Paasivaara M., Steineker (2016). Predicting 2016 US presidential election polls with online and media variables. *Designing Networks for Innovation and Improvisation.* Springer International Publishing, 45–53.

Karlin A. (2017). The Moscow municipal elections 2017. http://www.unz.com/akarlin/the-moscow-municipal-elections-2017/.

Kirkegaard O., Grigoriev A. (2017). Moscow districts analysis. http://rpubs.com/EmilOWK/Moscow_districts.

Liang F., Paulo R., Molina G., Clyde M., Berger J. (2008). Mixtures of *g* priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103 (481), 410–423.

Mavragani A., Tsagarakis K. (2016). YES or NO: Predicting the 2015 Greek referendum results using Google Trends. *Technological Forecasting and Social Change*, 109, 1–5.

Metropolis N., Rosenbluth A., Rosenbluth M. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21 (6), 1087–1092.

Metropolis N., Ulam S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44 (247), 335–341.

O'Connor B., Balasubramanyan R., Routledge B. R., Smith N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 122–129.

Oliveira D., Bermejo S., Santos P. (2017). Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology and Politics*, 14 (1), 34–45.

Polykalas S. E., Prezerakos G. N., Konidaris A. (2013). An algorithm based on Google Trends' data for future prediction. Case study: German Elections. In: *Signal Processing and Information Technology*, 69–73, IEEE International Symposium.

Preisendörfer P., Wolter F. (2014). Who is telling the truth? A validation study on determinants of response behavior in surveys. *Public Opinion Quarterly*, 78 (1), 126–146.

Scott S., Varian H. (2015). Bayesian variable selection for nowcasting economic time series. In: *Economic analysis of the digital economy*. Chicago, University of Chicago Press.

Whiteley P. (2016). Why do voters lie to the pollsters? *Political Insight*, 7 (1), 16–19.

Zeugner S. (2011). *Bayesian model averaging with BMS*. Tutorial to the R-package BMS 1e30. https://cran.r-project.org/web/packages/BMS/vignettes/bms.pdf.

Received 14.10.2017; accepted 08.06.2018.

Appendix 17 "The importance of being informed: forecasting market risk measures for the Russian RTS index future using online data and implied volatility over two decades."

Fantazzini D., Shangina T. (2019) The importance of being informed: forecasting market risk measures for the Russian RTS index future using online data and implied volatility over two decades, *Applied Econometrics* 55, 5-31. http://pe.cemi.rssi.ru/pe_2019_55_005-031.pdf

Permisson to post the Publisher's version/PDF: according to http://appliedeconometrics.cemi.rssi.ru/AppEc.files/contents_en.html, "*full texts are available for downloading with approximately one year lag*". In some cases, this 1-year lag can be shorter.

Прикладная эконометрика, 2019, т. 55, с. 4–31.
Applied Econometrics, 2019, v. 55, pp. 4–31.
DOI: 10.24411/1993-7601-2019-10008

D. Fantazzini, T. Shangina[1]

The importance of being informed: Forecasting market risk measures for the Russian RTS index future using online data and implied volatility over two decades

This paper focuses on the forecasting of market risk measures for the Russian RTS index future, and examines whether augmenting a large class of volatility models with implied volatility and Google Trends data improves the quality of the estimated risk measures. We considered a time sample of daily data from 2006 till 2019, which includes several episodes of large-scale turbulence in the Russian future market. We found that the predictive power of several models did not increase if these two variables were added, but actually decreased. The worst results were obtained when these two variables were added jointly and during periods of high volatility, when parameters estimates became very unstable. Moreover, several models augmented with these variables did not reach numerical convergence. Our empirical evidence shows that, in the case of Russian future markets, TGARCH models with implied volatility and Student's t errors are better choices if robust market risk measures are of concern.

Keywords: forecasting; Value-at-Risk; realized volatility; Google trends; implied volatility; GARCH; ARFIMA; HAR; realized-GARCH.

JEL classification: C22; C51; C53; G17; G32.

1. Introduction

The market risk is usually defined in the financial literature as the gains and losses on the value of a position or portfolio that can take place due to the movements in market variables (asset prices, interest rates, forex rates), see (Basel Committee…, 2009; Hartmann, 2010) and references therein for more details. The most well-known market risk measure is the Value-at-Risk (VaR), which can be defined as the maximum loss over a given time horizon that may be incurred by a position or a portfolio at a given level of confidence. The VaR is recognized by official bodies worldwide as an important market risk measurement tool, see (Jorion, 2007; Basel Committee…, 2013, 2016). The VaR has been criticized for not being sub-additive so that the risk of a portfolio can be larger than the sum of the stand-alone risks of its components, see e.g. (Artzner et al., 1997,

[1] **Fantazzini, Dean** — Moscow School of Economics, Moscow State University; fantazzini@mse-msu.ru.
Shangina, Tamara — ICEF, National Research University Higher School of Economics, Moscow; tbshangina@edu.hse.ru.

1999). For this reason, the Expected Shortfall (ES) has been proposed as an alternative risk measure able to satisfy the property of sub-addittivity and to be a coherent risk measure, see (Acerbi, Tasche, 2002). The ES measures the average of the worst α losses, where α is the percentile of the returns distribution. Unfortunately, Gneiting (2011) showed that the ES does not satisfy a mathematical property called elicitability (while VaR does have it), and it cannot be backtested. However, Emmer et al. (2015) showed that the ES is elicitable conditionally on the VaR, and that it can be backtested through the approximation of several VaR levels: this idea was further developed by Kratz et al. (2018) who proposed a multinomial test of VaR violations at multiple levels as an intuitive idea of backtesting the ES.

The goal of this paper is to examine whether augmenting a large class of volatility models with implied volatility from option prices and Google Trends data improves the quality of the estimated VaRs at multiple confidence levels for the Russian RTS index future. The RTS index is based on the 50 most liquid stocks of the Russian market, and it is an important indicator for the whole Russian market. Since the RTS index is not tradable, we considered the RTS index future for the purpose of our analysis. Google search data is an indicator of the people behavior and their attention, and it can be a driver of future volatility, see for example (Campos et al., 2017) and references therein for a large discussion. Implied volatility from option prices is a standard way to obtain a forward-looking estimate of the volatility which considers investors' beliefs, see the survey by Mayhew (1995) and references therein for more details. Internet searches come mostly from the general public and small investors (Da et al., 2011; Goddard et al. 2015, Vlastakis, Markellos, 2012 and Vozlyublennaia, 2014), whereas implied volatility captures the expectations of institutional investors and market makers who have access to premium and insider information (Martens, Zein, 2004; Busch et al., 2011; Bazhenov, Fantazzini, 2019). These two measures of investors' attention and expectations are used to augment a large class of volatility models to forecast the VaR at multiple levels for the Russian RTS index future, using daily data from 2006 till 2019. We considered four class of models: the Threshold-GARCH model by Glosten et al. (1993) and Zakoian (1994), the Heterogeneous Auto-Regressive (HAR) model by Corsi (2009), the AutoRegressive Fractional (ARFIMA) model by Andersen et al. (2003), and the Realized-GARCH model by Hansen et al. (2012). The forecasting performances of these models are compared using the forecasting diagnostics for market risk measurement, such as the tests by Kupiec (1995) and Christoffersen (1998), the multinomial test of VaR violations by Kratz et al. (2018), the asymmetric quantile loss (QL) function proposed by González-Rivera et al. (2004), and the Model Confidence Set by Hansen et al. (2011).

The first contribution of this paper is an evaluation of the contribution of both online search queries and options-based implied volatility to the modelling of the volatility of the Russian RTS index future, and how this dependence has changed over almost two decades (from 2006 till 2019). To our knowledge, this analysis has not been done elsewhere. The second contribution is an out-of-sample forecasting exercise of the Value-at-Risk for the RTS index future at multiple confidence levels using several alternative models' specifications, with and without Google data and implied volatility. The third contribution of the paper is a robustness check to measure the accuracy of Value-at-Risk forecasts obtained with a multivariate model.

The rest of this paper is organized as follows. Section 2 briefly reviews the literature devoted to Google Trends and implied volatility, while the methods proposed for forecasting the Value-at-Risk are discussed in Section 3. The empirical results are reported in Section 4, while a robustness check is discussed in Section 5. Section 6 briefly concludes.

2. Literature review

There is an increasing body of the financial literature which examines how implied volatility (IV) from option prices and Google Trends data influence volatility modelling and risk measures.

In the case of implied volatility, past research showed that it provides better forecasts for volatility than traditional GARCH models, see (Christensen, Prabhala, 1998; Corredor, Santamaría, 2004; Martens, Zein, 2004; Busch et al., 2011 and Haugom et al., 2014), just to name a few. However, there are some (few) cases where this is not true: for example, Agnolucci (2009) found that a Component-GARCH model performs better than IV in forecasting the volatility of crude oil futures, while Birkelund et al. (2015) examined the Nordic power forward market and found that the IV is a biased predictor of the realized volatility. In general, the financial literature usually shows that the best results are obtained when both the IV and other market variables are included in the forecasting model, with intraday volatility information being the only variable able to successfully complement IV, see (Taylor, Xu, 1997; Pong et al., 2004; Jeon, Taylor, 2013).

While implied volatility forecasts the future volatility well, the results are more mixed when forecasting the future quantiles of the returns' distribution is of concern. Giot (2005) showed that the VaR forecasts based on lagged implied volatility performed similarly to VaR estimates based on GARCH models, while Jeon and Taylor (2013) reported that the implied volatility has explanatory power for the left tail of the conditional distribution of SP500 daily returns. Instead, Chong (2004) found that time series models performed better than the model based on implied volatility when estimating the VaR for exchange rates. Similarly, Christoffersen and Mazzotta (2005) found that, while implied volatility provided the most accurate volatility forecasts for all currency exchange rates and forecast horizons considered, unfortunately, it did not capture the tail behavior of the returns distribution. Barone-Adesi et al. (2019) estimated and backtested the 1, 2.5 and 5% WTI crude oil futures value at risk and conditional value at risk for the years 2011–2016 and for both tails of the distribution, and they showed that the option-implied risk metrics are valid alternatives to filtered-historical simulation models.

The work by Bams et al. (2017) is the largest backtesting exercise to date, comparing the Value-at-Risk forecasts from implied volatility and historical volatility models, using more than 20 years of daily data from American markets. Their large-scale backtesting analysis shows that implied volatility based Value-at-Risk tends to be outperformed by simple GARCH based Value-at-Risk, and they explain the poor performance of the former due to the volatility risk premium embedded in implied volatilities. However, even when correcting for the variance risk premium, the VaR forecasts based on the IV cannot outperform the historical volatility models. Bams et al. (2017) explain this result by showing that, while the IV is useful for forecasting the future volatility, it is not useful for forecasting a quantile of the return distribution, due to the complex dependence structure between the volatility risk premium and the extreme returns which influence the quantile forecasting power of the implied volatility.

In the case of Google search data, a vast literature discussed how the Internet search activity captures investor attention and information demand, see (Ginsberg et al., 2009; Choi, Varian, 2012; Da et al., 2011; Vlastakis, Markellos, 2012; Vozlyublennaia, 2014; Goddard et al., 2015, Fantazzini, Toktamysova, 2015), just to name a few.

Vozlyublennaia (2014) showed that Google data does have short term and — in some cases — even long term effects on returns, whereas the effects on volatility are less pronounced. However, she does not investigate the predictability of volatility by using Google data. Dimpfl and Jank (2016)

were among the first to use daily Google data to forecast daily and weekly realized variances by including it as an additive component in Autoregressive (AR) and Heterogeneous Autoregressive (HAR) models, and they showed that Google searches improved both in- and out-of-sample performances. Campos et al. (2017) evaluated the marginal contribution of Google trends to forecast the Crude Oil Volatility index by using HAR models and several macro-finance variables and showed that Google data has a positive relationship with the oil volatility index, even accounting for macroeconomic variables. They also found that internet search volumes add valuable information to their forecasting models and their predictions generate more economic value than models without them. Xu et al. (2019) showed that Google Trends is a significant source of volatility besides macroeconomic fundamentals and its contribution to forecasting volatility can be enhanced by combining with other macroeconomic variables. Interestingly, they found that Google data with a higher observed frequency tends to be more useful in volatility forecasting than with a lower frequency. Moreover, the higher the stock market volatility is, the more important Google data are for volatility forecasting. Seo et al. (2019) proposed a set of hybrid models based on artificial neural networks with multi-hidden layers, combined with GARCH family models and Google Trends. They reported that their hybrid models with Google data statistically outperform the GARCH models and the hybrid models without Google data when forecasting the volatility of American markets.

The studies dealing with market risk measurement and Google Trends are much more limited, instead: Hamid and Heiden (2015) proposed an empirical similarity approach to forecast the weekly volatility and the VaR for the Dow Jones index by using Google data. They performed an out-of-sample exercise with 260 weekly data and found that their model delivered significantly more accurate forecasts than competing models (HAR and ARFIMA models with and without Google data), while requiring less capital due to fewer over-predictions. Basistha et al. (2019) was the first work that evaluated both the role of the Google data and implied volatility for forecasting the realized volatility of American financial markets, exchanges rates and commodity markets. They found that Google search data play a minor role in predicting the realized volatility once implied volatility is included in the set of regressors. They also performed a small risk management exercise involving the one-week-ahead 1% Value-at-Risk for the SP500 and the DJIA indices and found that adding the implied volatility produces an economically meaningful improvement in market risk measurement, while this is not the case when adding the Google search volume to the forecasting models. Bazhenov and Fantazzini (2019) performed a similar analysis with four Russian stocks for the 2016–2019 time sample, and they found that models including the implied volatility improved their forecasting performances, whereas models including Google Trends data worsened their performances. Interestingly, simple HAR and ARFIMA models without additional regressors often reported the best forecasts for the daily realized volatility and the daily Value-at-Risk at the 1% probability level, thus showing that efficiency gains more than compensate any possible model misspecifications.

3. Methodology

The goal of this paper is to verify whether adding the implied volatility from option prices and Google data to a large set of volatility models improves the quality of the estimated VaRs at multiple confidence levels for the Russian RTS index future. This procedure also has the benefit to indirectly test the quality of the models' Expected Shortfall, following the aforementioned ap-

proach by Kratz et al. (2018). Before presenting these market risk measures and the associated backtesting procedures, we briefly review the measures of volatility and the forecasting models that we will use to compute these market risk measures.

3.1. Measures of volatility

3.1.1. Realized variance

Suppose we have a stochastic process with the following form:

$$dp(t) = \mu(t)dt + \sigma(t)dW(t),$$

where $p(t)$ is the logarithm of instantaneous price, $\mu(t)$ is of finite variation, $\sigma(t)$ is strictly positive and square integrable and $W(t)$ is a Brownian motion. This is a continuous-time diffusive setting which rules out price jumps and assumes a frictionless market. For this diffusion process, the Integrated Variance associated with day t is defined as the integral of the instantaneous variance over the following one day integral:

$$IV_{t+1} = \int_t^{t+1} \sigma^2(s)ds.$$

It is possible to show that the previous integrated variance can be approximated to an arbitrary precision using the sum of intraday squared returns. This nonparametric estimator is called *Realized Variance* and it is a consistent estimator of the integrated variance as the sampling frequency increases, see (Meddahi, 2002) and (Andersen et al., 2001):

$$RV_{t+1} = \sum_{j=1}^{M} r_{t+j\Delta}^2,$$

where $\Delta = 1/M$ is the time interval of the intraday prices, M is the number of intraday returns, while $r_{t+j\Delta}$ is the intraday return. The previous estimator considers the daily realized variance, but different time horizons longer than a single day d can be computed. For example, the weekly realized variance w at time t is given by

$$RV_t^{(w)} = \frac{1}{5}\left(RV_t^{(d)} + RV_{t-1d}^{(d)} + \ldots + RV_{t-4d}^{(d)}\right),$$

where we considered a weekly time interval of five working days.

If we allow for the presence of jumps and consider the following continuous-time jump-diffusion process, $dp(t) = \mu(t)dt + \sigma(t)dW(t) + k(t)dq(t)$, where $q(t)$ is a counting process with $dq(t) = 1$ corresponding to a jump at time t and $dq(t) = 0$ otherwise and $k(t)$ refers to the size of the corresponding jumps, then it is possible to show that the realized volatility converges to the sum of the integrated variance and the cumulative squared jumps, see (Barndorff-Nielsen, Shephard, 2004a, 2006; Andersen et al., 2007):

$$\operatorname*{plim}_{\Delta \to 0} RV_{t+1}(\Delta) = \int_t^{t+1} \sigma^2(s)ds + \sum_{t<s\le t+1} k^2(s).$$

The continuous sample path variation measured by the integrated variance can be estimated non-parametrically using the standardized Realized Bipower Variation measure,

$$BV_{t+1}(\Delta) = \mu_1^{-2} \sum_{j=2}^{1/\Delta} |r_{t+j\Delta}| \cdot |r_{t+(j-1)\Delta}| = \mu_1^{-2} \sum_{i=2}^{M} |r_{t,i}| \cdot |r_{t,i-1}| = C_{t+1}(\Delta),$$

while the jump component can be consistently estimated by

$$J_{t+1}(\Delta) = \max\left[RV_{t+1}(\Delta) - BV_{t+1}(\Delta), 0\right],$$

where the non-negativity truncation on the actual empirical jump measurements was suggested by Barndorff-Nielsen and Shephard (2004b) because the difference between RV and BV can become negative in a given sample. However, Huang, Tauchen (2005) and Andersen et al. (2007) suggested to treat the small jumps as measurement errors and consider them as part of the continuous sample path variation process, whereas only abnormally large values of $RV_{t+1}(\Delta) - BV_{t+1}(\Delta)$ should be associated with the jump component. To achieve this goal, they proposed a test to identify the significant jumps and automatically guarantee that both J_{t+1} and C_{t+1} are positive. We employed this approach in our empirical analysis and we refer to Huang, Tauchen (2005) and Andersen et al. (2007) for the full description of this testing procedure.

Given that we also considered GARCH-type models with daily returns, we had to adjust the previous realized variance for the return in the overnight gap from the market close on day t to the market open on day $t+1$. We scaled up the market-open RV using the unconditional variance estimated with the daily squared returns:

$$RV_{t+1}^{24H} = \left(\frac{\sum_{t=1}^{T} r_t^2}{\sum_{t=1}^{T} RV_t^{OPEN}} \right) RV_{t+1}^{OPEN},$$

where r_t^2 are the daily squared returns computed using the close-to-close daily prices, while RV_{t+1}^{OPEN} is the realized variance computed with intraday data when the RTS future market is open. We chose this approach due to its better results in the financial empirical literature, see (Hansen, Lunde, 2005a; Christoffersen, 2012; Ahoniemi, Lanne, 2013).

3.1.2. Implied volatility

An implied volatility index estimates the market expectations for the future volatility implied by the stock index option prices. The first Russian volatility index named RTSVX (Russian Trading System Volatility Index) was introduced on 7 December 2010 and was discontinued on 12 December 2016. It was based on the volatility of the nearby and next option series for the RTS (Russian Trading System) Index futures, see the Moscow Exchange website for more details. The Moscow Exchange makes available a (partially reconstructed) time series of daily closed prices for the RTSVX index starting from January 2006. A new Russian Volatility Index (RVI) was introduced on 16 April 2014: this index measures the market expectations for volatility over a 30 day period, and it is computed using prices of nearby and next RTS Index option

series. The RVI is calculated in real time during both day and evening sessions (first values 19:00–23:50 Moscow time and then 10:00–18:45 Moscow time). There are three aspects where the RVI differs from the RTSVX: it is discrete, it uses actual option prices over 15 strikes, and it calculates a 30-day volatility. The RVI formula is reported below:

$$IV = 100\sqrt{\frac{T_{365}}{T_{30}} \cdot \left| T_1\sigma_1^2 \frac{T_2 - T_{30}}{T_2 - T_1} + T_2\sigma_2^2 \frac{T_{30} - T_1}{T_2 - T_1} \right|},$$

where T_{30} stands for 30 days expressed as a fraction of a calendar year, T_{365} for 365 days expressed as a fraction of a calendar year, T_1 is the time to expiration of the near-series options expressed as a fraction of a calendar year, T_2 is the time to expiration of the far-series options expressed as a fraction of a calendar year, σ_1^2 is the variance of the near-series options and σ_2^2 is the variance of the next-series of options, see http://fs.moex.com/files/6757 for the full description of the RVI methodology. A detailed comparison of the old and new Russian volatility indexes was performed by Caporale et al. (2019) and they found that the differences were minor. For this reason, we built a composite volatility index ranging from January 2006 till April 2019, which allowed us to cover almost two decades[2].

3.2. Volatility models

3.2.1. TGARCH model

The Generalized Auto-Regressive Conditional Heteroscedasticity (GARCH) models represent an important benchmark in empirical finance, see (Hansen, Lunde, 2005b) for a large-scale backtesting comparison involving more than 330 volatility models. A GARCH(p,q) model for the conditional variance σ_t^2 at time t can be defined as follows:

$$\sigma_{t+1}^2 = \alpha_0 + \sum_{i=0}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=0}^{q} \beta_j \sigma_{t-j}^2 .$$

The Threshold-GARCH (TGARCH) model proposed by Glosten et al. (1993) and Zakoian (1994) to take the leverage effect into account can be specified as follows:

$$\sigma_{t+1}^2 = \alpha_0 + \sum_{i=0}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=0}^{q} \beta_j \sigma_{t-j}^2 + \sum_{k=0}^{r} \gamma_k \varepsilon_{t-k}^2 I(\varepsilon_{t-k} < 0),$$

where $I(\cdot) = 1$ if $\varepsilon_{t-k} < 0$ and zero otherwise. A simple TGARCH(1,1) model with standardized errors following a Student's t-distribution was employed in this work. Moreover, we also considered a TGARCH(1,1) specification including the (implied) volatility index and Google Trends as additional regressors,

[2] We also tried to compute market risk measures using only one of the two volatility indexes, but the results did not change qualitatively, so that we stuck to our composite index. Note that both indexes are expressed in percentage form.

$$\sigma_{t+1}^2 = \alpha_0 + \alpha_1 \varepsilon_t^2 + \beta_1 \sigma_t^2 + \gamma_1 \varepsilon_t^2 I\left(\varepsilon_t < 0\right) + \delta IV_t + \psi GT_t$$

similarly to the specification proposed for Russian stocks by Bazhenov, Fantazzini (2019).

3.2.2. HAR model

Corsi (2009) proposed a model with a hierarchical process, where the future volatility depends on the past volatility over different lengths of periods. The HAR model has the following form:

$$RV_{t+1} = \beta_0 + \beta_D RV_t + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + \epsilon_{t+1},$$

where D, W and M stand for daily, weekly and monthly values of the realized volatility, respectively. Since we want to verify whether adding the implied volatility and Google data improve the estimation of the market risk measures for the RTS index future, we also included these additional regressors using the HAR specification by Bazhenov and Fantazzini (2019):

$$RV_{t+1} = \beta_0 + \beta_D RV_t + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + \delta IV_t + \psi GT_t + \epsilon_{t+1}. \quad (1)$$

Andersen et al. (2007) extended the HAR-RV model by decomposing the realized variances into the continuous sample path variability and the jump variation, and they proposed the so-called HAR-CJ model which is defined as follows:

$$RV_{t+1} = \beta_0 + \beta_{CD} C_t + \beta_{CW} C_{t-5,t} + \beta_{CM} C_{t-22,t} + \beta_{JD} J_t + \beta_{JW} J_{t-5,t} + \beta_{JM} J_{t-22,t} + \epsilon_{t+1}.$$

If we add the IV and Google data, the previous equation will transform into:

$$RV_{t+1} = \beta_0 + \beta_{CD} C_t + \beta_{CW} C_{t-5,t} + \beta_{CM} C_{t-22,t} + \beta_{JD} J_t + \beta_{JW} J_{t-5,t} + \beta_{JM} J_{t-22,t} + \delta IV_t + \psi GT_t + \epsilon_{t+1}.$$

Note that a large-scale forecast comparison of volatility models for the Russian stock market was performed by Aganin (2017), and he found that the HAR model showed a statistically significant superior performance compared to all other competing models.

3.2.3. ARFIMA model

The Auto-Regressive Fractional Integrated Moving Average, ARFIMA(p,d,q), model to forecast the realized volatility was proposed by Andersen et al. (2003), and it can be expressed as follows:

$$\Phi(L)(1-L)^d (RV_{t+1} - \mu) = \Theta(L)\varepsilon_{t+1},$$

where L is the lag operator, $\Phi(L) = 1 - \varphi_1 L - \ldots - \varphi_p L^p$, $\Theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q$ and $(1-L)^d$ is the fractional differencing operator defined by:

$$(1-L)^d = \sum_{k=0}^{\infty} \frac{\Gamma(k-d) L^k}{\Gamma(-d)\Gamma(k+1)},$$

where $\Gamma(\cdot)$ is the gamma function. Similarly to the HAR and GARCH models, we also added the implied volatility and Google Trends as external regressors:

$$\Phi(L)(1-L)^d(RV_{t+1}-\mu)=\delta IV_t+\psi GT_t+\Theta(L)\varepsilon_{t+1}.$$

Hyndman and Khandakar (2008) proposed an algorithm for the automatic selection of the optimal ARFIMA model, which is implemented in the *R* packages *forecast* and *rugarch* and which was used in this work.

3.2.4. Realized-GARCH model

The Realized-GARCH by Hansen et al. (2012) jointly models the returns and the realized measures of volatility. More specifically, this model connects the realized volatility to the latent volatility via a measurement equation, which also accommodates an asymmetric reaction to shocks. The Realized-GARCH model with a log-linear specification can be written as follows:

$$r_t=\mu+\sqrt{\sigma_t^2}\cdot z_t,\quad z_t\sim i.i.d.(0,1),$$

$$\log\sigma_t^2=\omega+\sum_{i=1}^{q}\gamma_i\log RV_{t-i}+\sum_{i=1}^{p}\beta_i\log\sigma_{t-i}^2,$$

$$\log RV_t=\xi+\psi\log\sigma_t^2+\tau_1 z_t+\tau_2\left(z_t^2-1\right)+u_t,\quad u_t\sim i.i.d.\left(0,\sigma_u^2\right),$$

where the three equations represent the return equation, the volatility equation, and the measurement equation, respectively. The latter equation models the contemporaneous dependence between the latent volatility and the realized measure, while the terms $\tau_1 z_t+\tau_2\left(z_t^2-1\right)$ accommodate potential leverage-type effects. Similarly to previous models, we also added the implied volatility and Google Trends as external regressors. The log-linear specification was used in this work due to its better numerical and statistical properties compared to the linear specification, see Hansen et al. (2012).

3.3. Market risk measures

The Value-at-Risk is the maximum market loss of a financial position over a time horizon *h* with at a pre-defined confidence level $(1-\alpha)$, or alternatively, the minimum loss of the α worst losses over the time horizon *h*. The VaR is a widely used measure of market risk in the financial sector, and we refer to (McNeil et al., 2015; Fantazzini, 2019) for a large discussion at the textbook level. In this work, we considered $h=1$.

In the case of GARCH and Realized-GARCH models with Student's *t* errors, the 1-day ahead VaR can be computed as follows:

$$VaR_{t+1,\alpha}=\hat{\mu}_{t+1}+t_{\alpha,\upsilon}^{-1}\cdot\sqrt{(\upsilon-2)/\upsilon}\cdot\sqrt{\hat{\sigma}_{t+1}^2},$$

where $\hat{\mu}_{t+1}$ is the 1-day-ahead forecast of the conditional mean, $\hat{\sigma}_{t+1}^2$ is the 1-day-ahead forecast of the conditional variance, while $t_{\alpha,\upsilon}^{-1}$ is the inverse function of the Student's *t* distribution with υ

degrees of freedom at the probability level α. The term $\sqrt{(\upsilon-2)/\upsilon}\cdot\hat{\sigma}_{t+1}$ is also known as the scale parameter of the Student's t distribution.

In the case of HAR and ARFIMA models, the 1-day ahead VaR can be computed as follows:

$$VaR_{t+1,\alpha}=\Phi_{\alpha}^{-1}\sqrt{\widehat{RV}_{t+1}},$$

where Φ_{α}^{-1} is the inverse function of a standard normal distribution function at the probability level α, while $\widehat{RV}_{t+1}$ is the 1-day-ahead forecast for the realized volatility.

The Expected Shortfall (ES) measures the average of the worst α losses, where α is a percentile of the returns' distribution, and it is computed as follows:

$$ES_{\alpha}=\frac{1}{\alpha}\int_{0}^{\alpha}F_{z}^{-1}(X)dz=\frac{1}{\alpha}\int_{0}^{\alpha}VaR_{z}(X)dz,$$

where F^{-1} is the inverse function of the returns' distribution, that is the Value-at-Risk.

The ES attracted a lot of attention when in October 2013 the Basel Committee on Banking Supervision issued the revision of the market risk framework, where the 99% VaR (that is, the α = 1% probability level VaR) was substituted with the 97.5% ES (Basel Committee…, 2013, p. 18). The main drawback of the ES is that it lacks a mathematical property called elicitability, while VaR does have it (see (Gneiting, 2011)): if a risk measure is elicitable, then it can be used within a scoring function to be minimized for comparative tests on models, thus allowing for the ranking of the risk models' performance. However, Emmer et al. (2015) and Kratz et al. (2018) showed that the ES is elicitable conditionally on the VaR, and that it can be back-tested through the approximation of several VaR levels. More specifically, Emmer et al. (2015) showed that,

$$ES_{\alpha}\approx\frac{1}{4}\left[q(\alpha)+q(0.75\alpha+0.25)+q(0.5\alpha+0.5)+q(0.25\alpha+0.75)\right],$$

where $q(\gamma)=VaR_{\gamma}$. For example, if $\alpha=2.5\%$ as proposed by the Basel Committee, then

$$ES_{2.5\%}\approx\frac{1}{4}\left[VaR_{2.5\%}+VaR_{2.125\%}+VaR_{1.75\%}+VaR_{1.375\%}\right].$$

In this case, a more convenient approximation was proposed by Wimmerstedt (2015):

$$ES_{2.5\%}\approx\frac{1}{5}\left[VaR_{2.5\%}+VaR_{2.0\%}+VaR_{1.5\%}+VaR_{1.0\%}+VaR_{0.5\%}\right].$$

Given these recent results, we decided to take a neutral stance towards the current (hot) debate between VaR and ES proponents: we computed the VaR_{α} at five probability levels (2.5%, 2%, 1.5%, 1%, 0.5%), so that we could also provide an approximate backtesting of the $ES_{2.5\%}$ which will be included in the future Basel 3 agreement (scheduled to be introduced on 1 January 2022).

3.4. Backtesting methods

The forecasting performance of different VaR models can be checked by comparing the forecasted values of the VaR with the actual returns for each day. The first step is to count the number of violations T_1 when the ex-ante forecasted VaR are smaller than the actual losses, with $T = T_1 + T_0$, while T_0 is the the number of no VaR violations. A «perfect VaR model» would show the fraction of actual violations $\hat{\pi} = T_1 / T$ equal to α%. The null hypothesis $\mathbf{H_0} : \hat{\pi} = \alpha$ can be tested using the *unconditional coverage test* by Kupiec (1995). He showed that the test statistic for this null hypothesis is given by:

$$LR_{uc} = -2\log\left[(1-\alpha)^{T_0}\alpha^{T_1} \Big/ \left\{(1-T_1/T)^{T_0}(T_1/T)^{T_1}\right\}\right] \overset{\mathbf{H_0}}{\sim} \chi_1^2 .$$

If we want to test the joint null hypothesis that the average number of VaR violations is correct and the violations are independent, then we can resort to the *conditional coverage test* by Christoffersen (1998). The main advantage of this test is that it can reject a model that forecasts either too many or too few clustered violations, while its main disadvantage is the need of at least several hundred observations to be accurate. The test statistic is reported below:

$$LR_{cc} = -2\log\left[(1-\alpha)^{T_0}\alpha^{T_1}\right] + 2\log\left[(1-\pi_{01})^{T_{00}}\pi_{01}^{T_{01}}(1-\pi_{11})^{T_{10}}\pi_{11}^{T_{11}}\right] \overset{\mathbf{H_0}}{\sim} \chi_2^2 ,$$

where T_{ij} is the number of observations with value i followed by j for $i, j = 0,1$ and $\pi_{ij} = T_{ij} \Big/ \sum_j T_{ij}$ are the corresponding probabilities.

Financial regulators are concerned not only with the number of VaR violations, but also with their magnitude. For this reason, we also computed the *asymmetric quantile loss* (*QL*) function proposed by González-Rivera et al. (2004):

$$QL_{t+1,\alpha} = \left(\alpha - I_{t+1}(\alpha)\right)\left(r_{t+1} - VaR_{t+1,\alpha}\right),$$

where $I_{t+1}(\alpha) = 1$ if $r_{t+1} < VaR_{t+1,\alpha}$ and zero otherwise. This loss function penalizes more heavily the realized losses below the α-th quantile level, so that it can be useful to compare the costs of different admissible choices.

The quantile loss function was subsequently used together with the *Model Confidence Set* (MCS) by Hansen et al. (2011) to select the best VaR forecasting models at a specified confidence level. Given the difference between the *QL*s of models i and j at time t (that is $d_{i,j,t} = QL_{i,t} - QL_{j,t}$), the MCS approach is used to test the following hypothesis of equal predictive ability, $\mathbf{H_{0,M}} : \mathrm{E}(d_{i,j,t}) = 0$, for all $i,j \in M$, where M is the set of forecasting models. The first step is to compute the following t-statistics:

$$t_{ij} = \frac{\bar{d}_{ij}}{\widehat{\mathrm{var}}\left(\bar{d}_{ij}\right)} \quad \text{for } i, j \in M,$$

where $\bar{d}_{ij} = T^{-1}\sum_{t=1}^{T} d_{ij,t}$ and $\widehat{\mathrm{var}}\left(\bar{d}_{ij}\right)$ is an estimate of $\mathrm{var}\left(\bar{d}_{ij}\right)$. Then, the following test statistic is computed: $T_{R,M} = \max_{i \in M}\left|t_{ij}\right|$. This statistic has a non-standard distribution, so the distribution under the null hypothesis is computed using bootstrap methods with 5000 replications and

a minimum block length equals to 5. If the null hypothesis is rejected, one model is eliminated from the analysis and the testing procedure starts from the beginning.

The *multinomial VaR test* by Kratz et al. (2018) implicitly backtests the Expected Shortfall using the previous idea by Emmer et al. (2015) to approximate the ES with several VaR levels. More specifically, Kratz et al. (2018) considered several VaR probability levels $\alpha_1,\ldots,\alpha_N$ defined by $\alpha_j = \alpha + \left[(j-1)/N\right](1-\alpha)$, $j=1,\ldots,N$, for some starting level α. If $I_{t,j} = 1_{(L_t > VaR_{t,\alpha_j})}$ is the usual indicator function for a VaR violation at the level α_j and $X_t = \sum_{j=1}^{N} I_{t,j}$, then the sequence $(X_t)_{t=1,\ldots,T}$ counts the number of VaR violations at the level α_j. If we denote with $MN\left(T,\left(p_0,\ldots,p_N\right)\right)$ the multinomial distribution with T trials, each of which may result in one of $N+1$ outcomes $\{0,1,\ldots,N\}$ according to probabilities $p_0,\ldots,p_N$ that sum to one, while the observed cell counts are denoted by $O_j = \sum_{t=1}^{T} I_{(X_t=j)}$, $j=0,1,\ldots,N$, then, under the assumptions of unconditional coverage and independence as in Christoffersen (1998), it is possible to show that the random vector $\left(O_0,\ldots,O_N\right)$ will follow the multinomial distribution $\left(O_0,\ldots,O_N\right) \sim MN\left(T,\left(\alpha_1-\alpha_0,\ldots,\alpha_{N+1}-\alpha_N\right)\right)$. Given an estimated multinomial distribution represented by $MN\left(T,\left(\theta_1-\theta_0,\ldots,\theta_{N+1}-\theta_N\right)\right)$ where θ_j ($j=1,\ldots,N$) are the distribution parameters estimated with the available data sample, Kratz et al. (2018) consider the following null and alternative hypotheses:

$$\mathbf{H_0}: \theta_j = \alpha_j, \quad \text{for } j=1,\ldots,N,$$
$$\mathbf{H_1}: \theta_j \neq \alpha_j, \quad \text{for at least one } j \in \{1,\ldots,N\}.$$

The null hypothesis can be tested with several test statistics, and we refer to (Cai, Krishnamoorthy, 2006) for a large simulations study to verify the exact size and power properties of five possible tests (three of them were later used by Kratz et al. (2018)). We employed the *exact method* in our empirical analysis: this is the fifth test statistic reviewed by Cai and Krishnamoorthy (2006), and it computes the probability of a given outcome under the null hypothesis using the multinomial probability distribution itself:

$$P\left(O_0,O_1,\ldots,O_N\right) = \frac{T!}{O_0!O_1!\ldots O_N!}(\alpha_1-\alpha_0)^{O_0}(\alpha_2-\alpha_1)^{O_1}\ldots(\alpha_{N+1}-\alpha_N)^{O_N}.$$

Cai and Krishnamoorthy (2006) found that the exact method performs very well, but it can be time-consuming if the number of cells N and the sample size T are large. In this latter case, simulation methods need be used[3]. A large discussion at the textbook level of all these backtesting methods and many others can be found in (Fantazzini, 2019, Ch. 11).

[3] Several tests reviewed by Cai and Krishnamoorthy (2006) are implemented in the *R* package *XNomial* available at https://cran.r-project.org/package=XNomial.

4. Empirical analysis

4.1. Data

Intraday data sampled every 5 minutes for the continuous RTS index future were downloaded from the website finam.ru, covering the period from January 2006 till April 2019. We then used the 5-minutes squared log-returns to calculate the daily, weekly and monthly realized variance measures, as previously discussed in Section 3.1.1. We remark that Liu et al. (2015) performed a large-scale forecasting analysis involving more than 400 estimators of realized measures and they found that it is difficult to significantly outperform the 5-minute RV estimator. For this reason, we employed this estimator in this work.

The daily historical data for the implied volatility of the RTS index were downloaded from the Moscow exchange (moex.com): as we discussed in Section 3.1.2, our volatility index is a composite index which consists of the RTSVX index for the period from January 2006 till November 2016, and the RVI index for the period from December 2016 till April 2019. The volatility index was rescaled to make it a daily volatility index, comparable with the other variables used in our empirical work.

Google Trends tracks the number of search queries for a topic or a keyword over a specific period and a specific region, and creates time-series reporting the relative popularity of the searched queries. More specifically, the amount of searches is divided by the total amount of searches for the same period and region, and the resulting time series is divided by its highest value and multiplied by 100. We remark that Google Trends creates a new time series for every period because its algorithm takes the highest value over the chosen period and normalizes all others to this peak point. Even though Google is not the main search engine in Russia (Yandex is, with a market share close to 56% in 2018 — all platforms), its market share is still very significant (over 40%). Moreover, Yandex search data are available only for the last year (in case of monthly data) or for the last 2 years (in case of weekly data), so that a reliable statistical analysis with these data is not possible. We used Google Trends data for the query «RTS index», both in English and in Russian, and we computed the average of these two series. All search volumes were downloaded from the Google Trends website using the *R* package *gtrendsR*. Google trends data are available since 2004, but if a multi-year sample is requested, only monthly data are obtained. To remedy this problem, we downloaded daily data for each month separately, and then we concatenated them in a single series by multiplying the separate daily data with the corresponding monthly data for the whole period.

The daily returns for the RTS index future, the implied volatility index (rescaled to show the daily implied variance), the Google Trends data and the daily realized variance are reported in Fig. 1.

4.2. In-sample analysis

Our time sample covers almost 15 years of daily data which includes several episodes of high volatility in the Russian financial market, like the global financial crisis in 2008–2009 and several rounds of sanctions since 2014: for example, Aganin and Peresetsky (2018) found that sanctions initially increased the volatility of the ruble exchange rate, but their impact has then

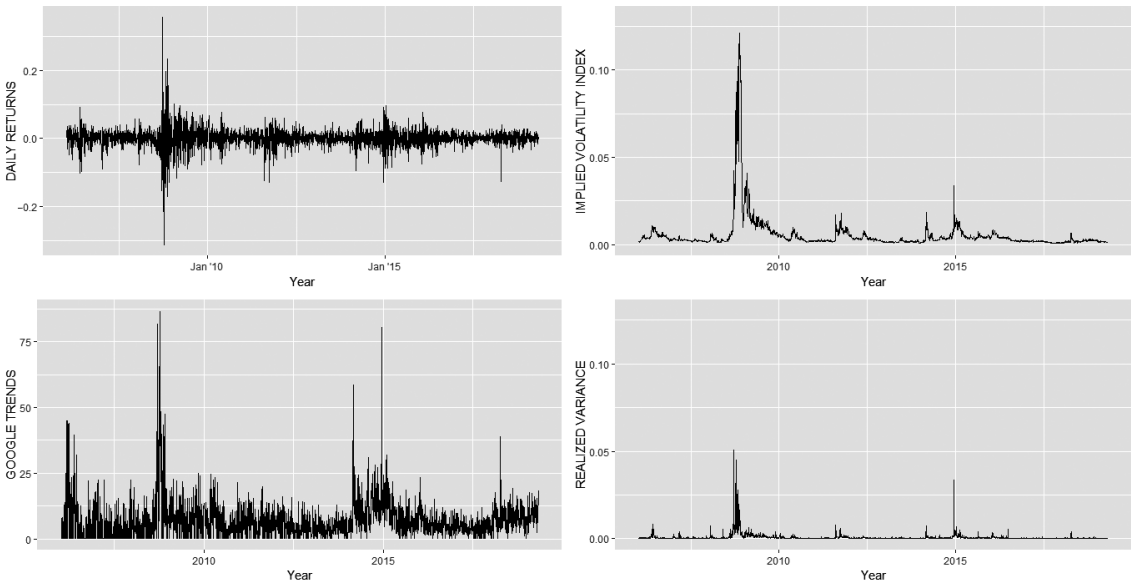


Fig. 1. The daily returns, the implied volatility index, the Google Trends data and the daily realized variance for the RTS index future

decreased with time. For these reasons, we do not report the models' estimates for the full sample because they would be misleading, being strongly impacted by structural breaks of different nature. Instead, we prefer to show the recursive estimates of the coefficients for the implied volatility and Google Trends in the HAR model of equation (1), which better convey the changing impacts of these regressors on the realized volatility of the RTS index future.

Figure 2 clearly shows the strong impact of the global financial crisis in 2008–2009, whereas the impact of sanctions since 2014 appear to be minor. Moreover, the (positive) effect of Google search queries seems to have decreased with time.

There is a vast literature dealing with multiple structural breaks in linear regression models, see (Zeileis et al., 2002; Zeileis, 2005; Perron, 2006) for extensive surveys. Among the several

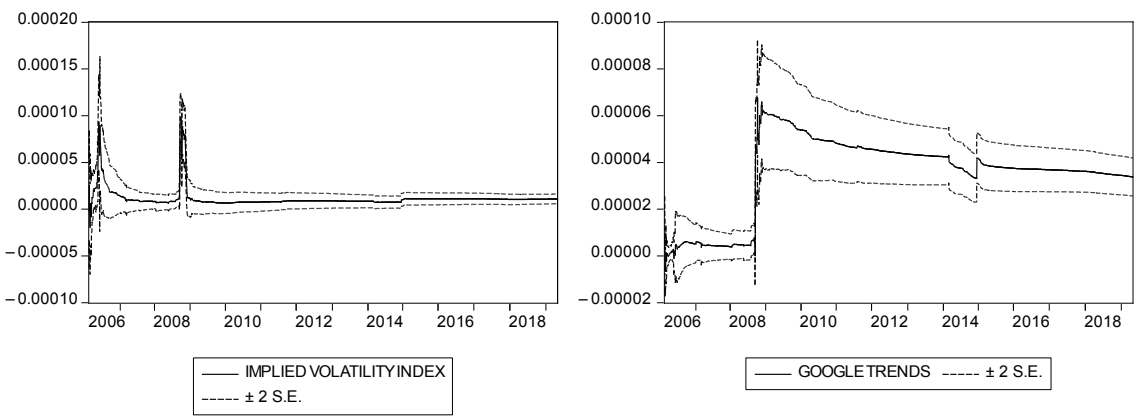


Fig. 2. Recursive estimates of the coefficients for the implied volatility and Google Trends in the HAR model of equation (1)

approaches proposed, we decided to employ the methodology based on information criteria proposed by Yao (1988), Liu et al. (1997), Bai and Perron (2003a) and Zeileis et al. (2010), which finds the optimal number of breakpoints by optimizing the Bayesian Information Criterion (BIC) and the modified BIC by Liu et al. (1997), known as the LWZ criterion. This approach has shown to be robust with different model setups and computationally tractable even with large datasets.

The multiple breakpoint test for the HAR model of equation (1) allowing for a maximum of 5 breaks and a dataset trimming of 15% is reported in Table 1. It employs heteroscedasticity and autocorrelation consistent (HAC) covariances using a Quadratic-Spectral kernel with a Newey–West bandwidth, see (Newey, West, 1994); Zeileis, 2006) and references therein for more details. The estimates of the model coefficients with breaks are reported in Table 2.

Table 1 shows that both the Schwarz and the LWZ information criteria select 1 break, coinciding with the beginning of the global financial crisis. Interestingly, the selected date is just 1 week after the bankruptcy of Lehman Brothers. Table 2 shows that both the sign and the size of the coefficients change significantly between the two time samples, particularly for the volatility components. Instead, the coefficients for the lagged implied volatility and Google Trends remain positive in both samples, but the RVI is statistically significant only in the first sample up to September 2008, whereas Google search queries are statistically significant only in the second sample (which makes sense given that Google was not very used in Russia during the first period). We also computed other tests for detecting breaks, like the sequential tests proposed by Bai (1997) and Bai, Perron (1998), and the global maximizer tests by Bai, Perron (1998, 2003a, b): in these cases, the number of significant breaks was higher — mostly 4 breaks — identified around September 2008, September 2010, December 2014 and December 2016, which we can be loosely interpreted as the beginning and the end of the global financial crisis in Russia (2008–2010) and the beginning and the end of the crisis related to sanctions and the oil price collapse (2014–2016)[4]. Finally, we remark that the variability of the parameters for the other models (GARCH and ARFIMA models) was even higher, which should not be a surprise, given the greater computational complexity of these models. However, we do not report them for the sake of interest and space, and we prefer to focus on VaR forecasting which is the main goal of this work.

4.3. Value-at-Risk forecasts

The previous empirical evidence of 1 or more breaks suggested to us to use a rolling window of 500 observations to estimate the volatility models and to compute the forecasted VaR. This time window should be a good compromise, given the numerical properties of GARCH models discussed by Hwang and Valls Pereira (2006) and Bianchi et al. (2011), together with the simulation evidence reported by Pesaran and Timmermann (2007), who showed that in a regression with multiple breaks the optimal window for estimation includes all of the observations after the last break, plus a limited number of observations before this break. Moreover, we considered the results for the full out-of-sample validation period (2008–2019) and for a rolling out-of-sample of 250 days (as requested by the Basel agreements), to examine how several structural breaks impacted the backtesting procedure. We employed the following models (see Table 3).

[4] These results are not reported for sake of space and are available from the authors upon request.

Table 1. Multiple breakpoint test for the HAR model of equation (1)

Number of breaks	Number of coefficients	Sum of squared residuals	Log-L	Schwarz* criterion	LWZ* criterion
0	6	0.006125	17081.35	–13.18175	–13.15066
1	13	0.005198	17351.93	–13.32864	–13.26128
2	20	0.005136	17371.72	–13.32345	–13.21981
3	27	0.005099	17383.45	–13.31337	–13.17344
4	34	0.005069	17393.26	–13.30213	–13.12591
5	41	0.005060	17396.41	–13.28684	–13.07433

Estimated break dates

1: 9/23/2008
2: 9/23/2008, 9/17/2010
3: 9/23/2008, 7/18/2014, 7/19/2016
4: 9/23/2008, 9/20/2010, 7/18/2014, 7/19/2016
5: 9/23/2008, 9/17/2010, 9/21/2012, 11/25/2014, 11/21/2016

* The minimum information criterion values are displayed with shading.

Table 2. Model estimates for the HAR model of equation (1) with breaks and HAC standard errors

	2/09/2006–9/22/2008	9/23/2008–4/29/2019
Constant	–0.00154***	–0.00016
	(0.00042)	(0.00014)
Realized Volatility (daily)	–0.77159**	0.49714***
	(0.27715)	(0.01829)
Realized Volatility (Weekly)	2.62269***	–0.06182
	(0.68857)	(0.04861)
Realized Volatility (Monthly)	–1.15187*	0.38608***
	(0.52013)	(0.07994)
RVI index	0.00005***	0.00000
	(0.00002)	(0.00000)
Google Trends	0.00004	0.00002**
	(0.00002)	(0.00001)

Note. * — $p < 0.05$, ** — $p < 0.01$, *** — $p < 0.001$

Table 3. Model specifications used in the backtesting analysis

Model	External regressors				Total	
	No external regressors	IV	GT	IV+GT		
TGARCH	✓	✓	✓	✓	4	
HAR	✓	✓	✓	✓	4	
HARCJ	✓	✓	✓	✓	4	
ARFIMA	✓	✓	✓	✓	4	
RG	✓	✓	✓	✓	4	
HAR LOG	✓	✓	✓	✓	4	
HARCJ LOG	✓	✓	✓	✓	4	TOTAL
ARFIMA LOG	✓	✓	✓	✓	4	32

4.3.1. Full out-of-sample validation

The *p*-values of the Kupiec and Christoffersen's tests values and the number of violations (in %) are reported in Table 4, while the models included in the Model Confidence Set (MCS) at the 10% confidence level and their associated asymmetric quantile loss are reported in Table 5. The *p*-values of the Multinomial VaR test by Kratz et al. (2018) with probability levels $\alpha_1 = 0.5\%$, $\alpha_2 = 1\%$, $\alpha_3 = 1.5\%$, $\alpha_4 = 2\%$, $\alpha_5 = 2.5\%$ are reported in Table 6. Only models who reached numerical convergence over the full out-of-sample are reported.

These tables show that only TGARCH models were able to pass the Kupiec and Christoffersen's tests for most quantiles, and to provide VaR violations (in %) close to the theoretical probability levels. TGARCH models also reported the lowest asymmetric losses for all quantiles up to the 2% probability level, and they provided the most precise VaR forecasts for the most extreme quantiles (0.5% and 1%), which are the most important quantiles for regulatory purposes. However, only the TGARCH model with implied volatility managed to pass most of the VaR tests, including the multinomial back-test with five quantiles, whereas TGARCH models with Google Trends or without any external regressors performed worse.

In general, we found that when both the implied volatility and Google data are added jointly, the parameters estimates of several models became very unstable (see the next Section 4.3.2 for more details), while six models out of 32 simply did not reach numerical convergence (four ARFIMA models and two realized-GARCH models). With the exception of TGARCH models, these results highlight that simpler models with no external regressors are a better choice when out-of-sample forecasting is the main concern, thanks to more efficient estimates. Our empirical evidence complements the results provided by Bams et al. (2017), who showed that implied volatility based Value-at-Risk could not outperform simple GARCH based Value-at-Risk due to the complex dependence structure between implied volatility, realized volatility and extreme returns. This is particularly true for Russian financial markets, where extreme returns take place more often than in American markets and they are caused by different type of shocks, from energy economics to geopolitics, see e.g. (Malakhovskaya, Minabutdinov, 2014; Aganin, Peresetsky, 2018). However, in the case of Russian markets, GARCH models augmented with IV do provide more precise VaR forecasts than simple GARCH models. Moreover, our results reveals another important factor that a model need to possess for successful VaR forecasting: computational robustness in case of frequent and extreme market returns.

4.3.2. Rolling out-of-sample of 250 days

After the previous results, we wanted to verify how the backtesting performance of the competing models changed over time. To achieve this goal, we computed the VaR violations in % for all competing models using a rolling out-of-sample of 250 days (as requested by the Basel agreements). The full color figure reporting the violations (in %) for the forecasted VaR at the 1% probability level can be found in the supplementary materials posted on the corresponding author's website[5].

First, the performance of TGARCH models remained remarkably stable over the full period, ranging between 1% and 2%, despite the several episodes of strong volatility in the RTS index

[5] https://sites.google.com/site/deanfantazzini/publications/recursive_VaR.png.

Table 4. Kupiec tests *p*-values, Christoffersen's tests *p*-values and number of violations (in %)

Model	VaR with $\alpha = 0.5\%$			VaR with $\alpha = 1\%$			VaR with $\alpha = 1.5\%$			VaR with $\alpha = 2\%$			VaR with $\alpha = 2.5\%$		
	Kupiec	Christ.	Violations	Kupiec	Christ.	Violations	Kupiec	Christ.	Violations	Kupiec	Christ.	Violations	Kupiec	Christ.	Violations
TGARCH	0.00	0.01	0.96	*0.00*	*0.00*	1.82	*0.00*	*0.00*	2.64	*0.00*	*0.00*	3.15	*0.00*	*0.00*	3.72
TGARCH IV	0.13	0.28	0.71	*0.03*	0.06	1.43	*0.04*	0.11	2.00	*0.00*	*0.02*	2.79	*0.04*	0.10	3.15
TGARCH GT	0.13	0.28	0.71	*0.02*	*0.04*	1.47	*0.01*	*0.04*	2.11	*0.00*	*0.00*	2.93	*0.01*	*0.03*	3.29
TGARCH IVGT	0.13	0.28	0.71	*0.00*	*0.01*	1.57	*0.03*	0.08	2.04	*0.01*	0.05	2.68	*0.04*	0.11	3.15
HAR	*0.00*	*0.00*	1.04	*0.01*	*0.02*	1.54	*0.00*	*0.00*	2.22	*0.01*	*0.02*	2.75	0.06	0.08	3.07
HAR IV	*0.00*	*0.00*	4.75	*0.00*	*0.00*	5.58	*0.00*	*0.00*	6.29	*0.00*	*0.00*	6.79	*0.00*	*0.00*	7.29
HAR GT	*0.00*	*0.00*	2.14	*0.00*	*0.00*	2.75	*0.00*	*0.00*	3.25	*0.00*	*0.00*	3.65	*0.00*	*0.00*	4.25
HAR IVGT	*0.00*	*0.00*	5.97	*0.00*	*0.00*	6.75	*0.00*	*0.00*	7.15	*0.00*	*0.00*	7.68	*0.00*	*0.00*	8.04
HARCJ	*0.00*	*0.00*	1.22	*0.00*	*0.00*	1.64	*0.01*	*0.01*	2.14	*0.04*	*0.04*	2.57	0.12	0.20	2.97
HARCJ IV	*0.00*	*0.00*	4.93	*0.00*	*0.00*	5.58	*0.00*	*0.00*	6.11	*0.00*	*0.00*	6.72	*0.00*	*0.00*	7.33
HARCJ GT	*0.00*	*0.00*	1.97	*0.00*	*0.00*	2.43	*0.00*	*0.00*	3.11	*0.00*	*0.00*	3.47	*0.00*	*0.00*	3.93
HARCJ IVGT	*0.00*	*0.00*	5.58	*0.00*	*0.00*	6.33	*0.00*	*0.00*	6.90	*0.00*	*0.00*	7.40	*0.00*	*0.00*	7.97
ARFIMA	*0.00*	*0.00*	1.14	*0.01*	*0.02*	1.54	*0.01*	*0.01*	2.18	*0.01*	*0.01*	2.68	0.05	0.07	3.11
ARFIMA GT	*0.00*	*0.00*	1.86	*0.00*	*0.00*	2.61	*0.00*	*0.00*	3.25	*0.00*	*0.00*	3.75	*0.00*	*0.00*	4.32
RG	*0.00*	*0.00*	1.11	*0.00*	*0.00*	1.82	*0.00*	*0.00*	2.25	*0.01*	*0.03*	2.72	0.06	0.17	3.07
RG IV	*0.00*	*0.00*	1.18	*0.00*	*0.00*	1.75	*0.00*	*0.00*	2.36	*0.00*	*0.02*	2.79	*0.01*	*0.03*	3.29
RG GT	*0.00*	*0.00*	1.11	*0.00*	*0.00*	1.79	*0.00*	*0.00*	2.25	*0.01*	*0.04*	2.68	0.05	0.14	3.11
RG IVGT	*0.00*	*0.00*	1.18	*0.00*	*0.00*	1.86	*0.00*	*0.00*	2.32	*0.01*	*0.04*	2.72	*0.01*	*0.04*	3.25
HAR LOG	*0.00*	*0.00*	1.82	*0.00*	*0.00*	2.61	*0.00*	*0.00*	3.18	*0.00*	*0.00*	3.47	*0.00*	*0.00*	3.93
HAR IV LOG	*0.00*	*0.00*	1.68	*0.00*	*0.00*	2.43	*0.00*	*0.00*	2.93	*0.00*	*0.00*	3.54	*0.00*	*0.00*	4.07
HAR GT LOG	*0.00*	*0.00*	1.72	*0.00*	*0.00*	2.57	*0.00*	*0.00*	3.15	*0.00*	*0.00*	3.57	*0.00*	*0.00*	4.00
HAR IVGT LOG	*0.00*	*0.00*	1.75	*0.00*	*0.00*	2.43	*0.00*	*0.00*	2.93	*0.00*	*0.00*	3.57	*0.00*	*0.00*	4.07
HARCJ LOG	*0.00*	*0.00*	1.79	*0.00*	*0.00*	2.43	*0.00*	*0.00*	2.79	*0.00*	*0.00*	3.22	*0.00*	*0.00*	3.90
HARCJ IV LOG	*0.00*	*0.00*	1.64	*0.00*	*0.00*	2.43	*0.00*	*0.00*	2.82	*0.00*	*0.00*	3.32	*0.00*	*0.00*	3.82
HARCJ GT LOG	*0.00*	*0.00*	1.79	*0.00*	*0.00*	2.43	*0.00*	*0.00*	2.82	*0.00*	*0.00*	3.22	*0.00*	*0.00*	3.93
HARCJ IVGT LOG	*0.00*	*0.00*	1.57	*0.00*	*0.00*	2.39	*0.00*	*0.00*	2.82	*0.00*	*0.00*	3.36	*0.00*	*0.00*	3.82

Note. *P*-values smaller than 0.05 are in italics font.

Table 5. Models included in the MCS at the 10% confidence level and associated asymmetric quantile loss

VaR with $\alpha = 0.5\%$		VaR with $\alpha = 1\%$		VaR with $\alpha = 1.5\%$		VaR with $\alpha = 2\%$		VaR with $\alpha = 2.5\%$	
Models in MCS	Loss	Models in MCS	Loss	Models in MCS	Loss	Models in MCS	Loss	Models in MCS	Loss
TGARCH IVGT	0.00047	TGARCH IVGT	0.00080	TGARCH IV	0.00110	TGARCH IVGT	0.00138	HARCJ	0.00163
TGARCH IV	0.00049	TGARCH IV	0.00081	TGARCH IVGT	0.00110	TGARCH IV	0.00138	TGARCH IVGT	0.00163
TGARCH	0.00049	TGARCH GT	0.00081	TGARCH GT	0.00111	HARCJ	0.00138	HAR	0.00163
TGARCH GT	0.00049	TGARCH	0.00083	HARCJ	0.00112	TGARCH GT	0.00139	ARFIMA	0.00164
HARCJ	0.00053	HARCJ	0.00085	HAR	0.00113	HAR	0.00139	TGARCH IV	0.00164
HAR	0.00054	HAR	0.00085	TGARCH	0.00114	ARFIMA	0.00139	HARCJ IV LOG	0.00164
HARCJ IV LOG	0.00055	ARFIMA	0.00086	ARFIMA	0.00114	HARCJ IV LOG	0.00140	HARCJ IVGT LOG	0.00164
HARCJ IVGT LOG	0.00055	HARCJ IV LOG	0.00087	HARCJ IV LOG	0.00115	HARCJ IVGT LOG	0.00140	TGARCH GT	0.00164
ARFIMA	0.00056	HARCJ IVGT LOG	0.00087	HARCJ IVGT LOG	0.00115	TGARCH	0.00142	HAR IV LOG	0.00166
HAR IV LOG	0.00057	HAR IV LOG	0.00089	HAR IV LOG	0.00116	HAR IV LOG	0.00142	HARCJ GT LOG	0.00166
HARCJ GT LOG	0.00057	HARCJ GT LOG	0.00089	HARCJ GT LOG	0.00117	HARCJ GT LOG	0.00142	TGARCH	0.00166
HARCJ LOG	0.00057	HAR IVGT LOG	0.00090	HAR IVGT LOG	0.00118	RG IVGT	0.00143	RG IV	0.00167
HAR IVGT LOG	0.00058	HARCJ LOG	0.00090	HARCJ LOG	0.00118	RG IV	0.00143	RG IVGT	0.00167
HAR LOG	0.00060	RG IVGT	0.00090	RG IVGT	0.00118	HAR IVGT LOG	0.00143	HARCJ LOG	0.00167
RG IVGT	0.00060	RG IV	0.00091	RG IV	0.00118	HARCJ LOG	0.00143	HAR IVGT LOG	0.00167
HAR GT LOG	0.00060	HAR LOG	0.00091	RG GT	0.00119	RG GT	0.00144	RG GT	0.00168
RG IV	0.00061	RG GT	0.00092	HAR LOG	0.00119	RG	0.00145	RG	0.00168
HARCJ GT	0.00061	HAR GT LOG	0.00092	RG	0.00120	HAR LOG	0.00145	HAR LOG	0.00168
RG GT	0.00061	RG	0.00093	HAR GT LOG	0.00120	HAR GT LOG	0.00146	HAR GT LOG	0.00169
RG	0.00062	HARCJ GT	0.00093	HARCJ GT	0.00121	ARFIMA GT	0.00147	ARFIMA GT	0.00170
ARFIMA GT	0.00063	ARFIMA GT	0.00094	ARFIMA GT	0.00122	HARCJ GT	0.00147	HARCJ GT	0.00171
HAR IVGT	eliminated	HAR IVGT	eliminated	HAR IVGT	eliminated	HAR IVGT	eliminated	HAR IVGT	eliminated
HARCJ IVGT	eliminated	HARCJ IVGT	eliminated	HARCJ IVGT	eliminated	HARCJ IVGT	eliminated	HARCJ IVGT	eliminated
HAR IV	eliminated	HARCJ IV	eliminated	HARCJ IV	eliminated	HARCJ IV	eliminated	HARCJ IV	eliminated
HARCJ IV	eliminated	HAR IV	eliminated	HAR IV	eliminated	HAR IV	eliminated	HAR IV	eliminated
HAR GT	eliminated	HAR GT	eliminated	HAR GT	eliminated	HAR GT	eliminated	HAR GT	eliminated

Table 6. Multinomial VaR test with probability levels
$\alpha_1 = 0.5\%, \alpha_2 = 1\%, \alpha_3 = 1.5\%, \alpha_4 = 2\%, \alpha_5 = 2.5\%$

Model	*P*-value	Model	*P*-value
TGARCH	*0.00*	ARFIMA GT	*0.00*
TGARCH IV	0.08	RG	*0.00*
TGARCH GT	*0.03*	RG IV	*0.00*
TGARCH IVGT	0.10	RG GT	*0.00*
HAR	*0.01*	RG IVGT	*0.00*
HAR IV	*0.00*	HAR LOG	*0.00*
HAR GT	*0.00*	HAR IV LOG	*0.00*
HAR IVGT	*0.00*	HAR GT LOG	*0.00*
HARCJ	*0.00*	HAR IVGT LOG	*0.00*
HARCJ IV	*0.00*	HARCJ LOG	*0.00*
HARCJ GT	*0.00*	HARCJ IV LOG	*0.00*
HARCJ IVGT	*0.00*	HARCJ GT LOG	*0.00*
ARFIMA	*0.00*	HARCJ IVGT LOG	*0.00*

Note. P-values smaller than 0.05 are in italics font.

future (see Fig. 1). Secondly, the HAR models with additional regressors performed very poorly and clearly suffered from computational problems, which resulted in VaR forecasts being strongly underestimated: particularly, the HAR models with both IV and Google data, and the HAR models with IV showed empirical violations higher than 5% and, after 2016, even higher than 15%. Using variables in logarithms solved this numerical problem, but the models' VaR violations were still quite high (between 2% and 4%) and unable to pass the Kupiec and Christoffersen tests. The few Realized-GARCH and ARFIMA models which managed to reach numerical convergence behaved similarly to HAR models with variables in logs. One of the main messages that our backtesting analysis conveys is to check the computational robustness of the model used to forecast the VaR or any other risk measure. This is important not only for the Russian market, but also for all emerging markets which may be subject to sudden market crashes due to a variety of reasons.

4.4. Robustness check: A hierarchical VAR model with LASSO

We wanted to check how our previous results changed with a multivariate model able to both accommodate a large number of regressors and to improve the model estimation and its forecasting performances. To achieve this goal, we employed the *Hierarchical Vector Autoregression* (HVAR) model estimated with the Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Nicholson et al. (2018). Let us consider the following vector autoregression:

$$Y_t = \boldsymbol{v} + \sum_{l=1}^{22} \Phi^l Y_{t-l} + \boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim WN\left(0, \Sigma_u\right),$$

where Y_t is a (4×1)-vector containing the daily returns, the daily realized volatility, the implied volatility and the Google data, $\boldsymbol{v}$ is an intercept vector, while Φ^l are the usual coefficient matrices.

The HVAR approach proposed by Nicholson et al. (2018) adds structured convex penalties to the least squares VAR problem, so that the optimization problem is given by

$$\min_{\boldsymbol{v},\Phi} \sum_{t=1}^{T} \left\| Y_t - \boldsymbol{v} - \sum_{l=1}^{22} \Phi^l Y_{t-l} \right\|_F^2 + \lambda\left(\mathcal{P}_Y(\Phi)\right),$$

where $\|A\|_F$ denotes the Frobenius norm of matrix A (that is, the elementwise 2-norm), $\lambda \geq 0$ is a penalty parameter, while $\mathcal{P}_Y(\Phi)$ is the group penalty structure on the endogenous coefficient matrices. The HVAR class of models solves the problem of an increasing maximum lag order by including the lag order into hierarchical group LASSO penalties, which induce sparsity and a low maximum lag order.

For our empirical work, we employed the *elementwise penalty function*

$$\mathcal{P}_Y(\Phi) = \sum_{i=1}^{4}\sum_{j=1}^{4}\sum_{l=1}^{22} \left\| \Phi_{ij}^l \right\|_2,$$

which is the most general structure, because every variable in every equation is allowed to have its own maximum lag resulting in 4^2 possible lag orders. The penalty parameter λ is estimated by sequential cross-validation, see Nicholson et al. (2018) for the full details. The *p*-values of the Kupiec and Christoffersen's tests, the number of violations (in %), and the *p*-values of the Multinomial VaR test by Kratz et al. (2018) with probability levels $\alpha_1 = 0.5\%$, $\alpha_2 = 1\%$, $\alpha_3 = 1.5\%$, $\alpha_4 = 2\%$, and $\alpha_5 = 2.5\%$ are reported in Table 7, while the VaR violations for the forecasted $\text{VaR}_{1\%}$ using a rolling out-of-sample of 250 days are reported in Fig. 3.

Table 7. Kupiec and Christoffersen's tests p-values, and Multinomial VaR test with probability levels $\alpha_1 = 0.5\%$, $\alpha_2 = 1\%$, $\alpha_3 = 1.5\%$, $\alpha_4 = 2\%$, $\alpha_5 = 2.5\%$

	Kupiec	Christoffersen	Violations %	Multinomial VaR test *p*-value
VaR with α = 0.5%	*0.00*	*0.00*	1.67	*0.00*
VaR with α = 1%	*0.00*	*0.00*	2.02	
VaR with α = 1.5%	*0.00*	*0.00*	2.30	
VaR with α = 2%	*0.01*	*0.00*	2.69	
VaR with α = 2.5%	*0.02*	*0.00*	3.19	

Note. P-values smaller than 0.05 are in italics font.

The HVAR model solves the numerical problems of the HAR model with additional variables in levels, but it has a backtesting performance similar to the HVAR models with variables in logarithms: that is, it underestimates the VaR following episodes of extremely high volatility.

5. Conclusions

We evaluated the contribution of both online search intensity and options-based implied volatility to the modelling of the volatility of the Russian RTS index future, and we examined how this dependence changed over almost two decades. We found that both the sign and the size of their

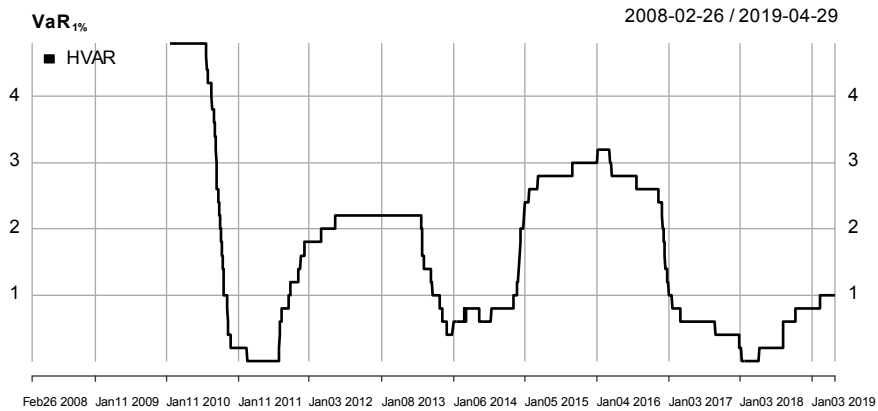


Fig. 3. Violations (in %) for the forecasted $VaR_{1\%}$ using a rolling out-of-sample of 250 days

coefficients changed significantly, particularly in the periods following the beginning of the global financial crisis in 2008 and (to a lower degree) after the introduction of sanctions in 2014.

We then performed a backtesting analysis involving the forecasting of the Value-at-Risk for the RTS index future at multiple confidence levels using several alternative models specifications, with and without Google data and implied volatility. We found that only TGARCH models were able to pass the Kupiec and Christoffersen's tests for most quantiles, and they also reported the lowest asymmetric losses for all quantiles up to the 2% probability level. However, only the TGARCH model with implied volatility managed to pass almost all back-tests, including the multinomial test with five quantiles needed to back-test the expected shortfall, whereas TGARCH models with Google Trends or without any external regressors did not. We noticed that when both the implied volatility and Google data were added jointly, the parameters estimates of several models became very unstable and several models did not reach numerical convergence (particularly, ARFIMA and Realized-GARCH models). Moreover, with the exception of TGARCH models, our results highlighted that simpler models with no additional regressors provided better VaR forecasts than augmented models. This empirical evidence complements the results provided by Bams et al. (2017), who showed that forecasting the volatility is different from forecasting a certain quantile of the return distribution, hence models forecasting well the former may not forecast well the latter. However, in the case of Russian markets, TGARCH models augmented with IV did provide better VaR forecasts than TGARCH models without it. We also evaluated the backtesting performance of the competing models using a rolling out-of-sample of 250 days: we found that the performance of TGARCH models remained remarkably stable over the full evaluation period, whereas HAR models with additional regressors performed very poorly and clearly suffered from computational problems, which resulted in VaR forecasts being strongly underestimated. Using variables in logarithms solved this numerical problem, but the models' VaR violations were still quite high and unable to pass the usual Kupiec and Christoffersen tests. The few Realized-GARCH and ARFIMA models which managed to reach numerical convergence behaved similarly to HAR models with variables in logs. Therefore, one of the main guidance that emerged from our backtesting analysis is to check the computational robustness of the model employed to forecast the VaR (or any other risk measure) in case of extreme and sudden market crashes. Finally, we also performed a robustness check to verify how

our previous results changed with a Hierarchical-VAR model with LASSO able to both accommodate a large number of regressors and to improve the model estimation and its forecasting performances. The HVAR model solved the numerical problems of the HAR models with additional variables in levels, but it still underestimated the VaR in the periods following episodes of extremely high volatility and abrupt market changes.

In general, models with implied volatility performed better than models with Google Trends data, thus confirming similar evidence reported by Basistha et al. (2019) and Bazhenov, Fantazzini (2019). These authors suggested two possible explanations for these results: first, the informational content included in Google search activity is also present in the implied volatility, but the opposite is not true, due to the fact that implied volatility is a forward-looking measure based on the expectations of large investors who have access to premium and insider information, while Google Trends data are mainly based on the expectations of small investors and uninformed traders. A second simpler explanation is that Yandex is the main search engine in Russia, so that Google Trends may not be the best proxy for Russian investors' interest and behavior. If Yandex will make available online search data at the daily frequency and for long periods, then this issue will definitely be an interesting avenue of future research.

References

Acerbi C., Tasche D. (2002). On the coherence of expected shortfall. *Journal of Banking and Finance*, 26 (7), 1487–1503.

Aganin A. (2017). Forecast comparison of volatility models on Russian stock market. *Applied Econometrics*, 48, 63–84 (in Russian).

Aganin A., Peresetsky A. (2018). Volatility of ruble exchange rate: Oil and sanctions. *Applied Econometrics*, 52, 5–21 (in Russian).

Agnolucci P. (2009). Volatility in crude oil futures: A comparison of the predictive ability of GARCH and implied volatility models. *Energy Economics*, 31 (2), 316–321.

Ahoniemi K., Lanne M. (2013). Overnight stock returns and realized volatility. *International Journal of Forecasting*, 29 (4), 592–604.

Andersen T. G., Bollerslev T., Diebold F. X., Ebens H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61 (1), 43–76.

Andersen T. G., Bollerslev T., Diebold F. X., Labys P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71 (2), 579–625.

Andersen T. G., Bollerslev T., Diebold F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics*, 89 (4), 701–720.

Artzner P., Delbaen F., Eber J. M., Heath D. (1997). Thinking coherently. *Risk*, 10 (11), 68–71.

Artzner P., Delbaen F., Eber J. M., Heath D. (1999). Coherent measures of risk. *Mathematical Finance*, 9 (3), 203–228.

Bai J. (1997). Estimating multiple breaks one at a time. *Econometric Theory*, 13, 315–352.

Bai J., Perron P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66, 47–78.

Bai J., Perron P. (2003a). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18, 1–22.

Bai J., Perron P. (2003b). Critical values for multiple structural change tests. *Econometrics Journal*, 18, 1–22.

Bams D., Blanchard G., Lehnert T. (2017). Volatility measures and Value-at-Risk. *International Journal of Forecasting*, 33 (4), 848–863.

Barndorff-Nielsen O. E., Shephard N. (2004a). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72 (3), 885–925.

Barndorff-Nielsen O. E., Shephard N. (2004b). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2 (1), 1–37.

Barndorff-Nielsen O. E., Shephard N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, 4 (1), 1–30.

Barone-Adesi G., Finta M. A., Legnazzi C., Sala C. (2019). WTI crude oil option implied VaR and CVaR: An empirical application. *Journal of Forecasting*, https://doi.org/10.1002/for.2580.

Basel Committee on Banking Supervision (2009). *Findings on the interaction of market and credit risk*. Bank for International Settlements, Working paper no. 16, May.

Basel Committee on Banking Supervision (2013). *Fundamental review of the trading book: A revised market risk framework*. Consultative Document, October.

Basel Committee on Banking Supervision (2016). *Minimum capital requirements for market risk*. Consultative Document, January.

Basistha A., Kurov A., Wolfe M. (2019). Volatility forecasting: The role of internet search activity and implied volatility. *Working paper*. West Virginia University. http://dx.doi.org/10.2139/ssrn.2812387.

Bazhenov T., Fantazzini D. (2019). Forecasting realized volatility of Russian stocks using Google Trends and implied volatility. *Russian Journal of Industrial Economics*, 12 (1), 79–88.

Bianchi C., De Giuli M. E., Fantazzini D., Maggi M. (2011). Small sample properties of copula-GARCH modelling: A Monte Carlo study. *Applied Financial Economics*, 21 (21), 1587–1597.

Birkelund O. H., Haugom E., Molnár P., Opdal M., Westgaard S. (2015). A comparison of implied and realized volatility in the Nordic power forward market. *Energy Economics*, 48, 288–294.

Busch T., Christensen B. J., Nielsen M. (2011). The role of implied volatility in forecasting future realized volatility and jumps in foreign exchange, stock, and bond markets. *Journal of Econometrics*, 160 (1), 48–57.

Cai Y., Krishnamoorthy K. (2006). Exact size and power properties of five tests for multinomial proportions. *Communications in Statistics. Simulation and Computation*, 35 (1), 149–160.

Campos I., Cortazar G., Reyes T. (2017). Modeling and predicting oil VIX: Internet search volume versus traditional variables. *Energy Economics*, 66, 194–204.

Caporale G. M., Gil-Alana L. A., Tripathy T. (2019). Volatility persistence in the Russian stock market. *Finance Research Letters*, https://doi.org/10.1016/j.frl.2019.06.014.

Choi H., Varian H. (2012). Predicting the present with Google Trends. *Economic Record*, 88, 2–9.

Chong J. (2004). Value at risk from econometric models and implied from currency options. *Journal of Forecasting*, 23 (8), 603–620.

Christensen B. J., Prabhala N. R. (1998). The relation between implied and realized volatility. *Journal of Financial Economics*, 50 (2), 125–150.

Christoffersen P. (1998). Evaluating interval forecasts. *International Economic Review*, 39, 841–862.

Christoffersen P. (2012). *Elements of financial risk management*. Academic Press.

Christoffersen P., Mazzotta S. (2005). The accuracy of density forecasts from foreign exchange options. *Journal of Financial Econometrics*, 3 (4), 578–605.

Corredor P., Santamaría R. (2004). Forecasting volatility in the Spanish option market. *Applied Financial Economics*, 14 (1), 1–11.

Corsi F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7 (2), 174–196.

Da Z., Engelberg J., Gao P. (2011). In search of attention. *Journal of Finance*, 66 (5), 1461–1499.

Dimpfl T., Jank S. (2016). Can internet search queries help to predict stock market volatility? *European Financial Management*, 22 (2), 171–192.

Emmer S., Kratz M., Tasche D. (2015). What is the best risk measure in practice? *Journal of Risk,* 18, 31–60.

Fantazzini D. (2019). *Quantitative finance with R and cryptocurrencies*. Amazon KDP.

Fantazzini D., Toktamysova Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics*, 170, 97–135.

Ginsberg J., Mohebbi M. H., Patel R. S., Brammer L., Smolinski M. S., Brilliant L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457 (7232), 1012–1014.

Giot P. (2005). Implied volatility indexes and daily value at risk models. *The Journal of Derivatives*, 12 (4), 54–64.

Glosten L., Jagannathan R., Runke D. (1993). Relationship between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48, 1779–1801.

Gneiting T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106 (494), 746–762.

Goddard J., Kita A., Wang Q. (2015). Investor attention and FX market volatility. *Journal of International Financial Markets, Institutions and Money,* 38, 79–96.

González-Rivera G., Lee T. H., Mishra S. (2004). Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. *International Journal of Forecasting*, 20 (4), 629–645.

Hamid A., Heiden M. (2015). Forecasting volatility with empirical similarity and Google Trends. *Journal of Economic Behavior and Organization*, 117, 62–81.

Hansen P. R., Lunde A. (2005a). A realized variance for the whole day based on intermittent high-frequency data. *Journal of Financial Econometrics*, 3, 525–554.

Hansen P. R., Lunde A. (2005b). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20 (7), 873–889.

Hansen P. R., Lunde A., Nason J. M. (2011). The model confidence set. *Econometrica*, 79 (2), 453–497.

Hansen P. R., Huang Z., Shek H. H. (2012). Realized GARCH: A joint model for returns and realized measures of volatility. *Journal of Applied Econometrics*, 27 (6), 877–906.

Hartmann P. (2010). Interaction of market and credit risk. *Journal of Banking and Finance*, 34 (4), 697–702.

Haugom E., Langeland H., Molnár P., Westgaard S. (2014). Forecasting volatility of the US oil market. *Journal of Banking and Finance*, 47, 1–14.

Huang X., Tauchen G. (2005). The relative contribution of jumps to total price variance. *Journal of Financial Econometrics*, 3 (4), 456–499.

Hyndman R. J., Khandakar Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27 (3), 1–22.

Hwang S., Valls Pereira P. L. (2006). Small sample properties of GARCH estimates and persistence. *The European Journal of Finance*, 12 (6–7), 473–494.

Jeon J., Taylor J. W. (2013). Using CAViaR models with implied volatility for Value-at-Risk estimation. *Journal of Forecasting*, 32 (1), 62–74.

Jorion P. (2007). *Financial risk manager handbook*. Vol. 406. Wiley.

Kratz M., Lok Y. H., McNeil A. J. (2018). Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. *Journal of Banking and Finance*, 88, 393–407.

Kupiec P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3 (2), 73–84.

Liu J., Wu S., Zidek J. V. (1997). On segmented multivariate regression. *Statistica Sinica,* 7, 497–525.

Liu L. Y., Patton A. J., Sheppard K. (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics*, 187 (1), 293–311.

Malakhovskaya O., Minabutdinov A. (2014). Are commodity price shocks important? A Bayesian estimation of a DSGE model for Russia. *International Journal of Computational Economics and Econometrics*, 4 (1/2), 148–180.

Martens M., Zein J. (2004). Predicting financial volatility: High-frequency time-series forecasts vis-à-vis implied volatility. *Journal of Futures Markets*, 24 (11), 1005–1028.

Mayhew S. (1995). Implied volatility. *Financial Analysts Journal*, 51 (4), 8–20.

McNeil A. J., Frey R., Embrechts P. (2015). *Quantitative risk management: Concepts, techniques and tools*. Princeton University Press.

Meddahi N. (2002). A theoretical comparison between integrated and realized volatility. *Journal of Applied Econometrics*, 17 (5), 479–508.

Newey W. K., West K. D. (1994). Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61 (4), 631–653.

Nicholson W. B., Wilms I., Bien J., Matteson D. S. (2018). High dimensional forecasting via interpretable vector autoregression. *ArXiv preprint*, arXiv:1412.5250.

Perron P. (2006). Dealing with structural breaks. *Palgrave Handbook of Econometrics*, 1 (2), 278–352.

Pesaran M. H., Timmermann A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137 (1), 134–161.

Pong S., Shackleton M. B., Taylor S. J., Xu X. (2004). Forecasting currency volatility: A comparison of implied volatilities and AR (FI) MA models. *Journal of Banking and Finance*, 28 (10), 2541–2563.

Seo M., Lee S., Kim G. (2019). Forecasting the volatility of stock market index using the hybrid models with google domestic trends. *Fluctuation and Noise Letters*, 18 (01), 1950006.

Taylor S. J., Xu X. (1997). The incremental volatility information in one million foreign exchange quotations. *Journal of Empirical Finance*, 4 (4), 317–340.

Vlastakis N., Markellos R. N. (2012). Information demand and stock market volatility. *Journal of Banking and Finance*, 36 (6), 1808–1821.

Vozlyublennaia N. (2014). Investor attention, index performance, and return predictability. *Journal of Banking and Finance*, 41, 17–35.

Xu Q., Bo Z., Jiang C., Liu Y. (2019). Does Google search index really help predicting stock market volatility? Evidence from a modified mixed data sampling model on volatility. *Knowledge-Based Systems*, 166, 170–185.

Wimmerstedt L. (2015). Backtesting expected shortfall: The design and implementation of different backtests. Technical report, Swedish Royal Institute of Technology.

Yao Y. C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics and Probability Letters*, 6 (3), 181–189.

Zakoian J. M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and Control*, 18 (5), 931–955.

Zeileis A., Leisch F., Hornik K., Kleiber C. (2002). Strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7 (2), 1–38.

Zeileis A. (2005). A unified approach to structural change tests based on ML scores, F statistics, and OLS residuals. *Econometric Reviews*, 24 (4), 445–466.

Zeileis A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16 (9), 1–16.

Zeileis A., Shah A., Patnaik I. (2010). Testing, monitoring, and dating structural changes in exchange rate regimes. *Computational Statistics and Data Analysis*, 54 (6), 1696–1706.

Received 19.08.2019; accepted 10.10.2019.

Appendix 18 "A multivariate approach for the simultaneous modelling of market risk and credit risk for cryptocurrencies."

Fantazzini, D. Zimin, S. (2020) A multivariate approach for the simultaneous modelling of market risk and credit risk for cryptocurrencies, *Journal of Industrial and Business Economics,* forthcoming DOI: https://doi.org/10.1007/s40812-019-00136-8

Permisson to post the Publisher's version/PDF: according to https://www.springer.com/gp/open-access/publication-policies/self-archiving-policy, it cannot be posted online, even in the case when it is included into a dissertation. ⇒ Instead, the authors' accepted manuscript must be used for online dissemination.

A multivariate approach for the simultaneous modelling of market risk and credit risk for cryptocurrencies *

Dean Fantazzini† Stephan Zimin ‡

Abstract

This paper proposes a set of models which can be used to estimate the market risk for a portfolio of crypto-currencies, and simultaneously to estimate also their credit risk using the Zero Price Probability (ZPP) model by Fantazzini et al (2008), which is a methodology to compute the probabilities of default using only market prices. For this purpose, both univariate and multivariate models with different specifications are employed. Two special cases of the ZPP with closed-form formulas in case of normally distributed errors are also developed using recent results from barrier option theory. A backtesting exercise using two datasets of 5 and 15 coins for market risk forecasting and a dataset of 42 coins for credit risk forecasting was performed. The Value-at-Risk and the Expected Shortfall for single coins and for an equally weighted portfolio were calculated and evaluated with several tests. The ZPP approach was used for the estimation of the probability of default/death of the single coins and compared to classical credit scoring models (logit and probit) and to a machine learning algorithm (Random Forest). Our results reveal the superiority of the t-copula/skewed-t GARCH model for market risk, and the ZPP-based models for credit risk.

Keywords: cryptocurrencies, market risk, credit risk, ZPP.

JEL classification: C32, C51, C53, C58, G12, G17, G32, G33.

Journal of Industrial and Business Economics, forthcoming

The published version is available at `https://link.springer.com/article/10.1007/s40812-019-00136-8`

*We would like to thank all the participants of the International Conference on Applied Research in Economics (ICare), which was held in September 2018 at the Higher School of Economics in Perm (Russia). We also want to thank an anonymous founder of a crypto-exchange, a cryptocurrency portfolio manager and several professional cryptocurrency traders, who provided important feedback for section 2. The second-named author gratefully acknowledges financial support from the Russian Academic Excellence Project '5-100'.

†Moscow School of Economics, Moscow State University, Leninskie Gory, 1, Building 61, 119992, Moscow, Russia. Fax: +7 4955105256 . Phone: +7 4955105267 . E-mail: `fantazzini@mse-msu.ru` .

‡Higher School of Economics, Moscow (Russia); smzimin@edu.hse.ru.

This is the working paper version of the paper *A multivariate approach for the simultaneous modelling of market risk and credit risk for cryptocurrencies*, forthcoming in the Journal of Industrial and Business Economics.

1 Introduction

Cryptocurrencies and the cryptomarket have become a popular trend in finance in the last years, as discussed by Antonopoulos (2014), Narayanan et al. (2016), Burniske and Tatar (2017), Fantazzini (2019) and Corbet et al. (2019). Bohr and Bashir (2014) and Yelowitz and Wilson (2015) showed that different people, from crypto and tech enthusiasts to investors, are interested in these new financial tools. Large market capitalizations for the most popular cryptocurrencies, increasing number of crypto funds[1], a growing cumulative number of Initial Coin Offerings (ICOs) and, at the same time, a frequent number of hacks and frauds, make the topic of cryptocurrencies' risk a really urgent problem, see Fantazzini et al. (2016), Fantazzini et al. (2017) and references therein. Hence, the main goal of this paper is to propose a unified framework for the simultaneous modelling of credit and market risk. If we use the formal definition of cryptocurrency proposed by Lansky (2018), then a cryptocurrency can be defined as a system that satisfies these six conditions:

> "*1) The system does not require a central authority, its state is maintained through distributed consensus. 2) The system keeps an overview of cryptocurrency units and their ownership. 3) The system defines whether new cryptocurrency units can be created. If new cryptocurrency units can be created, the system defines the circumstances of their origin and how to determine the ownership of these new units. 4) Ownership of cryptocurrency units can be proved exclusively cryptographically. 5) The system allows transactions to be performed in which ownership of the cryptographic units is changed. A transaction statement can only be issued by an entity proving the current ownership of these units. 6) If two different instructions for changing the ownership of the same cryptographic units are simultaneously entered, the system performs at most one of them.*"
>
> — Lansky (2018, p. 19)

Therefore, a cryptocurrency does not have debt, and it cannot default in a classical sense. However, its price and the investors' demand might drop dramatically because of a revealed scam, hack or other hidden problems that cannot be observed directly from the market data. Because of that, we believe that such risk is not exactly a market one and therefore we propose a definition of credit risk which is somewhat different from the classic one: credit risk for a crypto-coin is its "*death*", a situation when its price drops significantly and a coin becomes illiquid. In this regard, we have to admit that there is not an unambiguous definition of a dead coin, neither in the professional literature, nor in the academic literature. However, it is worth noting that even when considered dead, some coins do still have some negligible daily trading volumes. There are two reasons for this phenomenon: the possibility to recover

[1] `https://cryptofundresearch.com/cryptocurrency-funds-overview-infographic/`

at least a small amount of the initial investment, or to bet on the possible revamp of the dead coin. Differently from stock companies and small and medium enterprises (SMEs), the procedure to revamp a coin is much easier and faster: it only takes a new code or, even easier, an update of the previous code where the major flaws are corrected. This is why the "death" state for a cryptocurrency can be only a temporary situation rather than a permanent one: see, for example, Sid (2018) who discussed the story of *PENG*, which is a coin revamp of an old abandoned project. A brief review of the history and the financial literature devoted to cryptocurrencies is reported in Appendix A.

The first contribution of the paper is a large discussion about credit risk and market risk for cryptocurrencies and how these risks can be defined with these assets. In traditional finance, credit risk is usually defined as the gains and losses on a position (or portfolio) associated with the fulfillment (or not) of contractual obligations, whereas market risk as the gains and losses on the value of a position (or portfolio) due to the changes in market prices (Basel Committee on Banking Supervision (2009) p. 6-7, Hartmann (2010), p. 697). Instead, we argue that credit risk for cryptocurrencies can be defined as the gains and losses on the value of a position of a cryptocurrency that is abandoned and considered dead but which can be potentially revived, while market risk can be described as the gains and losses on the value of a position (or portfolio) of alive cryptocurrencies, due to the movements in market prices in centralized and decentralized exchanges.

The second contribution of the paper is a set of multivariate models which can be used to estimate the market risk for a portfolio of crypto-currencies by using the Value-at-Risk and the Expected Shortfall, and simultaneously to estimate also their credit risk using the Zero Price Probability (ZPP) model by Fantazzini et al. (2008), which is a methodology to compute the probabilities of default using only market prices. Recent papers by Su and Huang (2010), Li et al. (2016) and Dalla Valle et al. (2016) showed the ZPP dominance in terms of default probability estimation with respect to competing models.

The third contribution of this work is the development of closed-form formulas for the ZPP in two special cases, namely the random walk with drift and a GARCH(1,1) model with normal errors, using recent results from barrier option theory by Su and Rieger (2009). Even though crypto-assets are far from being normally distributed, these closed-form formulas can provide a quick estimate of the probability of the coin death and they can give an investor at least a rough idea of the crypto-asset credit risk.

The fourth contribution of the paper is a backtesting exercise using two datasets of 5 and 15 coins for market risk forecasting and a dataset of 42 coins for credit risk forecasting. For the purpose of the multivariate modelling of a portfolio of cryptocurrencies, we employ VAR-DCC and VAR-Copula-GARCH models with different specifications for the error terms. The Value-at-Risk and the Expected Shortfall for the single coins and for an equally weighted portfolio are calculated during a back-testing exercise and then evaluated with several tests. The ZPP approach is used for the estimation of the

probability of default/death for the single coins and compared to classical credit scoring models (logit and probit) and to a machine learning algorithm (Random Forest). Our results show that a t-copula/skewed-t GARCH model outperform the competing models for market risk, while ZPP-based models outperform the other models for credit risk.

The paper is organized as follows: the definitions of credit risk and market risk for cryptocurrencies are introducted in Section 2, while the methods proposed for market and credit risk modelling with cryptocurrencies are discussed in Section 3. Section 4 describes the empirical results, while robustness checks are discussed in Section 5. Section 6 briefly concludes.

2 Theoretical framework: Credit and Market risk for cryptocurrencies

The literature review in Appendix A shows that risk management is still in its early stages when cryptocurrencies are of concern. In this regard, there are important differences between cryptocurrency trading and traditional stock and forex trading:

- The *lack of financial oversight for many crypto-based companies and exchanges* means that coins prices can be susceptible to manipulations, pump and dump schemes and market frauds of various types. For example, Gandal et al. (2018) showed that the quick rise in the bitcoin price in 2013 from $150 to more than $1000 was probably due to price manipulation. Griffin and Shams (2018) showed that Tether, a digital currency pegged to US dollars, is likely used to provide price support and manipulate cryptocurrency prices. Moreover, the lack of regulatory oversight is one of the main reasons why there are large and recurring deviations in cryptocurrency prices across exchanges, as shown by Makarov and Schoar (2018)[2].

- *There are no traditional brokers*, and investors can trade cryptocurrencies directly either using centralized exchanges or decentralized exchanges[3].

- Differently from forex and stock markets, *cryptocurrency exchanges do not provide neither cash nor asset insurance* (with some exceptions like Coinbase and Gemini exchanges, which only insure cash deposits): in case of hacking attack, exchanges could lose part or all of investors' cryptocurrencies, and investors would not enjoy any type of government protection. Similar fate would await investors in the case the cryptocurrency turns out to be a scam. Note that if hackers steal the investors'

[2] A table showing the real-time price difference between the last trades across several bitcoin exchanges can be freely accessed at `https://data.bitcoinity.org/markets/arbitrage`

[3] A list of both centralized and decentralized exchanges can be found at `list.wiki/Cryptocurrency_Exchanges`. More information about decentralized exchanges is available at `github.com/distribuyed/index` and references therein.

private keys from the cryptocurrency exchange's wallet, then the investors will permanently lose their assets because cryptocurrency transactions are irreversible (due to the specific nature of a cryptocurrency blockchain): suing the exchange will not help because it will simply declare bankruptcy, leaving the investors with their losses. Therefore, choosing a good exchange is very important when trading cryptocurrencies.

Given this discussion, the separation between market and credit risk becomes even more blurred when dealing with cryptocurrencies than in traditional finance. In the latter case, *credit risk* is broadly defined as the gains and losses on a position or portfolio associated with the fulfillment (or not) of contractual obligations, whereas *market risk* is defined as the gains and losses on the value of a position or portfolio that can take place due to the movements in market prices (such as exchange rates, commodity prices, interest rates, etc.), see Basel Committee on Banking Supervision (2009), Hartmann (2010) and references therein for more details. However, the Basel Committee on Banking Supervision (2009) highlighted that the securitization trend in the last decade "*has diminished the scope for differences in measuring market and credit risk, as securitization transforms the latter into the former*"(Basel Committee on Banking Supervision (2009), p.14). Moreover, market and credit risk are driven by the same economic factors, as a large literature shows –see the special issue on the interaction of market and credit risk in the *Journal of Banking and Finance* in 2010–, so it is difficult to separate these two risks neatly in practical risk measurement and management.

In general, the definition of market risk can be extended rather easily to the case of cryptocurrencies market prices. Instead, credit risk requires a new definition because a cryptocurrency does not pay interest and does not involve the reimbursement of principal amounts, so the previous classical definition cannot be used. To this aim, the concept of a *dead coin* needs to be introduced: in simple terms, a dead coin is a cryptocurrency that does not offer any utility or feature, or it is no more developed and mining ceased, or it was created for hacking or for spreading of malware, or it was purely a scam, or simply a fun/parody project.

Unfortunately, there is not a unique definition of dead coins, neither in the professional literature, nor in the academic literature: in the professional literature, some define dead coins as those whose value drops below 1 cent[4], yet others stress, on top of that, no trading volume, no nodes running, no active community and de-listing from all exchanges[5]. In the academic literature, Feder et al. (2018) proposed the only definition (currently) available: they first define a "candidate peak" as a day in which the 7-day rolling price average is greater than any value 30 days before or after. Moreover, to choose only those peaks with sudden jumps, they define a candidate as a peak only if it is greater than or equal 50% of the

[4]https://www.investopedia.com/terms/d/dead-coin.asp
[5]https://steemit.com/beyondbitcoin/@freshfund/dead-coins-or-dormant-coins

minimum value in the 30 days prior to the candidate peak, and if its value is at least 5% as large as the cryptocurrency's maximum peak. Given these peak data, Feder et al. (2018) consider a coin *abandoned* (= dead), if the daily average volume for a given month is less than or equal to 1% of the peak volume. Besides, if the average daily trading volume for a month following a peak is greater than 10% of the peak value and that currency is currently abandoned, then Feder et al. (2018) change the coin status to *resurrected*.

Despite the differences in these criteria, they have one thing in common: the "death" state is not a permanent state and a dead coin can be revamped/revived, even several times during a relatively short period. For example, Feder et al. (2018) found that 44% of publicly traded coins are subsequently abandoned (18% permanently), while 85% of announced coins fail before being traded publicly. Moreover, they also found that most coins are not traded much, and these coins are more likely to die than their larger counterparts are. Interestingly, they showed that some coins could be revamped up to five times over a 5-year period.

Given this background, *credit risk for cryptocurrencies* can be defined as *the gains and losses on the value of a position of a cryptocurrency that is abandoned and considered dead according to professional and/or academic criteria, but which can be potentially revived and revamped*[6].

Substantially, the differences between credit and market risk in case of cryptocurrencies are of quantitative and temporal nature, not qualitative: if the financial losses and the technical problems can be dealt with the available financial and technical resources, then we have a market event. If the financial losses are too big and the technical problems cannot be solved, then we have a credit event and the cryptocurrency "dies". Moreover, it follows easily that the longer the time horizon, the more probable are large losses and/or technical problems and/or hacking attacks on the cryptocurrency blockchain, so credit risk becomes more important. Note that this latter result is already known in the classical financial literature, given that "*... the ratio of default and (normally distributed) market risk losses is proportional to the square-root of the holding period. Since the ratio goes to 0 as the holding period goes to 0, over short horizons market risk is relatively more important, while over longer horizons losses due to default become more important*"(Basel Committee on Banking Supervision (2009), p.16-17).

Once a credit event takes place, the development of a cryptocurrency stops, its price falls close to zero (or even to zero, if we consider the uninterrupted lack of trading for several days and weeks as evidence of a zero price[7]), but trading may still continue for the reasons discussed in the introduction. At a first glance, this situation may be similar to the so-called "zombie firms": these firms are characterized by

[6]In this paper, we deal only with the credit risk arising from the death of a cryptocurrency. The credit risk due to the possibility that a crypto-exchange is hacked and/or goes bankrupt is examined by Moore and Christin (2013) and Moore et al. (2018).

[7]Crypto-exchanges work 24 hours a day, 365 days a year.

a lack of profitability over an extended period, are rather old and their future expected profitability is rather low. There are no unique criteria to define a zombie firm but, in general, the ability to barely repay the interest on its debts coupled with the inability to repay the principal are common to several criteria proposed in the literature, see Caballero et al. (2008), Adalet et al. (2018) and Banerjee and Hofmann (2018). However, a dead cryptocurrency is a rather different phenomenon from a zombie firm: there is no financial institution providing financial support to continue the project, and the trading of the coin can continue even without any project development. Instead, a zombie firm is kept alive to guarantee that the firm continues its activities, thus saving jobs places.

Given the importance of the zero price barrier to determine the death of a cryptocurrencies, it would be natural to consider reduced-form and hybrid credit risk models in which the asset price jumps to zero at default: see, for example, Linetsky (2006), Carr and Linetsky (2006), Campi et al. (2009), Das and Hanouna (2009), Carr and Wu (2009), Bayraktar and Yang (2011) and Dyrssen et al. (2014). Unfortunately, all these models are based on derivatives data, bond data and/or accounting data, which are not available for cryptocurrencies[8]. This is why a different approach based only on market data is required.

Finally, Table 1 summarizes the main aspects of credit and market risk for cryptocurrencies[9].

	Market risk	**Credit risk**
Definition	Gains and losses on the value of a position or portfolio of ***alive*** cryptocurrencies, that can take place due to the movements in market prices in centralized and decentralized exchanges.	Gains and losses on the value of a position of a cryptocurrency that is ***abandoned and considered dead*** according to professional and/or academic criteria.
Differences from traditional finance	• Lack of financial oversight means that coins prices can be susceptible to manipulations, pump and dump schemes and other market frauds; • Lack of regulatory oversight also explain why prices can differ widely across exchanges; • Cryptocurrency exchanges do not provide neither cash nor asset insurance (but there are exceptions).	• Dead coins can be revamped several times; • Dead coins are very different from "zombie firms"; • Traditional credit risk models cannot be used due to the (current) lack of derivatives data, bond data and/or accounting data.

Table 1: Main aspects of credit and market risk for cryptocurrencies

[8]The CME and the CBOT introduced the first futures on bitcoin in December 2017, whereas options on cryptocurrencies are (currently) traded only on small and illiquid exchanges, with poor or no financial oversight.

[9]Updated lists of dead coins can be found at https://deadcoins.com, www.coinopsy.com/dead-coins. The first site employs a broad definition of dead coins, whereas the second site has stricter selection criteria.

3 Methods: Market and Credit risk models

We first briefly review the Value-at-Risk (VaR) and the Expected Shortfall (ES), as well as a set of statistical tests to measure the goodness-of-fit of these market risk measures from different models after a backtesting exercise[10]. Similarly, we will present several methods to compute the probability of death for a crypto-coin to measure its credit risk, as well as several metrics to evaluate the estimated death probabilities. Two special cases of the ZPP with closed-form formulas in case of normally distributed asset returns will be developed using recent results from barrier option theory. The multivariate time series models used to simultaneously estimate the market and credit risk for a portfolio of crypto-currencies are discussed in Appendix B.

3.1 Market Risk

Market risk arises due to movement in market variables − asset prices, interest rates, FX rates, or in our case − cryptocurrencies prices. In this work, we consider two measures of market risk: the Value-at-Risk (VaR) and the Expected Shortfall (ES).

Value-at-Risk. The Value at Risk at level α (VaR_α) is the minimum loss of the α worst losses an investor can expect to lose over a specific period of time. More formally, if we call $\Delta W(l)$ the change in the portfolio value from time t to $t+l$ and $F_l(x)$ the cumulative distribution function ("cdf") of ΔW_{t+l}, the VaR of a long position over the time horizon l with probability α can be defined as follows:

$$\alpha = Pr[\Delta W(l) \leq VaR_{\alpha,t+l}] = F_l(VaR_{\alpha,t+l}).$$

where VaR is defined as a negative value (loss), even though a part of the financial literature defines it as a positive value, see e.g. Jorion (2006) and Christoffersen (2011). If the cdf is known, then the VaR is its α-quantile times the value of the financial position. The previous definition of the VaR continues to apply to a short position if one uses the distribution of $-\Delta W(l)$, see Jorion (2007) for more details.

If we consider the profit and losses (P&L) for a single coin or for a portolio of crypto-currencies and the normal distribution is used, then the $\text{VaR}_{\alpha,t+l}$ is given by

$$VaR_{\alpha,t+l} = \mu_{t+l} + \sigma_{t+l}\,\Phi_\alpha^{-1}$$

where Φ_α^{-1} is the α-quantile of the standard normal distribution, μ_{t+l} is the forecasted mean of the P&L

[10] Backtesting is the process of assessing the performance of a model, by applying this model to a historical dataset to verify how accurately it would have predicted actual results. See Christoffersen (2011) and McNeil et al. (2015) for a discussion at the textbook level.

distribution at time $t+l$, and σ_{t+l} is the forecasted standard deviation of the P&L distribution at time $t+l$. If the Student's t distribution is used, then the $\text{VaR}_{\alpha,t+l}$ is given by

$$VaR_{\alpha,t+l} = \mu_{t+l} + s_{t+l}\, t^{-1}_{\alpha,\nu}$$

where s_{t+l} is the scale parameter of the Student's t distribution at time $t+l$ given by $s_{t+l} = \sigma_{t+l}\sqrt{(\nu-2)/\nu}$, while $t^{-1}_{\alpha,\nu}$ is the α-quantile of the Student's t distribution with ν degrees of freedom. If copula models are used, simulation methods have to be implemented, see Fantazzini (2008) - section 3.2 for a detailed description, and McNeil et al. (2015) for a general discussion.

Backtesting methods for the Value at Risk. For backtesting purposes, it is common practice to split the dataset into two parts (training set and testing set), and to perform a rolling window estimation to compute at each time step the forecasted $VaR_{\alpha,t+l}$ at the desired time horizon l and probability level α. The VaR forecasts and the actual realized series are then compared using the following back-testing techniques:

- Kupiec (1995) 's unconditional coverage test;
- Christoffersen (1998) 's conditional coverage test;
- Loss functions to evaluate VaR forecasts accuracy and to select the best models among a group of competing models using the model confidence set by Hansen, Lunde, and Nason (2011).

The *Kupiec's test* (also known as the *Unconditional Coverage test*) tests the difference between the observed and the expected number of VaR violations. Since the VaR is based on a confidence level α, when we observe T_1 actual exceedances out of T observations (with $T = T_0 + T_1$ and T_0 the number of no VaR violations), we observe an empirical frequency of $\hat{\pi} = T_1/T$ proportion of excessive losses. The Kupiec's test checks whether $\hat{\pi}$ is statistically significantly different from α. It is possible to show that the test of the null hypothesis $H_0 : \hat{\pi} = \alpha$ is given by the following likelihood ratio test statistic, see Kupiec (1995) for details:

$$LR_{uc} = -2\ln\left[(1-\alpha)^{T_0}\alpha^{T_1} \,/\, \{(1-T_1/T)^{T_0}(T_1/T)^{T_1}\}\right] \overset{H_0}{\sim} \chi^2_1$$

Christoffersen (1998) developed a likelihood ratio statistic to test the joint assumption of unconditional coverage and independence of VaR exceedances. The main advantage of this test named the *Conditional Coverage test* is that it can reject a VaR model that forecasts either too many or too few clustered exceedances, while its main limit is the need of (at least) several hundred observations to be accurate.

The test statistic is computed as follows:

$$LR_{cc} = -2\ln\left[(1-\alpha)^{T_0}\alpha^{T_1}\right] + 2\ln\left[(1-\pi_{01})^{T00}\pi_{01}^{T01}(1-\pi_{11})^{T10}\pi_{11}^{T11}\right] \overset{H_0}{\sim} \chi_2^2$$

where T_{ij} is the number of observations with value i followed by j for $i, j = 0, 1$ and

$$\pi_{ij} = \frac{T_{ij}}{\sum_j T_{ij}}$$

are the corresponding probabilities.

Loss functions are useful tools to compare the costs of different admissible choices. We will use the *asymmetric VaR loss function* proposed by Gonzalez-Rivera et al. (2004), which penalizes more heavily the realized losses below the α-th quantile level, that is $y_{t+l} < VaR_{\alpha,t+l}$:

$$l(y_{t+l}, VaR_{\alpha,t+l}) = (\alpha - d_{t+l}^{\alpha})(y_{t+l} - VaR_{\alpha,t+l})$$

where $d_{t+l}^{\alpha} = \mathbf{1}(y_{t+l} < VaR_{\alpha,t+l})$ is the indicator function for the VaR exceedances. This loss function will be used for the *Model Confidence Set* (MCS) by Hansen et al. (2011) to select the best forecasting models among a set of competing models, given a specified confidence level. The MCS approach is an iterative procedure which consists of a test for the null hypothesis of equal predictive ability and an elimination rule: at each iteration, an equivalence test is performed to check if all models in the set of forecasting models $M = M_0$ have an equal forecasting accuracy by testing the following null hypothesis for a given confidence level $1-\beta$:

$$H_{0,M} = E(d_{ij,t}) = 0, \quad \forall i, j \in M, \quad vs \quad H_{A,M} = E(d_{ij,t}) \neq 0$$

where $d_{ij,t} = L_{i,t} - L_{j,t}$ is the sample loss differential between forecasting models i and j and $L_{i,t}$ stands for the loss function of model i at time t. If the null hypothesis cannot be rejected, then $\widehat{M}^*_{1-\beta} = M$. If the null hypothesis is rejected, the elimination rule e_M is used to remove the worst forecasting models from the set M. The procedure is repeated until the null hypothesis cannot be rejected, and the final set of models define the so-called model confidence set $\widehat{M}^*_{1-\beta}$. Hansen et al. (2011) propose different equivalence tests and we discuss here the T-max statistic which we used in the paper. First, the following t-statistics are computed:

$$t_{i\cdot} = \frac{\overline{d}_{i\cdot}}{\widehat{var}(\overline{d}_{i\cdot})}, \quad \text{for} \quad i \in M,$$

where $\overline{d}_{i\cdot} = m^{-1}\sum_{j\in M}\overline{d}_{ij}$ is the simple loss of the i-th model relative to the average losses across

models in the set M, and $\overline{d}_{ij} = T^{-1}\sum_{t=1}^{T} d_{ij,t}$ measures the sample loss differential between model i and j. Then, the T-max statistic is computed as $T_{max} = \max_{i\in M}(t_{i\cdot})$. The distribution of this test statistic is non-standard and is estimated using bootstrapping methods, see Hansen et al. (2011) for details. The elimination rule for the T-max statistic is $e_{max,M} = \arg\max_{i\in M}\left(\overline{d}_{i\cdot}/\widehat{var}(\overline{d}_{i\cdot})\right)$.

Expected Shortfall. The Expected Shortfall (ES) measures the average of the worst α losses, where α is a percentile of the P&L distribution. More formally, the expected shortfall at probability level $\alpha \in (0,1)$ is defined as

$$ES_\alpha = \frac{1}{\alpha}\int_0^\alpha F_z^{-1}(X)dz = \frac{1}{\alpha}\int_0^\alpha VaR_z(X)dz$$

where F^{-1} is the inverse function of the cdf of the P&L, which is the Value-at-Risk.

If the normal distribution is used, then the $\text{ES}_{\alpha,t+l}$ is given by,

$$ES_{\alpha,t+l} = \mu_{t+l} + \sigma_{t+l}\,\frac{\phi(\Phi_\alpha^{-1})}{\alpha}$$

where $\phi(\cdot)$ denotes the standard normal density function. Instead, if the Student's t distribution is employed, then the $\text{ES}_{\alpha,t+l}$ is given by

$$ES_{\alpha,t+l} = \mu_{t+l} + s_{t+l}\left(\frac{t_\nu(t_{\alpha,\nu}^{-1})}{\alpha}\left[\frac{\nu + (t_{\alpha,\nu}^{-1})^2}{\nu - 1}\right]\right)$$

where t_ν is the density function of the Student's t distribution with ν degrees of freedom. If copula models are used, similarly to the case of the VaR, simulation methods have to be implemented.

Backtesting methods for the Expected Shortfall. Backtesting the ES is an active field of research since the work of McNeil and Frey (2000), but it has become a very hot topic after Gneiting (2011) showed that ES lacks a mathematical property called elicitability (while VaR does have it), so that it may not be backtested[11]. However, Acerbi and Szekely (2014) showed that the ES can be back-tested and that the property of elicitability is useful only for model selection to choose the best model among a set of competitors. Moreover, Emmer et al. (2015) showed that the ES is elicitable conditionally on the VaR, and that it can be backtested through the approximation of several VaR levels: Kratz et al. (2018) used this latter approach for an empirical application. In this paper, we will use the tests by McNeil and Frey (2000) and by Kratz et al. (2018).

[11] In simple terms, a statistic $\psi(Y)$ of a random variable Y is *elicitable* if it minimizes the expected value of a scoring function S, $\psi(Y) = \arg\min_x E[S(x,Y)]$, where S can be, for the case of the Value-at-Risk, the asymmetric loss function of Gonzalez-Riviera et.al. (2004), while x are the model forecasts. Given a vector of forecasted VaR x_t and a vector of realized P&L y_t, the forecasting model can then be evaluated by minimizing the mean score $\bar{S} = \frac{1}{T}\sum_{t=1}^{T} S(x_t, y_t)$. Elicitability allows for the ranking of the risk models' performance because the scoring function can be used for comparative tests of the models.

McNeil and Frey (2000) proposed a one-sided test based on the *Exceedance Residuals (ER)* that exceed the VaR, that is $ER_{t+l} = (Y_{t+l} - \widehat{ES}_{\alpha,t+l})\mathbf{1}_{Y_{t+l} \leq \widehat{VaR}_{\alpha,t+l}}$, which should form a martingale difference sequence if $\widehat{ES}_{\alpha,t+l}$ and $\widehat{VaR}_{\alpha,t+l}$ are the true ES and VaR, respectively. More specifically, McNeil and Frey (2000) implement the standardized version of this test, where the exceedance residuals are standardized by the forecasted volatility (that is, $ER_{t+l}/\hat{\sigma}_{t+l}$). The null hypothesis is that the expected value of the standardized ER is zero against the alternative that the residuals have mean greater than zero, i.e. the expected shortfall is underestimated,

$$H_0 : \mu = 0, \quad vs \quad H_A : \mu > 0$$

where $\mu = E(ER_{t+l})$. The distribution of this test is computed using the bootstrap method by Efron and Tibshirani (1994) p. 224.

Kratz et al. (2018) proposed a *multilevel VaR backtest* to implicitly backtest the expected shortfall using the initial idea by Emmer et al. (2015), who showed that

$$ES_\alpha \approx \frac{1}{4}[q(\alpha) + q(0.75\alpha + 0.25) + q(0.5\alpha + 0.5) + q(0.25\alpha + 0.75)]$$

where $q(\gamma) = VaR_\gamma$. For example, if $\alpha = 0.05$ then

$$ES_{5\%} \approx \frac{1}{4}[VaR_{5\%} + VaR_{4\%} + VaR_{3\%} + VaR_{2\%}]$$

An estimated ES_α can be considered reliable if the estimates of the four VaR values obtained from the same model are reliable. In the general case, Kratz et al. (2018) consider VaR probability levels $\alpha_1, \ldots, \alpha_N$ defined by

$$\alpha_j = \alpha + \frac{j-1}{N}(1-\alpha), \quad j = 1, \ldots, N,$$

for some starting level α. Kratz et al. (2018) suppose to have a set of $VaR_{\alpha,t}$ and $ES_{\alpha,t}$ forecasts and a series of ex-post losses $\{L_t, t = 1, \ldots, T\}$. If $I_{t,j} = \mathbf{1}_{(L_t > VaR_{\alpha_j,t})}$ is the usual indicator function for a VaR violation at the level α_j and $X_t = \sum_{j=1}^{N} I_{t,j}$, then the sequence $(X_t)_{t=1,\ldots,T}$ counts the number of VaR levels that were breached. If we define $MN(T, (p_0, \ldots, p_N))$ as the multinomial distribution with T trials, each of which may result in one of $N+1$ outcomes $\{0, 1, \ldots, N\}$ according to probabilities $p_0, \ldots, p_N$ that sum to one, while the observed cell counts are defined by

$$O_j = \sum_{t=1}^{T} I_{(X_t = j)}, \quad j = 0, 1, \ldots, N$$

then, under the assumptions of unconditional coverage and independence as in Christoffersen (1998), the

random vector $(O_0, \ldots, O_N)$ should follow the multinomial distribution $(O_0, \ldots, O_N) \sim MN(T, (\alpha_1 - \alpha_0, \ldots, \alpha_{N+1} - \alpha_N))$. Assuming the multinomial distribution estimated from the data is $MN(T, (\theta_1 - \theta_0, \ldots, \theta_{N+1} - \theta_N))$, Kratz et al. (2018) test the following null and alternative hypotheses:

$$\begin{aligned} H_0 : &\quad \theta_j = \alpha_j, \quad \text{for} \quad j = 1, \ldots, N \\ H_1 : &\quad \theta_j \neq \alpha_j, \quad \text{for at least one } j \in \{1, \ldots, N\} \end{aligned}$$

Several test statistics have been proposed in the statistical literature to test the previous hypotheses: Cai and Krishnamoorthy (2006) performed a large simulations study to verify the exact size and power properties of five possible tests, while Kratz et al. (2018) employed three of these five tests. We will employ the *exact method* which is the fifth test statistic reviewed by Cai and Krishnamoorthy (2006), which computes the probability of a given outcome under the null hypothesis using the multinomial probability distribution itself:

$$P(O_0, O_1, \ldots, O_N) = \frac{T!}{O_0! O_1! \ldots O_N!} (\alpha_1 - \alpha_0)^{O_0} (\alpha_2 - \alpha_1)^{O_1} \ldots (\alpha_{N+1} - \alpha_N)^{O_N}$$

Cai and Krishnamoorthy (2006) found out that the exact method performs very well, but it can be time-consuming if the number of cells and the sample size T are large. In this latter case, simulation methods can be used to decrease the computational burden[12].

3.2 Credit Risk

In traditional finance, credit risk is the risk of a change in the value of positions associated with an unexpected deterioration in their credit quality, either downgrade or default. In the case of cryptocurrencies, as we previously discussed in section 3, credit risk is the risk of an extreme drop in the coin price following its "death", due to the discovery of a scam or a hacker attack or for other reasons. In this work, we will use the Probability of Default (PD) as the main measure of credit risk. However, when dealing with cryptocurrencies, this probability is usually referred to as the *probability of death of the coin i*, over a period of time $t+T$, given that it is alive at the time t:

$$PD_{i,t} = \mathcal{P}(D^i_{t+T} = 1 | D^i_t = 0)$$

Several approaches will be considered to estimate this probability using only available data, namely credit scoring models and the Zero Price Probability by Fantazzini, De Giuli, and Maggi (2008). We want to

[12]Several tests reviewed by Cai and Krishnamoorthy (2006) are implemented in the R package `XNomial` available at `cran.r-project.org/web/packages/XNomial` .

emphasize that structural models, like the well known Merton and KMV models, cannot be used because they require accounting data which are not available for cryptocurrencies. Similary to traditional credit risk, we will consider a 1-year time horizon.

Scoring models. Scoring models use statistical techniques to combine different factors into a quantitative score which is used to compute the probability of default, see Mester (1997), Fuertes and Kalotychou (2006), Rodriguez and Rodriguez (2006), Altman and Sabato (2007), Fantazzini and Figini (2008) and Fantazzini and Figini (2009). Credit scoring models can be applied even if ratings and accounting data are not available. Several scoring models typically take the following form:

$$PD_{i,t} = \mathcal{P}(D^i_{t+T} = 1 | D^i_t = 0; \mathbf{X}_{i,t}) = F(\beta' \mathbf{X}_{i,t})$$

In case of a logit model, $F(\beta' \mathbf{X}_{i,t})$ is given by the logistic cdf,

$$F(\beta' \mathbf{X}_{i,t}) = \frac{1}{1 + e^{-(\beta' \mathbf{X}_{i,t})}}$$

while for the probit model $F(\beta' \mathbf{X}_{i,t}) = \Phi(\beta' \mathbf{X}_{i,t})$ where $\Phi(\cdot)$ is a standard normal cdf.

We can also employ machine learning algorithms for solving classification tasks with two classes (dead coin/alive coin): in this work, we will use the *Random Forest* algorithm proposed by Ho (1995) and Breiman (2001). A Random Forest is an ensemble method using a large number of decision trees, where a decision tree is an instrument for decision making with the structure of a tree, i.e. it has branches and leaves. Each branch includes attributes which separate the alternative cases and following the tree through the branches, we end with a (pre-specified) value of an objective function. In case of a classification problem, each leave places an object either in one class or in the competing one. A single decision tree may provide a poor classification and can suffer from over-fitting and model instability: this is why the Random Forest algorithm is commonly used. Such an approach has multiple advantages: it is not sensitive to monotone transformations, it can cope with datasets with a large number of features and classes, and it can be used with continuous variables. See Friedman et al. (2016) and Smith and Koning (2017) for more details about decision trees and random forests.

The Zero Price Probability (ZPP). The Zero Price Probability was firstly introduced in Fantazzini et al. (2008) and relies on the fact that the PD can be estimated by computing the market-implied probability $\mathcal{P}(P_\tau \leq 0)$ with $t < \tau \leq t + T$. Because for a stock price (or a coin price) P_τ is a truncated variable and cannot become less than zero, the Zero-Price Probability is the probability that P_τ goes below the truncation level of zero. Fantazzini et al. (2008) discussed in details why the null price can be

used as a default barrier.

A general estimation procedure of the ZPP for multivariate time series is reported below:

1. Consider a generic conditional model for the differences of prices levels $\mathbf{X}_t = \mathbf{P}_t - \mathbf{P}_{t-1}$ without the log-transformation:

$$\mathbf{X}_t = \boldsymbol{\mu}_t + \mathbf{D}_t \mathbf{z}_t \tag{1}$$

   where $\boldsymbol{\mu}_t$ is a vector of conditional means, $\mathbf{D}_t = diag(\sigma_{1,t}, \ldots, \sigma_{n,t})$ a diagonal matrix of conditional standard deviations, while $\mathbf{z}_t$ is a vector of standardized errors with a conditional joint distribution given by H_t, which can be any of those discussed in Section 3.1 .

2. Simulate a high number N of price trajectories up to time $t+T$, using the estimated time series model (3) at step 1. We will compute the 1-year ahead probability of death for each coin using $T = 365$ days, given that crypto-assets are traded every day of the week (whereas $T = 250$ would be used in the case of traditional financial assets).

3. The probability of default/death for a crypto-coin i is simply the ratio n/N, where n is the number of times out of N when the simulated price $P^k_{\tau,i}$ touched or crossed the zero barrier along the simulated trajectory:

$$PD_{i,t} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}\{P^k_{\tau,i} \leq 0 \quad \text{for some } t < \tau \leq t+T\}$$

It follows easily from (1) that the models used in section 3.1 for measuring market risk can also be used for measuring credit risk.

A special case: ZPP with normally distributed price differences. The original method proposed in Fantazzini et al. (2008) dealt with general univariate models, where the ZPP can be estimated using only simulation methods. However, if we assume that the price differences are normally distributed and we use some recent theoretical results from barrier option theory, then it is possible to compute the ZPP using a closed-form solution.

Su and Rieger (2009) (Theorem 2.2) showed that if a stock price S_t follows a standard Brownian motion, then the ratio of the probability that the asset hits a barrier $H \in [0, \infty)$ before the maturity T and the probability that the asset is below the barrier at maturity T is equal to 2 :

$$P(\tau \leq T) = 2P(S_T \leq H) = 2\frac{1}{\sqrt{2\pi T}} \int_{-\infty}^{H} e^{-\frac{x^2}{2t}} dx$$

where τ is the first t when the underlying asset price touches the barrier H, that is $\min_{0\leq t\leq T}\{t|S_t \leq H\}$. If the asset follows a geometric Brownian motion with drift, they showed that the ratio of the previous two probabilities is not constant anymore, but for $H \to 0$ it still converges to 2 (see Su and Rieger (2009), Lemma 2.1):

$$\lim_{H\to 0} \frac{P(\tau \leq T)}{P(S_T \leq H)} = 2$$

Using these theoretical results and assuming that the coin price differences $X_t = P_t - P_{t-1}$ are normally distributed, we can find the closed-form solutions for the ZPP in two special cases:

- *Naive constant variance model (NCV).* This model was initially proposed by Li, Yang, and Zou (2016) and it is simply the random walk with drift:

  $$X_t = \mu + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma^2)$$

  If we use the previous results by Su and Rieger (2009) with $H = 0$ and a geometric Brownian motion with drift, it is straightforward to show that the 1-yead ahead ZPP can be computed at time t as follows:

  $$ZPP_{NCV} = P_{NCV}[P_\tau \leq 0] \approx 2\Phi\left(\frac{P_t - \mu T}{\sigma\sqrt{T}}\right), \quad \text{with } t < \tau \leq t + T \tag{2}$$

  where P_t is the last price. Note that this formula is an approximation which is valid only in the limit with $\Delta t \to 0$ and $H \to 0$.

- *GARCH(1,1) with normal errors.* If we assume that X_t follows a model with a constant mean and a GARCH(1,1) with normally distributed errors,

  $$\begin{aligned} X_t &= \mu + \varepsilon_t \\ \varepsilon_t &= \sigma_t^{1/2}\eta_t, \quad \eta_t \sim NID(0,1) \\ \sigma_t^2 &= \omega + \alpha\varepsilon_t^2 + \beta\sigma_{t-1}^2 \end{aligned}$$

  the ZPP can be approximated as follows

  $$ZPP_{GARCH11} = P_{GARCH11}[P_\tau \leq 0] \approx 2\Phi\left(\frac{P_t - \mu T}{\sigma_T}\right), \quad \text{with } t < \tau \leq t + T \tag{3}$$

  where σ_T^2 is sum of the forecasted variances of all shocks ε_t from $t+1$ till $t+T$, which can be shown to be[13]

[13] Compute the recursive forecasts of the conditional variance from time $t+1$ till time $t+T$; then collect all the common

$$\sigma_T^2 = \omega \cdot [A \times B] + \frac{[1-(\alpha+\beta)^T]}{[1-(\alpha+\beta)]} \cdot \hat{\sigma}^2_{t+1|t} \tag{4}$$

where $\hat{\sigma}^2_{t+1|t}$ can be either the forecasted variance at time $t+1$ conditional on information at time t, or the unconditional variance forecast[14], while A and B are two vectors defined below:

$$\underbrace{A}_{1\times(T-1)} = [(T-1) \quad (T-2) \ \dots \ 2 \quad 1], \qquad \underbrace{B}_{(T-1)\times 1} = \begin{bmatrix} (\alpha+\beta)^0 \\ (\alpha+\beta)^1 \\ \vdots \\ (\alpha+\beta)^{T-3} \\ (\alpha+\beta)^{T-2} \end{bmatrix}$$

There is an increasing empirical literature showing that the statistical distributions of crypto-assets are far from being normally distributed, see e.g. Chu et al. (2015), Osterrieder (2016), Chu et al. (2017), and Osterrieder and Lorenz (2017). Nevertheless, the previous closed-form formulas for the ZPP can provide a quick estimate of the probability of coin death and they can give an investor at least a rough idea of the crypto-asset credit risk. This is why they can be useful instruments for online data providers, who can publish the quotes of traded crypto-assets together with these market-implied PDs[15]. Note that the ZPP can highlight potential fraudulent behavior, like the famous pump-and-dump fraud that involves artificially inflating the price of a stock through various means, to later sell the purchased stock at a higher price to (un-informed) investors who lose all their money. In these cases, the ZPP tends to show very large swings of the estimated PD in short periods of time and any investor should consider such unusual behavior as a potential red flag. See Fantazzini et al. (2008) for some examples with the ZPPs of bankrupt companies which committed different forms of fraud.

Evaluation of prediction quality The main instrument used to check the forecasting performance of a model with binary data is the *confusion matrix* (Kohavi and Provost (1998)). The confusion matrix for a two class classifier is reported in Table 2. In the specific context of our analysis, the entries in the confusion matrix have the following meaning: a is the number of correct predictions that a coin is dead, b is the number of incorrect predictions that a coin is dead, c is the number of incorrect predictions that a coin is alive, while d is the number of correct predictions that a coin is alive.

The confusion matrix is then used to compute the *Area Under the Receiver Operating Characteristic* curve (AUC or AUROC) by Metz and Kronman (1980), Goin (1982) and Hanley and McNeil (1982)

components and use the property of the geometric series. The result will be given by (4).

[14]This second option is preferable in case of risk management.

[15]The ZPP is implemented in the R package `bitcoinFinance` available at `github.com/deanfantazzini/bitcoinFinance` .

$Observed/Predicted$	$DEAD\ COIN$	$ALIVE$
$DEAD\ COIN$	a	b
$ALIVE$	c	d

Table 2: Theoretical confusion matrix. Number of: a true positive, b false positive, c false negative, d true negative.

for all forecasting models: the receiver operating characteristic curve is obtained by plotting, for any probability threshold between 0 and 1, the proportion of correctly predicted dead coins $a/(a+b)$ on the y-axis (also known as the sensitivity or hit rate), and the proportion of alive coins predicted as dead coins $c/(c+d)$ (also known as the false positive rate). The AUC lies between zero and one and the closer it is to one the more accurate the forecasting model is, see Sammut and Webb (2011), pp. 869-875, and references therein for more details. We will also compute the model confidence set by Hansen et al. (2011) which was extended by Fantazzini and Maggi (2015) to binary models, to assess the prediction power of our competing models. We will employ the squared loss which is called the *Brier score* when applied on binary outcomes, see Brier (1950).

4 Empirical analysis

We combined a collection of over 1500 coins available from `coinmarketcap.com` with a list of 42 dead coins from `deadcoins.com`, which is a trusted source of information about scams, parody, and other junk coins. Merging both datasets and taking only the coins that are traded from at least the 1st January 2016, we obtained a dataset with 149 alive and 13 dead coins. It is worth noting that dead coins still have some negligible daily trading volumes for the reasons discussed in the Introduction.

The focus of this work is to estimate the market risk and the credit risk for small and medium-sized portfolios of crypto-currencies, so that we considered the following three cases:

- estimation of the market risk measures for a portfolio of the top-5 cryptocurrencies by market capitalization;
- estimation of the market risk measures for a portfolio of 15 coins, consisting of the top-5 coins, 5 random average coins, and 5 junk cryptocurrencies;
- estimation of the credit risk measures for 42 coins, consisting of 29 alive coins and 13 dead coins[16].

[16] We used a sample of 42 coins for credit risk and not a larger sample for two reasons: 1) considering that we used multivariate models for the simultaneous estimation of both market and credit risk, the models discussed in this paper would be unable to estimate the market risk of a portfolio of hundreds (or thousands) of coins. In this case, a completely different set of multivariate models need to be used, the DECO model by Engle and Kelly (2012) being (potentially) one of them. 2) The lack of historical data for the vast majority of dead coins: this problem would make any model suffer from massive selection bias. For these two reasons, the case of large portfolios will be considered in a separate paper.

4.1 Portfolio with top-5 coins

We considered the top-5 coins by market capitalization at the time we started writing this paper in 2018, with daily data ranging between June 2016 and May 2018: BTC (Bitcoin), ETH (Ethereum), XRP (Ripple), LTC (Litecoin) and XLM (Stellar). This period of time witnessed the fast rise of crypto-currencies (with bitcoin reaching almost $20000), followed by a violent price fall. The descriptive statistics of the daily profits and losses for each coin are presented in Table 3. All coins generally exhibited large skewness and kurtosis, very high coefficients of variations, and unconditional distributions very far from the Gaussian.

	Mean	*Std.Dev.*	*Min*	*Max*	*Skewness*	*Kurtosis*	*Coeff.of Var.*
BTC	9.60	381.65	-2329.30	3608.20	0.97	20.16	39.74
ETH	0.79	27.94	-238.23	151.21	-0.58	15.20	35.27
XRP	0.00	0.07	-0.92	0.78	-0.26	76.95	83.89
XLM	0.00	0.02	-0.17	0.33	4.07	87.40	54.78
LTC	0.16	8.42	-52.53	101.96	3.10	42.05	53.05

Table 3: Descriptive statistics for the daily P&L for each coin (June 2016 - May 2018).

Following Giacomini and Komunjer (2005), Gonzalez-Rivera et al. (2004) and Fantazzini (2009a), we used a rolling forecasting scheme of 522 observations and we left 200 data for the out-of-sample evaluation. We used the first differences of the data without log transformation (that is the P&L) to compute not only their market risk measures, but also their credit risk measures using the Zero Price Probability (ZPP) by Fantazzini et al. (2008). More specifically, we used a VAR(1) model for the conditional means, univariate GARCH(1,1) models for the conditional variances, and the following six multivariate distributions for the standardized errors:

- a DCC(1,1) model with multivariate normal distribution;
- a DCC(1,1) model with multivariate Student's t distribution;
- a normal copula/skewed-t GARCH model with constant correlation matrix;
- a t-copula/skewed-t GARCH model with constant correlation matrix;
- a normal copula/skewed-t GARCH model with a correlation matrix following a DCC(1,1) model;
- a t-copula/skewed-t GARCH model with a correlation matrix following a DCC(1,1) model;

We calculated the 1-day-ahead forecasts for the 1%, 2%, 3%, 4% and 5% VaR levels and for the 5% ES, for each coin and for an equally-weighted portfolio with the five coins. The actual VaR exceedances T_1/T , the p-values of the Kupiec's unconditional coverage test and the p-values of the Christoffersen's conditional coverage test for the VaR forecasts at all quantile levels are reported in Tables 4-6. The asymmetric VaR losses and the forecasting models included in the MCS are reported in Table 7.

	DCC MVN	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*
VaR exceedances (T_1/T): 1\% quantile						
BTC	3.52%	1.51%	1.51%	1.01%	1.51%	1.51%
ETH	3.02%	3.52%	3.02%	1.51%	2.51%	3.02%
XRP	1.51%	1.51%	0.50%	0.00%	0.00%	0.00%
XLM	2.01%	1.51%	1.01%	0.50%	1.01%	1.01%
LTC	0.50%	0.00%	0.00%	0.00%	0.00%	0.00%
portfolio	4.02%	2.01%	5.03%	2.51%	2.51%	2.01%
VaR exceedances (T_1/T): 2\% quantile						
BTC	4.02%	4.52%	3.02%	2.51%	4.02%	4.02%
ETH	4.52%	4.02%	4.02%	3.52%	4.02%	4.02%
XRP	3.52%	4.52%	1.01%	1.01%	1.01%	1.01%
XLM	3.52%	4.02%	1.01%	1.01%	1.01%	1.01%
LTC	1.51%	2.51%	0.50%	0.50%	0.50%	0.50%
portfolio	7.54%	6.53%	6.03%	5.53%	6.03%	6.03%
VaR exceedances (T_1/T): 3\% quantile						
BTC	5.53%	6.03%	4.52%	4.02%	4.52%	4.52%
ETH	5.53%	6.03%	6.03%	4.52%	6.03%	6.03%
XRP	4.02%	6.03%	1.51%	2.01%	2.01%	2.01%
XLM	3.52%	6.03%	2.51%	1.01%	3.02%	3.02%
LTC	3.02%	5.03%	1.01%	1.01%	1.51%	1.01%
portfolio	8.54%	8.54%	9.55%	10.05%	8.54%	8.54%
VaR exceedances (T_1/T): 4\% quantile						
BTC	6.03%	6.53%	5.03%	5.03%	6.03%	5.53%
ETH	7.04%	7.54%	8.04%	7.54%	8.54%	8.54%
XRP	4.02%	8.54%	2.51%	2.51%	2.51%	2.51%
XLM	3.52%	6.03%	3.02%	3.02%	3.02%	3.02%
LTC	4.52%	10.05%	2.01%	1.51%	2.01%	2.01%
portfolio	9.55%	11.06%	10.55%	10.05%	10.05%	9.55%
VaR exceedances (T_1/T): 5\% quantile						
BTC	6.53%	7.54%	6.53%	6.03%	6.53%	6.53%
ETH	8.04%	8.04%	8.54%	8.54%	9.05%	9.05%
XRP	5.03%	10.55%	3.02%	4.02%	3.52%	3.52%
XLM	4.52%	7.54%	4.52%	3.52%	4.02%	4.02%
LTC	5.53%	12.56%	2.51%	2.01%	3.02%	2.51%
portfolio	11.06%	12.56%	12.06%	11.06%	10.55%	11.06%

Table 4: VaR exceedances T_1/T for each coin and for the equally-weighted portfolio.

	DCC MVN	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*
Kupiec's test p-value: VaR 1\%						
BTC	**0.01**	0.50	0.50	0.99	0.50	0.50
ETH	**0.02**	**0.01**	**0.02**	0.50	0.07	**0.02**
XRP	0.50	0.50	0.44	0.05	0.05	0.05
XLM	0.21	0.50	0.99	0.44	0.99	0.99
LTC	0.44	0.05	0.05	0.05	0.05	0.05
portfolio	**0.00**	0.21	**0.00**	0.07	0.07	0.21
Kupiec's test p-value: VaR 2\%						
BTC	0.07	**0.03**	0.34	0.62	0.07	0.07
ETH	**0.03**	0.07	0.07	0.17	0.07	0.07
XRP	0.17	**0.03**	0.27	0.27	0.27	0.27
XLM	0.17	**0.07**	0.27	0.27	0.27	0.27
LTC	0.60	0.62	0.07	0.07	0.07	0.07
portfolio	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Kupiec's test p-value: VaR 3\%						
BTC	0.06	**0.03**	0.24	0.42	0.24	0.24
ETH	0.06	**0.03**	**0.03**	0.24	**0.03**	**0.03**
XRP	0.42	**0.03**	0.17	0.38	0.38	0.38
XLM	0.68	**0.03**	0.68	0.06	0.99	0.99
LTC	0.99	0.13	0.06	0.06	0.17	0.06
portfolio	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Kupiec's test p-value: VaR 4\%						
BTC	0.17	0.09	0.48	0.48	0.17	0.30
ETH	0.05	**0.02**	**0.01**	**0.02**	**0.00**	**0.00**
XRP	0.99	**0.00**	0.25	0.25	0.25	0.25
XLM	0.72	0.17	0.46	0.46	0.46	0.46
LTC	0.71	**0.00**	0.11	0.11	0.11	0.11
portfolio	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Kupiec's test p-value: VaR 5\%						
BTC	0.34	0.13	0.34	0.52	0.34	0.34
ETH	0.07	0.07	0.04	**0.04**	**0.02**	**0.02**
XRP	0.99	**0.00**	0.17	0.51	0.31	0.31
XLM	0.75	0.13	0.75	0.31	0.51	0.51
LTC	0.74	**0.00**	0.08	**0.03**	0.17	0.08
portfolio	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Table 5: Kupiec's tests for each coin and for the equally-weighted portfolio.

	DCC MVN	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*
Christoffersen's test p-value: VaR 1\%						
BTC	0.02	0.76	0.76	0.98	0.76	0.76
ETH	0.03	**0.01**	**0.03**	0.76	0.05	**0.03**
XRP	0.76	0.76	0.73	0.14	0.14	0.14
XLM	0.42	0.76	0.98	0.73	0.98	0.98
LTC	0.73	0.14	0.14	0.14	0.14	0.14
portfolio	**0.00**	0.42	**0.00**	0.17	0.17	0.42
Christoffersen's test p-value: VaR 2\%						
BTC	0.14	0.06	0.53	0.78	0.14	0.14
ETH	0.01	0.12	0.12	0.18	0.12	0.12
XRP	0.30	0.06	0.53	0.53	0.53	0.53
XLM	0.30	0.14	0.53	0.53	0.53	0.53
LTC	0.83	0.78	0.20	0.20	0.20	0.20
portfolio	0.00	0.00	0.00	0.01	0.00	0.00
Christoffersen's test p-value: VaR 3\%						
BTC	0.05	**0.04**	0.33	0.52	0.33	0.33
ETH	0.05	**0.04**	0.08	0.36	0.08	0.08
XRP	0.52	**0.04**	0.38	0.63	0.63	0.63
XLM	0.71	0.08	0.81	0.16	0.83	0.83
LTC	0.83	0.18	0.16	0.16	0.38	0.16
portfolio	0.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Christoffersen's test p-value: VaR 4\%						
BTC	0.16	0.13	0.63	0.63	0.16	0.52
ETH	0.09	0.05	**0.03**	0.07	0.02	0.02
XRP	0.71	**0.01**	0.46	0.46	0.46	0.46
XLM	0.73	0.37	0.63	0.63	0.63	0.63
LTC	0.61	**0.00**	0.26	0.26	0.26	0.26
portfolio	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Christoffersen's test p-value: VaR 5\%						
BTC	0.33	0.22	0.33	0.33	0.33	0.33
ETH	0.16	0.16	0.10	0.10	0.06	0.06
XRP	0.59	**0.00**	0.32	0.58	0.46	0.46
XLM	0.62	0.31	0.62	0.46	0.58	0.58
LTC	0.49	**0.00**	0.18	0.08	0.32	0.18
portfolio	**0.00**	**0.00**	**0.00**	**0.00**	**0.01**	**0.00**

Table 6: Christoffersen's tests for each coin and for the equally-weighted portfolio.

	DCC MVN	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*
Asymmetric VaR Loss and inclusion in the MCS: 1\% quantile						
BTC	0.004	0.004	0.005	0.005	0.004	0.005
ETH	0.005	0.004	0.005	0.005	0.005	0.005
XRP	0.008	0.008	0.007	0.007	0.007	0.007
XLM	0.005	0.007	0.007		0.007	0.007
LTC	0.004	0.004				
portfolio	0.003	0.003	0.003	0.003	0.003	0.003
Asymmetric VaR Loss and inclusion in the MCS: 2\% quantile						
BTC	0.008	0.008	0.008	0.008		0.008
ETH	0.009	0.009	0.009	0.009	0.009	0.009
XRP	0.012	0.014	0.013	0.013	0.013	0.013
XLM	0.010	0.012	0.012	0.012		0.012
LTC	0.008	0.008				
portfolio	0.005	0.005	0.005	0.005	0.005	0.004
Asymmetric VaR Loss and inclusion in the MCS: 3\% quantile						
BTC	0.011	0.012	0.012	0.012	0.012	0.012
ETH	0.012	0.012	0.012	0.012	0.012	0.012
XRP	0.016	0.018	0.017	0.017	0.017	0.017
XLM	0.014	0.016	0.016	0.016	0.016	0.016
LTC	0.011	0.011				
portfolio	0.006	0.006	0.006	0.006	0.006	0.006
Asymmetric VaR Loss and inclusion in the MCS: 4\% quantile						
BTC	0.014	0.014	0.015	0.015	0.015	0.015
ETH	0.015	0.015	0.015	0.015	0.015	0.015
XRP	0.020	0.021	0.021	0.021	0.021	0.021
XLM	0.018	0.019	0.020			
LTC	0.014	0.015				
portfolio	0.007	0.008	0.008	0.008	0.008	0.008
Asymmetric VaR Loss and inclusion in the MCS: 5\% quantile						
BTC	0.017	0.017	0.017	0.017	0.017	0.017
ETH	0.017	0.018	0.018	0.018	0.018	0.018
XRP	0.023	0.024	0.024	0.024	0.024	0.024
XLM	0.021	0.022				
LTC	0.016	0.019				
portfolio	0.009	0.009	0.009	0.009	0.009	0.009

Table 7: Asymmetric VaR Loss and MCS: an empty cell means the model is not included into the MCS.

Table 4-7 show that the t-copula with skewed-t GARCH marginals can be a good compromise for precise VaR estimates across different quantile levels, particularly for the most extreme quantiles (1% and 2%) which are the most important for regulatory purposes, see BIS (2013) and BIS (2016). This evidence is consistent with other results reported in the financial literature using copulas, see Cherubini et al. (2004), Fantazzini (2008), Fantazzini (2009a), Patton (2009), Weiss (2011), Weiss (2013), Patton (2013) and McNeil et al. (2015). The worse VaR results for central quantiles (4% and 5%) were expected, due to the computational problems associated with the estimation of GARCH models with small samples (≤ 500 observations), see Fantazzini (2009a) for more details. Unfortunately, increasing the estimation window is not advisable when using crypto-assets, due to the potential presence of structural breaks caused by changes in local regulations, the arrival of new investors, hacking attacks and massive improvements in mining hardware, see e.g. Bouri et al. (2016), Fantazzini et al. (2016), Fantazzini et al. (2017) and Mensi et al. (2018). This is why we employed a rolling estimation window of 522 observations, which is similar to the window size suggested by Hwang and Valls Pereira (2006), who investigated the small sample properties of the maximum likelihood estimates of ARCH and GARCH models. The asymmetric VaR losses of the competing models are rather close and almost all models are included into the MCS (for both the single coins and the portfolio case). This latter result was expected because Fantazzini (2009a) showed that, when small samples are considered and the data are skewed and leptokurtic, the biases in the GARCH parameters are so large that they deliver conservative VaR estimates, even with a simple multivariate normal distribution.

The p-values of the exceedance residuals test by McNeil and Frey (2000) and of the multilevel VaR backtest with $N = 4$ levels by Kratz et al. (2018) for the forecasted ES 5% are reported in Table 8.

	DCC MVN	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*
ES 5% test 1: Exceedance Residuals test by McNeil and Frey (2000)						
BTC	1.00	1.00	0.78	0.75	0.66	0.67
ETH	0.85	1.00	0.49	0.70	0.52	0.52
XRP	0.28	1.00	0.66	0.91	0.83	0.81
XLM	**0.00**	0.99	0.87	0.93	0.86	0.85
LTC	0.96	1.00	0.99	0.98	1.00	0.99
portfolio	0.25	1.00	0.12	0.21	0.15	0.44
ES 5% test 2: multilevel VaR backtest by Kratz et al. (2018)						
BTC	0.09	**0.01**	0.51	0.81	0.17	0.22
ETH	**0.01**	**0.01**	**0.00**	**0.02**	**0.00**	**0.00**
XRP	0.68	**0.00**	0.22	0.51	0.42	0.42
XLM	0.70	**0.03**	0.76	0.21	0.73	0.73
LTC	0.97	**0.00**	**0.02**	**0.00**	0.08	**0.02**
portfolio	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Table 8: ER test by McNeil and Frey (2000) and multilevel VaR backtest by Kratz et al. (2018).

The ER test by McNeil and Frey (2000) does not reject the null hypothesis for almost any model, except for a normal DCC model. Instead, the multilevel VaR backtest by Kratz et al. (2018) is much more selective, and it rejects the null hypothesis of a correctly specified multinomial distribution for both

DCC models and (to a lesser extent) copula models. These results confirm the evidence reported in table 4 that DCC models work poorly with extreme quantiles, while copula models have problems with central quantiles.

4.2 Portfolio with 15 coins

We considered the following 15 coins with daily data ranging between June 2016 and May 2018: BTC (Bitcoin), ETH (Ethereum), XRP (Ripple), LTC (Litecoin), XLM (Stellar), XCP (Counterparty), SHIFT (Shift), RVR (RevolutionVR), THC (HempCoin), CLAM (Clams), GAME (GameCredits), MINT (MintCoin), ABY (ArtByte), GLD (GoldCoin), EFL (e-Gulden). The actual VaR exceedances T_1/T and the p-values of the Kupiec's unconditional coverage test are reported in Table 9, while the p-values of the Christoffersen's conditional coverage test, the asymmetric VaR losses and the forecasting models included in the MCS are reported in Table 10. The p-values of the exceedance residuals test by McNeil and Frey (2000) and of the multilevel VaR backtest with $N = 4$ levels by Kratz et al. (2018) for the forecasted ES 5% are reported in Table 11.

The results with 15 coins are rather similar to the previous analysis with 5 coins: copula-GARCH models pass the vast majority of Kupiec and Christoffersen tests (whereas DCC models do not) and are usually more precise for extreme quantiles, particularly the t-copula/skewed-t GARCH model. However, the asymmetric VaR losses are not very different and in several cases, all models are included in the MCS. The ER test by McNeil and Frey (2000) rejects the null hypothesis for very few models, while the multilevel VaR backtest by Kratz et al. (2018) again rejects the null hypothesis much more for DCC models than for copula models. In general, DCC models seem to provide worse model fits of the multivariate distribution tails compared to copula models, thus confirming recent simulation studies by Muller and Righi (2018).

4.3 Credit risk for 42 coins

We computed the probability of death/default for a set of 42 coins reported in Table 12 , using the methods described in section 3.2.2 .

We split our dataset into a training set containing all coins up to May 2017, and a validation (out-of-time) set ranging from June 2017 till May 2018. Note that the coins which were considered dead in May 2017, were also included in the validation set because they kept on trading and they could be potentially revived, see the previous discussion in Section 3. We estimated all models using the training subset and then we computed the 1-year-ahead forecasts for the probability of death for each coin: these forecasts were used with the validation dataset to compute the AUC scores, Brier scores and the MCS.

	DCC MVN	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*	*DCC MVN*	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*
	VaR exceedances (T_1/T): 1\% quantile						**Kupiec's test p-value: VaR 1\%**					
BTC	4.02%	1.01%	2.01%	0.50%	1.01%	1.01%	**0.00**	0.99	0.21	0.44	0.99	0.99
ETH	4.02%	2.51%	3.02%	1.01%	1.51%	1.51%	**0.00**	0.07	**0.02**	0.99	0.50	0.50
XRP	4.52%	2.01%	1.51%	0.50%	0.50%	0.50%	**0.00**	0.21	0.50	0.44	0.44	0.44
XLM	5.03%	3.02%	3.52%	1.51%	2.51%	2.51%	**0.00**	**0.02**	**0.01**	0.50	0.07	0.07
LTC	2.51%	0.50%	1.01%	0.00%	0.00%	0.00%	0.07	0.44	0.99	0.05	0.05	0.05
XCP	4.02%	3.02%	3.02%	3.52%	3.52%	3.52%	**0.00**	**0.02**	**0.02**	0.01	**0.01**	**0.01**
SHIFT	5.03%	3.02%	2.51%	1.51%	1.51%	1.51%	**0.00**	**0.02**	0.07	0.50	0.50	0.50
RVR	4.52%	4.02%	2.51%	1.51%	2.01%	2.01%	**0.00**	**0.00**	0.07	0.50	0.21	0.21
THC	3.02%	3.02%	1.01%	0.50%	0.50%	0.50%	**0.02**	**0.02**	0.99	0.44	0.44	0.44
CLAM	3.02%	1.01%	1.51%	0.50%	0.50%	0.50%	**0.02**	0.99	0.50	0.44	0.44	0.44
GAME	2.01%	3.02%	2.01%	1.01%	1.01%	1.01%	0.07	**0.02**	0.21	0.99	0.99	0.99
MINT	4.02%	1.51%	1.51%	1.01%	1.01%	1.01%	**0.00**	0.50	0.50	0.99	0.99	0.99
ABY	4.52%	3.02%	1.51%	1.51%	1.51%	1.51%	**0.00**	**0.02**	0.50	0.50	0.50	0.50
GLD	2.01%	1.51%	1.51%	0.50%	0.50%	0.50%	0.21	0.50	0.50	0.44	0.44	0.44
EFL	2.51%	2.01%	1.01%	0.00%	0.00%	0.00%	0.07	0.21	0.99	0.05	0.05	0.05
portfolio	7.04%	6.03%	7.04%	4.52%	5.03%	5.03%	**0.00**	**0.00**	0.00	**0.00**	**0.00**	**0.00**
	VaR exceedances (T_1/T): 2\% quantile						**Kupiec's test p-value: VaR 2\%**					
BTC	7.04%	4.02%	3.02%	1.01%	1.51%	1.51%	**0.00**	0.07	0.34	0.27	0.60	0.60
ETH	5.53%	3.52%	4.52%	1.51%	4.02%	4.02%	**0.00**	0.17	**0.03**	0.60	0.07	0.07
XRP	5.03%	4.52%	3.52%	1.51%	2.01%	2.01%	**0.01**	**0.03**	0.17	0.60	0.99	0.99
XLM	6.53%	6.03%	4.02%	4.02%	4.02%	4.02%	**0.00**	**0.00**	0.07	0.07	0.07	0.07
LTC	3.52%	4.02%	1.01%	0.50%	0.50%	0.50%	0.17	0.07	0.27	0.07	0.07	0.07
XCP	5.53%	3.52%	4.02%	6.03%	6.03%	6.03%	**0.00**	0.17	0.07	**0.00**	**0.00**	**0.00**
SHIFT	5.03%	4.02%	4.52%	3.52%	3.52%	3.52%	**0.01**	0.07	**0.03**	0.17	0.17	0.17
RVR	5.03%	5.03%	4.02%	3.02%	3.02%	3.02%	**0.01**	**0.01**	0.07	0.34	0.34	0.34
THC	3.52%	3.02%	2.51%	0.50%	1.51%	1.51%	0.17	0.34	0.62	0.07	0.60	0.60
CLAM	3.02%	1.51%	2.01%	1.01%	1.01%	1.01%	0.34	0.60	0.99	0.27	0.27	0.27
GAME	4.52%	4.52%	3.02%	2.51%	3.02%	3.02%	**0.01**	**0.03**	0.34	0.62	0.34	0.34
MINT	5.53%	4.52%	3.02%	2.51%	2.51%	2.51%	**0.00**	**0.03**	0.34	0.62	0.62	0.62
ABY	5.53%	5.53%	4.02%	4.52%	4.02%	4.52%	**0.00**	**0.00**	0.07	**0.03**	0.07	**0.03**
GLD	2.51%	3.02%	2.01%	0.50%	0.50%	0.50%	0.62	0.34	0.99	0.07	0.07	0.07
EFL	2.51%	4.52%	1.51%	0.00%	0.00%	0.00%	0.62	**0.03**	0.60	**0.00**	**0.00**	**0.00**
portfolio	8.04%	8.04%	8.04%	7.04%	7.04%	7.04%	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	VaR exceedances (T_1/T): 3\% quantile						**Kupiec's test p-value: VaR 3\%**					
BTC	8.04%	5.03%	5.03%	3.02%	5.03%	5.03%	**0.00**	0.13	0.13	0.99	0.13	0.13
ETH	6.03%	5.53%	5.03%	4.02%	4.52%	4.52%	0.03	0.06	0.13	0.42	0.24	0.24
XRP	5.03%	6.03%	6.03%	3.02%	3.02%	3.02%	0.13	**0.03**	**0.03**	0.99	0.99	0.99
XLM	7.54%	6.53%	5.03%	4.02%	4.52%	4.52%	**0.00**	**0.01**	0.13	0.42	0.24	0.24
LTC	5.03%	7.04%	2.01%	1.01%	1.51%	2.01%	0.13	**0.00**	0.38	0.06	0.17	0.38
XCP	6.53%	5.53%	6.53%	6.53%	7.04%	7.04%	**0.01**	0.06	**0.01**	**0.01**	**0.00**	**0.00**
SHIFT	6.03%	6.53%	5.53%	4.52%	4.52%	4.52%	**0.03**	**0.01**	0.06	0.24	0.24	0.24
RVR	5.53%	6.03%	5.03%	5.03%	5.03%	5.03%	0.06	**0.03**	0.13	0.13	0.13	0.13
THC	3.52%	6.03%	3.02%	2.01%	2.01%	2.01%	0.68	**0.03**	0.99	0.38	0.38	0.38
CLAM	4.02%	2.51%	2.01%	1.51%	1.51%	1.51%	0.42	0.68	0.38	0.17	0.17	0.17
GAME	5.53%	5.53%	4.02%	4.52%	5.03%	5.03%	**0.03**	0.06	0.42	0.24	0.13	0.13
MINT	5.53%	6.03%	4.02%	4.52%	4.52%	4.52%	0.06	**0.03**	0.42	0.24	0.24	0.24
ABY	6.53%	6.53%	8.04%	6.03%	6.03%	6.03%	**0.01**	**0.01**	**0.00**	**0.03**	**0.03**	**0.03**
GLD	3.02%	3.52%	2.51%	2.01%	1.51%	1.51%	0.99	0.68	0.68	0.38	0.17	0.17
EFL	2.51%	5.53%	2.01%	0.50%	0.50%	0.50%	0.68	0.06	0.38	**0.01**	**0.01**	**0.01**
portfolio	9.55%	10.05%	10.55%	7.04%	7.54%	7.54%	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	VaR exceedances (T_1/T): 4\% quantile						**Kupiec's test p-value: VaR 4\%**					
BTC	9.05%	7.04%	6.03%	6.03%	7.04%	7.04%	**0.00**	0.05	0.17	0.17	0.05	0.05
ETH	7.04%	6.53%	6.03%	4.52%	5.03%	5.53%	0.05	0.09	0.17	0.71	0.48	0.30
XRP	5.53%	8.54%	6.53%	3.52%	4.02%	4.02%	0.30	**0.00**	0.09	0.72	0.99	0.99
XLM	7.54%	7.54%	5.53%	4.52%	4.52%	5.03%	**0.02**	**0.02**	0.30	0.71	0.71	0.48
LTC	5.03%	8.04%	2.01%	2.01%	2.51%	3.02%	0.48	**0.01**	0.11	0.11	0.25	0.46
XCP	6.53%	6.03%	8.04%	7.54%	7.54%	8.04%	0.09	0.17	**0.01**	**0.02**	**0.02**	**0.01**
SHIFT	6.53%	7.04%	6.53%	5.53%	5.53%	5.53%	0.09	0.05	0.09	0.30	0.30	0.30
RVR	6.53%	8.54%	5.53%	5.03%	5.53%	5.53%	0.09	**0.00**	0.30	0.48	0.30	0.30
THC	3.52%	7.04%	3.02%	2.51%	2.51%	2.51%	0.72	0.05	0.46	0.25	0.25	0.25
CLAM	5.53%	5.03%	3.02%	2.01%	2.01%	2.51%	0.30	0.48	0.46	0.11	0.11	0.25
GAME	6.03%	7.04%	5.53%	5.53%	6.03%	6.03%	0.09	0.05	0.30	0.30	0.17	0.17
MINT	7.04%	6.53%	4.52%	4.52%	4.52%	4.52%	0.05	0.09	0.71	0.71	0.71	0.71
ABY	7.04%	10.05%	8.04%	7.54%	7.54%	7.54%	0.05	**0.00**	0.01	0.02	**0.02**	**0.02**
GLD	4.02%	6.53%	3.52%	3.02%	2.51%	3.02%	0.99	0.09	0.72	0.46	0.25	0.46
EFL	3.02%	6.53%	2.51%	1.01%	1.01%	1.51%	0.46	0.09	0.25	**0.01**	**0.01**	0.04
portfolio	11.06%	12.56%	10.55%	9.55%	9.55%	9.55%	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	VaR exceedances (T_1/T): 5\% quantile						**Kupiec's test p-value: VaR 5\%**					
BTC	9.05%	9.05%	7.04%	7.04%	7.54%	7.54%	**0.02**	0.02	0.21	0.21	0.13	0.13
ETH	8.04%	7.54%	7.04%	4.52%	6.53%	6.53%	0.07	0.13	0.21	0.75	0.34	0.34
XRP	8.04%	10.05%	8.04%	4.02%	4.02%	4.02%	0.07	**0.00**	0.07	0.51	0.51	0.51
XLM	7.54%	8.04%	7.54%	5.03%	5.53%	5.53%	0.13	0.07	0.13	0.99	0.74	0.74
LTC	6.53%	10.55%	4.02%	2.51%	4.02%	4.52%	0.34	**0.00**	0.51	0.08	0.51	0.75
XCP	7.54%	8.04%	8.54%	8.54%	8.54%	8.54%	0.13	0.07	0.04	0.04	**0.04**	0.04
SHIFT	7.04%	9.05%	7.54%	6.03%	6.03%	6.03%	0.21	**0.02**	0.13	0.52	0.52	0.52
RVR	6.53%	9.55%	7.04%	5.53%	6.03%	6.03%	0.34	**0.01**	0.21	0.74	0.52	0.52
THC	4.52%	8.54%	3.52%	2.51%	2.51%	2.51%	0.75	**0.04**	0.31	0.08	0.08	0.08
CLAM	6.03%	7.04%	3.52%	3.02%	2.51%	3.02%	0.52	0.21	0.31	0.17	0.08	0.17
GAME	6.03%	7.54%	6.03%	6.03%	6.03%	6.03%	0.34	0.13	0.52	0.52	0.52	0.52
MINT	7.54%	8.04%	4.52%	5.03%	5.03%	5.03%	0.13	0.07	0.75	0.99	0.99	0.99
ABY	7.04%	11.06%	9.55%	7.54%	7.54%	7.54%	0.21	**0.00**	**0.01**	0.13	0.13	0.13
GLD	4.02%	8.54%	5.53%	4.52%	4.02%	4.02%	0.51	**0.04**	0.74	0.75	0.51	0.51
EFL	3.52%	8.54%	3.02%	1.51%	1.51%	1.51%	0.31	**0.04**	0.17	**0.01**	**0.01**	**0.01**
portfolio	12.06%	13.57%	12.06%	10.05%	10.05%	10.05%	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Table 9: VaR exceedances T_1/T (left) and Kupiec's tests (right) for each coin and for the equally-weighted portfolio. P-values smaller than 0.05 are in bold font.

	DCC MVN	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*	*DCC MVN*	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*
	Christoffersen's test p-value: VaR 1\%						**Asymmetric VaR 1% Loss and MCS**					
BTC	**0.00**	0.98	0.42	0.73	0.98	0.98	0.010	0.006	0.008	0.007	0.006	0.006
ETH	**0.00**	0.17	0.06	0.98	0.76	0.76	0.010	0.006	0.007	0.007	0.007	0.007
XRP	**0.00**	0.42	0.76	0.73	0.73	0.73	0.006	0.006	0.007	0.006	0.006	0.006
XLM	**0.00**	0.06	**0.02**	0.76	0.17	0.17	0.009	0.008	0.009	0.007	0.007	0.007
LTC	0.17	0.73	0.98	0.14	0.14	0.14	0.010	0.005	0.008	0.007	0.007	0.006
XCP	**0.00**	0.06	0.06	**0.02**	**0.02**	**0.02**	0.013	0.013	0.010	0.014	0.014	0.014
SHIFT	**0.00**	0.06	0.17	0.76	0.76	0.76	0.007	0.007	0.007		0.007	0.007
RVR	**0.00**	0.00	0.17	0.76	0.42	0.42	0.015	0.010	0.010	0.009	0.009	0.009
THC	0.06	0.06	0.98	0.73	0.73	0.73	0.008	0.009	0.008	0.009		0.009
CLAM	0.06	0.98	0.76	0.73	0.73	0.73	0.015	0.014		0.014	0.014	0.014
GAME	0.17	0.06	0.42	0.98	0.98	0.98	0.012	0.008	0.009	0.008	0.008	0.008
MINT	**0.00**	0.76	0.76	0.98	0.98	0.98	0.014	0.014	0.015	0.015	0.015	0.015
ABY	**0.00**	0.06	0.76	0.76	0.76	0.76	0.012	0.007	0.009	0.008	0.008	0.008
GLD	0.42	0.76	0.76	0.73	0.73	0.73	0.015	0.014		0.014	0.014	0.014
EFL	0.17	0.42	0.98	0.14	0.14	0.14	0.016	0.009	0.014	0.015	0.014	0.014
portfolio	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**		0.005		0.005	0.005	0.005
	Christoffersen's test p-value: VaR 2\%						**Asymmetric VaR 2% Loss and MCS**					
BTC	**0.00**	0.14	0.53	0.53	0.83	0.83	0.015	0.012	0.013	0.011	0.011	0.011
ETH	**0.01**	0.18	0.06	0.83	0.14	0.14	0.015	0.010	0.012	0.012	0.012	0.012
XRP	**0.02**	0.06	0.30	0.83	0.92	0.92	0.012	0.012	0.012	0.010	0.011	0.010
XLM	**0.00**	**0.00**	0.14	0.14	0.14	0.14	0.013	0.013		0.012	0.012	0.012
LTC	0.30	0.14	0.53	0.20	0.20	0.20	0.014	0.009	0.011	0.011	0.011	0.010
XCP	**0.01**	0.30	0.12	**0.00**	**0.00**	**0.00**	0.020	0.019	0.017	0.022	0.023	0.023
SHIFT	**0.02**	0.14	0.06	0.30	0.30	0.30	0.013	0.013	0.014	0.013	0.014	0.013
RVR	**0.03**	**0.03**	0.14	0.53	0.53	0.53	0.020	0.018	0.018	0.017	0.018	0.018
THC	0.30	0.53	0.78	0.20	0.83	0.83	0.012	0.013	0.015			
CLAM	0.53	0.83	0.92	0.53	0.53	0.53	0.025	0.025	0.030	0.024	0.024	0.023
GAME	**0.02**	0.06	0.53	0.78	0.53	0.53	0.018	0.016	0.014	0.014	0.014	0.014
MINT	**0.01**	0.06	0.53	0.78	0.78	0.78	0.024	0.024	0.025	0.025	0.025	0.026
ABY	**0.01**	**0.01**	0.14	0.06	0.14	0.06	0.019	0.014	0.014	0.016	0.017	0.017
GLD	0.78	0.53	0.92	0.20	0.20	0.20	0.022	0.020		0.020	0.021	0.020
EFL	0.78	0.06	0.83	**0.02**	**0.02**	**0.02**	0.023	0.018	0.022	0.023	0.023	0.022
portfolio	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.009	0.008	0.009	0.007	0.008	0.007
	Christoffersen's test p-value: VaR 3\%						**Asymmetric VaR 3% Loss and MCS**					
BTC	**0.00**	0.07	0.25	0.83	0.07	0.07	0.017			0.016	0.016	0.016
ETH	0.08	0.05	0.25	0.43	0.36	0.36	0.019	0.015	0.016	0.016	0.017	0.016
XRP	0.18	**0.04**	**0.04**	0.83	0.83	0.83	0.016	0.017	0.016	0.015	0.015	0.015
XLM	**0.01**	**0.04**	0.18	0.52	0.33	0.33	0.017	0.017		0.017	0.017	0.017
LTC	0.18	**0.02**	0.63	0.16	0.38	0.63	0.017	0.014	0.015	0.015	0.014	0.014
XCP	**0.04**	0.15	**0.02**	**0.04**	**0.01**	**0.01**	0.026	0.025	0.023	0.029	0.029	0.029
SHIFT	**0.04**	**0.02**	0.09	0.33	0.33	0.33	0.018	0.018	0.019	0.018	0.019	0.018
RVR	0.15	**0.04**	0.18	0.25	0.25	0.25	0.025	0.024	0.024	0.023	0.024	0.023
THC	0.71	0.08	0.83	0.63	0.63	0.63	0.016	0.018				
CLAM	0.52	0.81	0.63	0.38	0.38	0.38	0.033	0.035	0.040	0.032	0.032	0.032
GAME	0.08	0.09	0.52	0.33	0.18	0.18	0.023	0.021	0.018	0.020	0.021	0.020
MINT	0.15	**0.04**	0.52	0.33	0.33	0.33	0.033	0.032	0.034	0.035	0.035	0.035
ABY	**0.02**	**0.02**	**0.00**	**0.04**	**0.04**	**0.04**	0.033	0.032	0.034	0.035	0.035	0.035
GLD	0.83	0.44	0.81	0.63	0.38	0.38	0.028	0.025		0.026	0.026	0.026
EFL	0.81	0.09	0.63	0.04	**0.04**	**0.04**	0.029	0.025		0.031	0.030	0.029
portfolio	**0.00**	**0.00**	**0.00**	**0.02**	**0.00**	**0.00**	0.010	0.010	0.010	0.009	0.009	0.009
	Christoffersen's test p-value: VaR 4\%						**Asymmetric VaR 4% Loss and MCS**					
BTC	**0.00**	**0.03**	0.16	0.16	**0.03**	**0.03**	0.025	0.022	0.023	0.020	0.021	0.021
ETH	0.09	**0.03**	**0.03**	0.66	0.17	0.18	0.022	0.019	0.020	0.020	0.020	0.020
XRP	0.30	**0.02**	0.10	0.73	0.71	0.71	0.020	0.021	0.020	0.019	0.020	0.019
XLM	0.07	0.07	0.30	0.61	0.61	0.46	0.021	0.020		0.021	0.021	0.020
LTC	0.46	0.04	0.26	0.26	0.46	0.63	0.021	0.019	0.018	0.018	0.017	0.017
XCP	0.24	0.16	**0.03**	0.05	0.05	**0.03**	0.031	0.030	0.029	0.034	0.035	0.034
SHIFT	0.10	0.05	0.10	0.30	0.30	0.30	0.022	0.022	0.023	0.023	0.023	0.023
RVR	0.24	**0.02**	0.30	0.63	0.52	0.52	0.030	0.028	0.028	0.028	0.028	0.028
THC	0.73	0.14	0.63	0.46	0.46	0.46	0.019	0.021				
CLAM	0.30	0.46	0.63	0.26	0.26	0.46	0.041	0.043	0.048	0.041	0.041	0.040
GAME	0.24	0.14	0.52	0.52	0.37	0.37	0.027	0.026	0.022	0.026	0.026	0.025
MINT	0.14	0.10	0.61	0.61	0.61	0.61	0.040	0.040	0.041	0.043		
ABY	0.05	**0.00**	0.04	**0.02**	**0.02**	**0.02**	0.029	0.025	0.028	0.030	0.031	0.031
GLD	0.71	0.13	0.45	0.63	0.46	0.63	0.033	0.030		0.031	0.031	0.031
EFL	0.63	0.24	0.46	0.04	0.04	0.12	0.035	0.031		0.037	0.036	0.036
portfolio	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.012	0.012	0.012	0.011	0.011	0.011
	Christoffersen's test p-value: VaR 5\%						**Asymmetric VaR 5% Loss and MCS**					
BTC	**0.03**	**0.00**	0.09	0.09	0.08	0.08	0.029	0.026	0.027	0.025	0.026	0.025
ETH	0.16	0.08	0.09	0.67	0.33	0.33	0.026	0.023	0.024	0.024	0.024	0.024
XRP	0.07	**0.01**	0.05	0.58	0.58	0.58	0.023	0.025	0.024	0.023	0.023	0.023
XLM	0.31	0.18	0.31	0.59	0.49	0.49	0.024	0.023		0.024	0.024	0.024
LTC	0.63	**0.01**	0.58	0.18	0.58	0.62	0.023	0.022	0.020	0.021	0.020	0.020
XCP	0.22	0.16	0.10	0.10	0.10	0.10	0.036	0.034	0.034	0.039	0.039	0.039
SHIFT	0.16	**0.01**	0.31	0.37	0.37	0.37	0.025	0.025	0.027	0.026	0.027	0.026
RVR	0.63	**0.02**	0.29	0.84	0.77	0.77	0.033	0.033	0.033	0.032	0.032	0.032
THC	0.62	0.10	0.46	0.18	0.18	0.18	0.022	0.024				
CLAM	0.37	0.16	0.46	0.32	0.18	0.32	0.048	0.051	0.055	0.048	0.048	0.047
GAME	0.63	0.31	0.77	0.77	0.77	0.77	0.031	0.029	0.026	0.030	0.030	0.030
MINT	0.31	0.05	0.62	0.59	0.59	0.59	0.047	0.046	0.048			
ABY	0.16	**0.00**	**0.02**	0.09	0.09	0.09	0.033	0.030	0.032	0.035	0.036	0.036
GLD	0.58	0.10	0.84	0.62	0.58	0.58	0.037	0.035		0.036	0.037	0.036
EFL	0.46	0.10	0.32	**0.03**	**0.03**	**0.03**	0.039	0.036			0.043	
portfolio	**0.00**	**0.00**	**0.00**	**0.01**	**0.01**	**0.01**	0.013	0.013	0.014	0.012	0.012	0.012

Table 10: Christoffersen's tests, asymmetric VaR Loss and MCS (an empty cell means the model is not included) for each coin and for the equally-weighted portfolio.

	DCC MVN	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*	*DCC MVN*	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*
	ES 5\% test 1: Exceedance Residuals test						**ES 5\% test 2: multilevel VaR backtest**					
BTC	0.99	1.00	0.76	1.00	1.00	1.00	**0.00**	**0.01**	0.21	0.36	0.10	0.10
ETH	0.97	0.99	0.70	0.32	0.49	0.46	**0.00**	0.06	0.06	0.90	0.26	0.22
XRP	0.67	0.96	0.91	0.98	0.96	0.96	**0.02**	**0.00**	**0.03**	0.94	0.98	0.98
XLM	**0.01**	1.00	0.04	0.11	0.14	0.12	**0.00**	**0.00**	0.10	0.47	0.37	0.34
LTC	0.26	1.00	0.85	0.99	1.00	1.00	0.29	**0.00**	0.36	**0.00**	**0.03**	0.07
XCP	0.96	1.00	0.54	0.06	**0.04**	**0.03**	**0.00**	0.06	**0.00**	**0.00**	**0.00**	**0.00**
SHIFT	0.63	1.00	0.35	0.07	0.10	0.09	**0.01**	**0.00**	**0.02**	0.39	0.39	0.39
RVR	0.49	0.99	0.34	0.17	0.12	0.10	**0.02**	**0.00**	0.13	0.50	0.39	0.39
THC	0.10	0.99	0.12	0.25	0.20	0.20	0.71	**0.02**	0.79	0.10	0.28	0.28
CLAM	0.90	0.97	0.15	0.94	0.93	0.96	0.62	0.66	0.71	0.15	0.10	0.22
GAME	0.87	0.98	0.47	0.33	0.23	0.22	**0.01**	**0.02**	0.62	0.61	0.32	0.32
MINT	0.95	1.00	0.52	0.63	0.65	0.79	**0.00**	**0.01**	0.82	0.82	0.82	0.82
ABY	0.90	1.00	0.86	0.39	0.42	0.29	**0.00**	**0.00**	**0.00**	**0.01**	**0.01**	**0.01**
GLD	0.87	1.00	0.78	0.91	0.86	0.85	0.96	0.11	0.98	0.38	0.19	0.24
EFL	0.94	1.00	0.93	0.97	0.98	0.97	0.76	**0.01**	0.42	**0.00**	**0.00**	**0.00**
portfolio	0.54	0.99	**0.00**	0.05	**0.02**	**0.02**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Table 11: ER test by McNeil and Frey (2000) and multilevel VaR backtest by Kratz et al. (2018).

Ticker	*Name*	*Ticker*	*Name*	*Ticker*	*Name*
BTC	Bitcoin	SC	Siacoin	MUE	MonetaryUnit
ETH	Ethereum	GAME	GameCredits	RADS	Radium
XRP	XRP	DMD	Diamond	OMNI	Omni
XLM	Stellar	THC	HempCoin	MINT	MintCoin
LTC	Litecoin	BLK	BlackCoin	ABY	ArtByte
USDT	Tether	AMP	Synereo	GLD	GoldCoin
XMR	Monero	CLAM	Clams	HUC	HunterCoin
DASH	Dash	PND	Pandacoin	EFL	e-Gulden
XEM	NEM	GRC	GridCoin	EGC	EverGreenCoin
DOGE	Dogecoin	POT	PotCoin	ORB	Orbitcoin
LSK	Lisk	IOC	I/O Coin	WDC	WorldCoin
BTS	BitShares	XMY	Myriad	GUN	Guncoin
DCR	Decred	FLO	FLO	TTC	TittieCoin
DGB	DigiByte	XST	Stealth	QTL	Quatloo

Table 12: Set of coins used for credit risk measurement

The last month trading volume, the one-year trading volume and the average yearly search volume index as provided by Google Trends were used as regressors for the logit, probit and random forest models[17]. For computing the ZPP, we employed the random walk model with the closed-formula (2), the univariate GARCH(1,1) model with normal errors and the closed-formula (3), univariate GARCH(1,1) models with student's t and skewed-t errors, DCC(1,1) models with standardized errors following either a multivariate normal or a multivariate t distribution, and Copula-GARCH models with skewed-t marginals and either a normal copula or a t-copula (both with constant and dynamic parameters). The AUC scores, the Brier score and the models included in the MCS, for both the training and the validation datasets, are reported in Table 13.

Classical credit scoring models performed better in the training sample, whereas the models perfor-

[17]The *Search Volume Index* by Google Trends computes how many searches have been done for a keyword or a topic on Google over a specific period of time and a specific region. See `https://support.google.com/trends/?hl=en` for more details.

Model	*AUC (training)*	*AUC (validation)*	*Brier score/MCS (training)*	*Brier score/MCS (validation)*
Logit	0.91	0.91	0.097	0.125
Probit	0.91	0.90	0.097	0.124
Random forest	0.80	0.90	0.154	0.103
ZPP RW	0.75	0.81	0.180	0.300
ZPP GARCH n	0.40	0.52		
ZPP GARCH t	0.56	0.63	0.187	0.301
ZPP GARCH sk.t	0.59	0.66	0.186	0.303
ZPP DCC n	0.60	0.76	0.168	0.277
ZPP DCC t	0.48	0.66	0.143	0.233
ZPP N.Copula sk.t-G.	0.56	0.73	0.230	0.262
ZPP T.Copula sk.t-G.	0.58	0.70		0.297
ZPP N.Copula DCC sk.t-G	0.59	0.72	0.246	0.280
ZPP T.Copula DCC sk.t-G	0.57	0.74	0.200	0.228

Table 13: AUC, Brier scores and MCS (an empty cell means the model is not included).

mances are much closer when the validation/out-of-time sample is considered. In the latter case, the ZPPs computed with multivariate models improved their performances in term of AUC, but they are still worse then credit scoring models. However, almost all models are included in the MCS, thus showing that there is insufficient information in the data to separate good and bad models: this outcome was expected due to the small sample size involved. Interestingly, the ZPP computed with the simple random walk with drift performed remarkably well, thus confirming the empirical evidence by Li et al. (2016). This result and the improved performance of the ZPP computed with multivariate models can probably be explained using again the simulation results in Fantazzini (2009a), who showed that the biased parameters of a multivariate distribution computed with a small sample of skewed and leptokurtic data can deliver conservative VaR estimates in the left tail of the distribution: considering that the left tail of the P&L distribution is the key element to compute the ZPP, a positively biased probability may have helped these models to forecast the riskiest coins. Another reason which could have contributed to the better performance of all multivariate models in the validation sample is the different coins' dependence structure in the training and validation time samples. Figure 1 reports the heat maps of the correlation matrices for all coins in the training sample (left) and the validation sample (right): it is evident that the second sample witnessed a much stronger dependence among coins than the first sample, and we would have a similar picture if we used measures of rank correlation like the Kendall' tau and the Spearman's rho (not reported). For example, the average correlation between coins in the training sample was 0.15, whereas it was 0.32 in the validation sample. This stronger dependence may explain why multivariate models performed better than the univariate models thanks to a larger information set.

Finally, we remark that the AUCs estimated with an out-of-time validation dataset can be higher than those estimated with an in-sample dataset at a previous time, because new future data may well

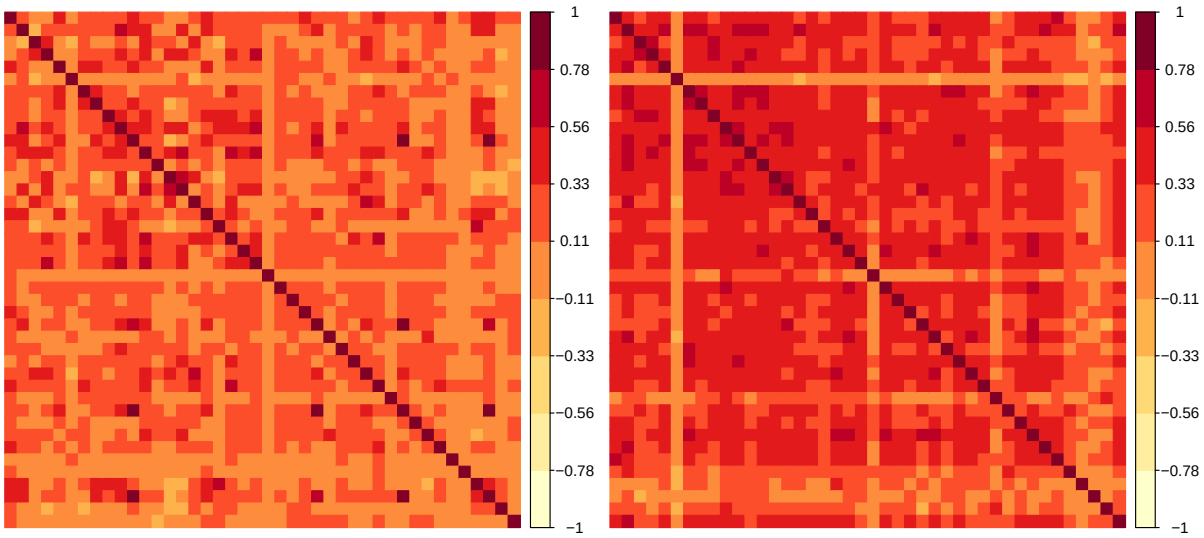


Figure 1: Heat maps of the correlation matrices in the training sample (June 2016 - May 2017, left) and the validation sample (June 2017 - May 2018, right)

come from a different distribution, see Figure 1 for an explicative example. Moreover, an estimated AUC may widely differ from the true metric in a small sample. In this regard, Hanczar et al. (2010) performed an extensive simulation study with several models and cross-section training and test datasets, and they showed that "*(i) for small samples the root mean square differences of the estimated and true metrics are considerable; (ii) even for large samples, there is only weak correlation between the true and estimated metrics; and (iii) generally, there is weak regression of the true metric on the estimated metric*" (Hanczar et al. (2010), p. 822). Interestingly, they found that with $N = 50$ the estimated AUCs may differ from the true AUC with amounts up to ± 0.2: given that we worked with temporal datasets and parameter-intensive models, this range may be even higher. Besides, this simulation evidence also helps to explain why the MCS included almost all models. All metrics computed by Hanczar et al. (2010) with small simulated samples showed very large variances and any model selection based on these metrics would be difficult: hence the inability of the MCS approach to distinguish between good and bad models.

5 Robustness checks

We wanted to verify that our previous results hold also with different portfolios and different models. Therefore, we performed a series of robustness checks, considering a capitalization-weighted portfolio instead of an equally-weighted portfolio, and we computed the market risk measures assuming the cryptocurrencies P&L to be comonotonic.

5.1 Capitalization-weighted portfolios

The equally-weighted portfolio is the main benchmark of the financial industry and literature, see e.g. DeMiguel et al. (2007) and references therein. Moreover, Hu et al. (2018) and Brauneis and Mestel (2019) analyzed cryptocurrency-portfolios and showed that the simple 1/N portfolio outperforms all competing portfolio strategies. However, most investors in crypto-currencies use capitalization-weighted portfolios, where individual assets are weighted according to their total market capitalization see, for example, the monthly trading volume rankings published at `coinmarketcap.com/currencies/volume/monthly`. Moreover, capitalization-weighted portfolios are important benchmarks in the financial literature, see Sharpe (1992), Black and Litterman (1992) and Reilly and Brown (2002). Therefore, we performed the previous back-testing analysis with capitalization-weighted portfolios considering 5 and 15 coins, respectively. The actual VaR exceedances T_1/T, the p-values of the Kupiec's unconditional coverage test, the p-values of the Christoffersen's conditional coverage test, the p-values of the exceedance residuals test by McNeil and Frey (2000) and of the multilevel VaR backtest by Kratz et al. (2018) for the forecasted 5%ES are reported in Table 14.

	DCC MVN	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*	*DCC MVN*	*DCC MVT*	*Const N-Copula GARCH*	*Const T-Copula GARCH*	*DCC N-Copula GARCH*	*DCC T-Copula GARCH*
	VaR exceedances (T_1/T): 1\% quantile						**Kupiec's test p-value: VaR 1\%**					
5 coins	3.52%	1.51%	3.52%	2.51%	2.51%	2.01%	**0.01**	0.50	0.01	0.07	0.07	0.21
15 coins	6.53%	4.52%	6.53%	3.52%	4.52%	4.02%	**0.00**	**0.00**	**0.00**	**0.01**	**0.00**	**0.00**
	VaR exceedances (T_1/T): 2\% quantile						**Kupiec's test p-value: VaR 2\%**					
5 coins	5.53%	5.53%	5.53%	5.03%	5.53%	5.03%	**0.00**	**0.00**	**0.00**	**0.01**	**0.00**	**0.01**
15 coins	7.04%	6.53%	7.04%	5.53%	5.53%	5.53%	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	VaR exceedances (T_1/T): 3\% quantile						**Kupiec's test p-value: VaR 3\%**					
5 coins	6.53%	7.04%	8.54%	6.53%	6.53%	6.53%	**0.01**	**0.00**	**0.00**	**0.01**	**0.01**	**0.01**
15 coins	8.04%	8.54%	7.54%	6.03%	6.53%	6.53%	**0.00**	**0.00**	**0.00**	**0.03**	**0.01**	**0.01**
	VaR exceedances (T_1/T): 4\% quantile						**Kupiec's test p-value: VaR 4\%**					
5 coins	8.04%	9.55%	9.55%	9.05%	8.04%	8.54%	**0.01**	**0.00**	**0.00**	**0.00**	**0.01**	**0.00**
15 coins	8.54%	11.56%	9.05%	7.04%	6.53%	6.53%	**0.00**	**0.00**	**0.00**	0.05	0.09	**0.09**
	VaR exceedances (T_1/T): 5\% quantile						**Kupiec's test p-value: VaR 5\%**					
5 coins	9.05%	12.06%	9.55%	9.55%	10.55%	10.55%	**0.02**	**0.00**	**0.01**	**0.01**	**0.00**	**0.00**
15 coins	10.05%	12.06%	10.55%	9.05%	9.05%	9.05%	**0.00**	**0.00**	**0.00**	**0.02**	**0.02**	**0.02**
	Christoffersen's test p-value: VaR 1\%						**ES 5\% test 1: Exceedance Residuals test**					
5 coins	**0.02**	0.76	0.02	0.17	0.17	0.42	0.45	0.99	0.21	0.50	0.54	0.68
15 coins	**0.00**	**0.00**	**0.00**	**0.02**	**0.00**	**0.00**	0.74	0.99	0.06	0.38	0.23	0.26
	Christoffersen's test p-value: VaR 2\%						**ES 5\% test 2: multilevel VaR backtest**					
5 coins	0.01	0.01	0.01	0.02	0.01	0.02	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
15 coins	**0.00**	**0.00**	**0.00**	**0.01**	**0.01**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Christoffersen's test p-value: VaR 3\%											
5 coins	**0.02**	**0.01**	**0.00**	**0.02**	**0.02**	**0.02**						
15 coins	**0.00**	**0.00**	**0.01**	0.08	**0.04**	**0.04**						
	Christoffersen's test p-value: VaR 4\%											
5 coins	**0.03**	**0.00**	**0.00**	**0.00**	**0.01**	**0.01**						
15 coins	**0.02**	**0.00**	**0.00**	0.14	0.24	0.24						
	Christoffersen's test p-value: VaR 5\%											
5 coins	0.06	**0.00**	**0.02**	**0.02**	**0.01**	**0.01**						
15 coins	**0.02**	**0.00**	**0.00**	0.06	0.06	0.06						

Table 14: Capitalization-weighted portfolios risk measures: Kupiec's unconditional coverage test, Christoffersen's conditional coverage test, ER test by McNeil and Frey (2000), multilevel VaR backtest by Kratz et al. (2018).

The risk estimates are slightly more precise than the corresponding estimates with equally-weighted portfolios, probably due to the less extreme distributions of top cryptocurrencies compared to coins with small capitalizations. Nevertheless, the null hypotheses of the Kupiec's unconditional coverage test,

Christoffersen's conditional coverage test, and the multilevel VaR backtest are again rejected for most portfolio risk estimates, while the ER test by McNeil and Frey (2000) does not highlight any particular misspecification for the ES 5%. T-copula based models often provide the most precise estimates and, in general, copula models tend to fare better than DCC models with large portfolios than with small portfolios.

5.2 Comonotonic assets

A set of random variables is comonotonic if they are perfectly positively dependent, which means that they are almost surely strictly increasing functions of a single random variable, see McNeil et al. (2015) for a detailed discussion at the textbook level. In this case, the Value-at-Risk and the expected shortfall are additive for a sum of comonotonic random variables, see Dhaene et al. (2006) for a proof. Therefore, portfolio risk measures can be computed using only the marginal models without the need to estimate the dependence structure (which is represented by the so-called *comonotonicity copula*).

The actual VaR exceedances T_1/T, the p-values of the Kupiec's unconditional coverage test, the p-values of the Christoffersen's conditional coverage test, the p-values of the exceedance residuals test by McNeil and Frey (2000) and of the multilevel VaR backtest by Kratz et al. (2018) for the forecasted 5%ES in case of an equally-weighted portfolio are reported in Table 15.

	Marginal models			***Marginal models***			***Marginal models***			***Marginal models***		
	Normal GARCH	*Student GARCH*	*Skewed-t GARCH*	*Normal GARCH*	*Student GARCH*	*Skewed-t GARCH*	*Normal GARCH*	*Student GARCH*	*Skewed-t GARCH*	*Normal GARCH*	*Student GARCH*	*Skewed-t GARCH*
	VaR exceedances (T_1/T): 1%			**Kupiec's test p.: VaR 1%**			**C. test p-value: VaR 1%**			**ES 5%: E.R. test**		
5 coins	1.01%	1.01%	0.00%	0.99	0.99	0.05	0.98	0.98	0.14	0.85	0.98	0.99
15 coins	1.51%	0.00%	1.01%	0.50	0.05	0.99	0.76	0.14	0.98	0.62	1.00	0.83
	VaR exceedances (T_1/T): 2%			**Kupiec's test p.: VaR 2%**			**C. test p-value: VaR 2%**			**ES 5%: m.VaR backtest**		
5 coins	2.01%	1.51%	1.01%	0.99	0.60	0.27	0.92	0.83	0.53	0.96	0.79	0.12
15 coins	1.51%	1.01%	1.01%	0.60	0.27	0.27	0.83	0.53	0.53	0.49	0.75	0.12
	VaR exceedances (T_1/T): 3%			**Kupiec's test p.: VaR 3%**			**C. test p-value: VaR 3%**					
5 coins	3.02%	4.02%	1.01%	0.99	0.42	0.06	0.83	0.52	0.16			
15 coins	2.51%	2.01%	1.01%	0.68	0.38	0.06	0.81	0.63	0.16			
	VaR exceedances (T_1/T): 4%			**Kupiec's test p.: VaR 4%**			**C. test p-value: VaR 4%**					
5 coins	3.02%	5.03%	2.01%	0.46	0.48	0.11	0.63	0.63	0.26			
15 coins	2.51%	4.02%	2.01%	0.25	0.99	0.11	0.46	0.71	0.26			
	VaR exceedances (T_1/T): 5%			**Kupiec's test p.: VaR 5%**			**C. test p-value: VaR 5%**					
5 coins	4.52%	6.03%	3.02%	0.75	0.52	0.17	0.67	0.33	0.32			
15 coins	3.02%	4.52%	3.52%	0.17	0.75	0.31	0.32	0.62	0.46			

Table 15: Equally-weighted portfolios risk measures with comonotonic assets: Kupiec's unconditional coverage test, Christoffersen's conditional coverage test, ER test by McNeil and Frey (2000), multilevel VaR backtest by Kratz et al. (2018).

All estimated portfolio risk measures pass the backtesting specification tests, with the student's t GARCH marginals providing the most accurate measures, whereas the other marginal models are slightly more conservative. These results will definitely not surprise traders and financial professionals dealing with crypto-currencies, who several times repeated that the crypto-market is still a "one-man show" driven by the bitcoin price, see e.g. Mitsuru et al. (2014), Bitconnect (2017) and DeMichele (2018).

6 Conclusions

Credit and market risks for cryptocurrencies are more interlinked than for traditional assets, and their differences are of quantitative and temporal nature, not qualitative. Credit risk for cryptocurrencies can be defined as the gains and losses on the value of a position of a cryptocurrency that is abandoned and considered dead but which can be potentially revived, while market risk can be described as the gains and losses on the value of a position (or portfolio) of alive cryptocurrencies, due to the movements in market prices in centralized and decentralized exchanges.

This paper proposed a set of models to estimate simultaneously both the market risk and the credit risk for a portfolio of crypto-currencies. Moreover, two closed-form formulas for the ZPP model in case of normally distributed errors were also developed using recent results from barrier option theory. These formulas can be useful to get a rough idea of the crypto-asset credit risk and may be interesting to online data providers, who can publish the quotes of crypto-assets together with their market-implied credit risk measures.

A backtesting exercise for market risk modelling using two datasets of 5 and 15 coins was performed. Our results showed that the t-copula with skewed-t GARCH marginals can be a good compromise for precise VaR estimates across different quantile levels, particularly the most extreme quantiles (1% and 2%) which are the most important for regulatory purposes. However, the asymmetric VaR losses of the competing models were rather close and all models were included in the MCS for almost all coins and for the portfolio case: this result was expected because Fantazzini (2009a) showed that, when small samples are considered and the data are skewed and leptokurtic, the biases in the GARCH parameters are so large that they can deliver conservative VaR estimates, even with a simple multivariate normal distribution. The backtesting of the expected shortfall estimates showed that DCC models often underestimated the true ES, whereas t-copula/skewed-t GARCH models were generally fine.

A backtesting exercise for credit risk modelling was performed using a dataset of 42 coins. The empirical analysis showed that classical credit scoring models performed better in the training sample, whereas the models' performances were much closer in the validation sample, with the simple ZPP computed using a random walk with drift performing remarkably well. In general, all multivariate models performed much better in the validation sample than in the training sample, thanks to the much stronger dependence shown by coins in the validation sample.

Finally, we performed a set of robustness checks to verify that our results also hold under different forecasting setups. We found out that risk estimates using capitalization-weighted portfolios are slightly more precise than those with equally-weighted portfolios, probably due to the less extreme distributions of top cryptocurrencies compared to coins with small capitalizations. Moreover, market risk measures

computed assuming the cryptocurrencies P&L to be comonotonic passed all the backtesting specification tests, thus confirming financial professional literature showing that crypto-currencies are mainly driven by the bitcoin price.

The extreme volatility, skewness, kurtosis, and dependence of cryptocurrencies due to the frequent presence of structural breaks and price manipulations pose a serious challenge to any multivariate risk model. Skewed-t marginals and copulas allowing for dynamic dependence are a good starting point, but the length of the estimation window plays also an important role due to multiple breaks of different nature. In this regard, a rolling estimation window of approximately 500 observations can be considered a good compromise, considering the numerical properties of GARCH models discussed by Hwang and Valls Pereira (2006), together with the simulation evidence reported by Pesaran and Timmermann (2007), who showed that in a regression with multiple breaks the optimal window for estimation includes all of the observations after the last break, plus a limited number of observations before this break. This setup is viable with portfolios of small sizes, but it is hardly feasible with portfolios of medium and large sizes, due to a large number of parameters involved. Moreover, the strong dependence of cryptocurrencies from the bitcoin price becomes much more complex and unstable to model with larger portfolios. In such a situation, a numerically efficient solution is to assume the assets to be comonotonic.

The number of coins that we used for backtesting is rather low, and this can be a potential limitation of our analysis. However, we want to remark that we considered only small and medium-sized portfolios to avoid additional model complexity and due to the lack of historical data for the vast majority of dead coins. The last issue is certainly a major stumbling block for fully developing models for credit risk measurement and management with crypto-assets, because the lack of these data would make any model suffer from massive selection bias: for example, in the middle of 2017, there were more than 1500 traded alive coins and almost 800 dead coins, but historical market data were available only for a few dozens dead coins. The retrieval and the analysis of such data are left as an avenue for future research. Another potential future development is the analysis of the interactions effects between market risk and credit risk for cryptocurrencies and how they change over time.

References

Acerbi, Carlo, and Balazs Szekely. 2014. "Back-Testing Expected Shortfall." *Risk* 27: 76–81.

Adalet McGowan, Muge, Dan Andrews, and Valentine Millot. 2018. "The Walking Dead? Zombie Firms and Productivity Performance in Oecd Countries." *Economic Policy* 33 (96): 685–736.

Aielli, Gian Piero. 2013. "Dynamic Conditional Correlation: On Properties and Estimation." *Journal of Business & Economic Statistics* 31 (3): 282–99.

Altman, Edward I, and Gabriele Sabato. 2007. "Modelling Credit Risk for Smes: Evidence from the Us Market." *Abacus* 43 (3): 332–57.

Antonopoulos, Andreas M. 2014. *Mastering Bitcoin: Unlocking Digital Cryptocurrencies.* " O'Reilly Media, Inc."

Back, Adam. 2002. "Hashcash-a Denial of Service Counter-Measure."

Balcilar, Mehmet, Elie Bouri, Rangan Gupta, and David Roubaud. 2017. "Can Volume Predict Bitcoin Returns and Volatility? A Quantiles-Based Approach." *Economic Modelling* 64: 74–81.

Bali, Turan G, and Hao Zhou. 2016. "Risk, Uncertainty, and Expected Returns." *Journal of Financial and Quantitative Analysis* 51 (3): 707–35.

Banerjee, Ryan, and Boris Hofmann. 2018. "The Rise of Zombie Firms: Causes and Consequences." *BIS Quarterly Review, September*, 67–78.

Basel Committee on Banking Supervision. 2009. "Findings on the Interaction of Market and Credit Risk." Bank for International Settlements, Working paper n. 16, May.

Bauwens, Luc, and Sebastien Laurent. 2005. "A New Class of Multivariate Skew Densities, with Application to Generalized Autoregressive Conditional Heteroscedasticity Models." *Journal of Business & Economic Statistics* 23 (3): 346–54.

Bauwens, Luc, Christian M Hafner, and Sebastien Laurent. 2012. *Handbook of Volatility Models and Their Applications.* Vol. 3. Wiley.

Bauwens, Luc, Sebastien Laurent, and Jeroen VK Rombouts. 2006. "Multivariate Garch Models: A Survey." *Journal of Applied Econometrics* 21 (1): 79–109.

Bayraktar, Erhan, and Bo Yang. 2011. "A Unified Framework for Pricing Credit and Equity Derivatives." *Mathematical Finance* 21 (3): 493–517.

BIS. 2013. "Fundamental Review of the Trading Book: A Revised Market Risk Framework." *Basel Committee on Banking Supervision.*

———. 2016. "Minimum Capital Requirements for Market Risk - Publication No. 352." *Basel*

Committee on Banking Supervision.

Bitconnect. 2017. "How Is the Price of Cryptocurrency Defined?" Bitconnect editorial board.

Black, Fischer, and Robert Litterman. 1992. "Global Portfolio Optimization." *Financial Analysts Journal* 48 (5): 28–43.

Bohr, Jeremiah, and Masooda Bashir. 2014. "Who Uses Bitcoin? An Exploration of the Bitcoin Community." In *2014 Twelfth Annual International Conference on Privacy, Security and Trust*, 94–101. IEEE.

Bouoiyour, Jamal, and Refk Selmi. 2015. "What Does Bitcoin Look Like?" *Annals of Economics & Finance* 16 (2).

Bouoiyour, Jamal, Refk Selmi, and Aviral Kumar Tiwari. 2015. "Is Bitcoin Business Income or Speculative Foolery? New Ideas Through an Improved Frequency Domain Analysis." *Annals of Financial Economics* 10 (01): 1550002.

Bouri, Elie, Georges Azzi, and Anne Haubo Dyhrberg. 2017. "On the Return-Volatility Relationship in the Bitcoin Market Around the Price Crash of 2013." *Economics: The Open-Access, Open-Assessment E-Journal* 11 (2): 1–16.

Bouri, Elie, Luis A Gil-Alana, Rangan Gupta, and David Roubaud. 2016. "Modelling Long Memory Volatility in the Bitcoin Market: Evidence of Persistence and Structural Breaks." *International Journal of Finance & Economics*, 24(1): 412-426 .

Brandvold, Morten, Peter Molnar, Kristian Vagstad, and Ole Christian Andreas Valstad. 2015. "Price Discovery on Bitcoin Exchanges." *Journal of International Financial Markets, Institutions and Money* 36: 18–35.

Brauneis, Alexander, and Roland Mestel. 2019. "Cryptocurrency-Portfolios in a Mean-Variance Framework." *Finance Research Letters* forthcoming.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Buchholz, Martis, Jess Delaney, Joseph Warren, and Jeff Parker. 2012. "Bits and Bets, Information, Price Volatility, and Demand for Bitcoin." *Economics* 312.

Burniske, Chris, and Jack Tatar. 2017. *Cryptoassets:The Innovative Investors Guide to Bitcoin and Beyond*. McGraw Hill Professional.

Caballero, Ricardo J, Takeo Hoshi, and Anil K Kashyap. 2008. "Zombie Lending and Depressed Restructuring in Japan." *American Economic Review* 98 (5): 1943–77.

Cai, Yong, and K Krishnamoorthy. 2006. "Exact Size and Power Properties of Five Tests for

Multinomial Proportions." *Communications in Statistics-Simulation and Computation* 35 (1): 149–60.

Campi, Luciano, Simon Polbennikov, and Alessandro Sbuelz. 2009. "Systematic Equity-Based Credit Risk: A Cev Model with Jump to Default." *Journal of Economic Dynamics and Control* 33 (1): 93–108.

Caporin, Massimiliano, and Michael McAleer. 2013. "Ten Things You Should Know About the Dynamic Conditional Correlation Representation." *Econometrics* 1 (1): 115–26.

———. 2014. "Robust Ranking of Multivariate Garch Models by Problem Dimension." *Computational Statistics & Data Analysis* 76: 172–85.

Carr, Peter, and Vadim Linetsky. 2006. "A Jump to Default Extended Cev Model: An Application of Bessel Processes." *Finance and Stochastics* 10 (3): 303–30.

Carr, Peter, and Liuren Wu. 2009. "Stock Options and Credit Default Swaps: A Joint Framework for Valuation and Estimation." *Journal of Financial Econometrics* 8 (4): 409–49.

Catania, Leopoldo, Stefano Grassi, and Francesco Ravazzolo. 2018. "Predicting the Volatility of Cryptocurrency Time–Series." In *Mathematical and Statistical Methods for Actuarial Sciences and Finance*.

Chan, Stephen, Jeffrey Chu, Saralees Nadarajah, and Joerg Osterrieder. 2017. "A Statistical Analysis of Cryptocurrencies." *Journal of Risk and Financial Management* 10 (2): 12.

Chaum, David. 1983. "Blind Signatures for Untraceable Payments." In *Advances in Cryptology*, 199–203.

Chaum, David, and Stefan Brands. 1997. "'Minting'electronic Cash." *IEEE Spectrum* 34 (2). IEEE: 30–34.

Cheah, Eng-Tuck, and John Fry. 2015. "Speculative Bubbles in Bitcoin Markets? An Empirical Investigation into the Fundamental Value of Bitcoin." *Economics Letters* 130: 32–36.

Cherubini, Umberto, Elisa Luciano, and Walter Vecchiato. 2004. *Copula Methods in Finance*. John Wiley & Sons.

Christoffersen, Peter. 2011. *Elements of Financial Risk Management*. Academic Press.

Christoffersen, Peter F. 1998. "Evaluating Interval Forecasts." *International Economic Review*, 841–862.

Chu, Jeffrey, Stephen Chan, Saralees Nadarajah, and Joerg Osterrieder. 2017. "GARCH Modelling of Cryptocurrencies." *Journal of Risk and Financial Management* 10 (4): 17.

Chu, Jeffrey, Saralees Nadarajah, and Stephen Chan. 2015. "Statistical Analysis of the Exchange Rate of Bitcoin." *PloS One* 10 (7): e0133678.

Ciaian, Pavel, Miroslava Rajcaniova, and d'Artis Kancs. 2016. "The Digital Agenda of Virtual Currencies: Can Bitcoin Become a Global Currency?" *Information Systems and E-Business Management*

14 (4): 883–919.

Corbet, Shaen, Brian Lucey, and Larisa Yarovaya. 2018. "Datestamping the Bitcoin and Ethereum Bubbles." *Finance Research Letters* 26: 81–88.

Corbet, Shaen, Brian Lucey, Andrew Urquhart, and Larisa Yarovaya. 2019. "Cryptocurrencies as a Financial Asset: A Systematic Analysis." *International Review of Financial Analysis* forthcoming.

Dalla Valle, Luciana, Maria Elena De Giuli, Claudia Tarantola, and Claudio Manelli. 2016. "Default Probability Estimation via Pair Copula Constructions." *European Journal of Operational Research* 249 (1): 298–311.

Das, Sanjiv R, and Paul Hanouna. 2009. "Implied Recovery." *Journal of Economic Dynamics and Control* 33 (11): 1837–57.

DeMichele, Thomas. 2018. "Why Do Altcoin Prices Often Follow Bitcoin Price?" Cryptocurrency facts.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal. 2007. "Optimal Versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy?" *The Review of Financial Studies* 22 (5): 1915–53.

Dhaene, Jan, Steven Vanduffel, Marc J Goovaerts, Rob Kaas, Qihe Tang, and David Vyncke. 2006. "Risk Measures and Comonotonicity: A Review." *Stochastic Models* 22 (4): 573–606.

Dyhrberg, Anne Haubo. 2016a. "Bitcoin, Gold and the Dollar–A Garch Volatility Analysis." *Finance Research Letters* 16: 85–92.

———. 2016b. "Hedging Capabilities of Bitcoin. Is It the Virtual Gold?" *Finance Research Letters* 16: 139–44.

Dyrssen, Hannah, Erik Ekstrom, and Johan Tysk. 2014. "Pricing Equations in Jump-to-Default Models." *International Journal of Theoretical and Applied Finance* 17 (03): 1450019.

Efron, Bradley, and Robert J Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC press.

Emmer, S, M Kratz, and D Tasche. 2015. "What Is the Best Risk Measure in Practice?" *Journal of Risk* 18: 31–60.

Engle, Robert. 2002. "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models." *Journal of Business & Economic Statistics* 20 (3): 339–50.

Engle, Robert F, and Kevin Sheppard. 2001. "Theoretical and Empirical Properties of Dynamic Conditional Correlation Multivariate Garch." National Bureau of Economic Research.

Fantazzini, Dean. 2008. "Dynamic Copula Modelling for Value at Risk." *Frontiers in Finance and Economics* 5 (2): 72–108.

———. 2009a. "The Effects of Misspecified Marginals and Copulas on Computing the Value at Risk:

A Monte Carlo Study." *Computational Statistics & Data Analysis* 53 (6): 2168–88.

———. 2009b. "Value at Risk for High-Dimensional Portfolios: A Dynamic Grouped T-Copula Approach." In *The Var Implementation Handbook*, edited by Greg Gregoriou, 253–82. McGraw-Hill.

Fantazzini, Dean. 2019. *Quantitative Finance with R and Cryptocurrencies.* Amazon KDP, ISBN-13: 978-1090685315.

Fantazzini, Dean, and Silvia Figini. 2008. "Default Forecasting for Small-Medium Enterprises: Does Heterogeneity Matter?" *International Journal of Risk Assessment and Management* 11 (1-2): 138–63.

———. 2009. "Random Survival Forests Models for Sme Credit Risk Measurement." *Methodology and Computing in Applied Probability* 11 (1): 29–45.

Fantazzini, Dean, and Mario Maggi. 2015. "Proposed Coal Power Plants and Coal-to-Liquids Plants in the Us: Which Ones Survive and Why?" *Energy Strategy Reviews* 7: 9–17.

Fantazzini, Dean, Maria Elena De Giuli, and Mario Maggi. 2008. "A New Approach for Firm Value and Default Probability Estimation Beyond Merton Models." *Computational Economics* 31 (2): 161–80.

Fantazzini, Dean, Erik Nigmatullin, Vera Sukhanovskaya, and Sergey Ivliev. 2016. "Everything You Always Wanted to Know About Bitcoin Modelling but Were Afraid to Ask. Part 1." *Applied Econometrics* 44: 5–24.

———. 2017. "Everything You Always Wanted to Know About Bitcoin Modelling but Were Afraid to Ask. Part 2." *Applied Econometrics* 45: 5–28.

Feder, Amir, Neil Gandal, JT Hamrick, Tyler Moore, and Marie Vasek. 2018. "The Rise and Fall of Cryptocurrencies." In *Proc. of the Workshop on the Economics of Information Security (Weis).*

Fernandez, Carmen, and Mark FJ Steel. 1998. "On Bayesian Modeling of Fat Tails and Skewness." *Journal of the American Statistical Association* 93 (441) Group: 359–71.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2016. *The Elements of Statistical Learning.* Vol. 1.

Fuertes, Ana-Maria, and Elena Kalotychou. 2006. "Early Warning Systems for Sovereign Debt Crises: The Role of Heterogeneity." *Computational Statistics & Data Analysis* 51 (2): 1420–41.

Gandal, Neil, JT Hamrick, Tyler Moore, and Tali Oberman. 2018. "Price Manipulation in the Bitcoin Ecosystem." *Journal of Monetary Economics* 95: 86–96.

Garcia, David, and Frank Schweitzer. 2015. "Social Signals and Algorithmic Trading of Bitcoin." *Royal Society Open Science* 2 (9): 150288.

Garcia, David, Claudio J Tessone, Pavlin Mavrodiev, and Nicolas Perony. 2014. "The Digital Traces of Bubbles: Feedback Cycles Between Socio-Economic Signals in the Bitcoin Economy." *Journal of the Royal Society Interface* 11 (99): 20140623.

Gerlach, Jan-Christian, Guilherme Demos, and Didier Sornette. 2018. "Dissection of Bitcoin's Mul-

tiscale Bubble History from January 2012 to February 2018." *arXiv Preprint arXiv:1804.06261.*

Giacomini, Raffaella, and Ivana Komunjer. 2005. "Evaluation and Combination of Conditional Quantile Forecasts." *Journal of Business & Economic Statistics* 23 (4): 416–31.

Gkillas, Konstantinos, and Paraskevi Katsiampa. 2018. "An Application of Extreme Value Theory to Cryptocurrencies." *Economics Letters* 164: 109–11.

Glaser, Florian, Kai Zimmermann, Martin Haferkorn, Moritz Weber, and Michael Siering. 2014. "Bitcoin-Asset or Currency? Revealing Users' Hidden Intentions."

Gneiting, Tilmann. 2011. "Making and Evaluating Point Forecasts." *Journal of the American Statistical Association* 106 (494): 746–62.

Goin, James E. 1982. "ROC Curve Estimation and Hypothesis Testing: Applications to Breast Cancer Detection." *Pattern Recognition* 15 (3): 263–69.

Gonzalez-Rivera, Gloria, Tae-Hwy Lee, and Santosh Mishra. 2004. "Forecasting Volatility: A Reality Check Based on Option Pricing, Utility Function, Value-at-Risk, and Predictive Likelihood." *International Journal of Forecasting* 20 (4): 629–45.

Griffin, John M, and Amin Shams. 2018. "Is Bitcoin Really Un-Tethered?" University of Texas at Austin.

Hanczar, Blaise, Jianping Hua, Chao Sima, John Weinstein, Michael Bittner, and Edward R Dougherty. 2010. "Small-Sample Precision of Roc-Related Estimates." *Bioinformatics* 26 (6): 822–30.

Hanley, James A, and Barbara J McNeil. 1982. "The Meaning and Use of the Area Under a Receiver Operating Characteristic (Roc) Curve." *Radiology* 143 (1): 29–36.

Hansen, Bruce E. 1994. "Autoregressive Conditional Density Estimation." *International Economic Review.* JSTOR, 705–30.

Hansen, Peter R, Asger Lunde, and James M Nason. 2011. "The Model Confidence Set." *Econometrica* 79 (2): 453–97.

Hartmann, Philipp. 2010. "Interaction of Market and Credit Risk." *Journal of Banking and Finance* 4 (34): 697–702.

Hayes, Adam. 2015. "A Cost of Production Model for Bitcoin." New School for Social Research, Department of Economics.

———. 2017. "Cryptocurrency Value Formation: An Empirical Study Leading to a Cost of Production Model for Valuing Bitcoin." *Telematics and Informatics* 34 (7): 1308–21.

Ho, Tin Kam. 1995. "Random Decision Forests." In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, 1:278–82. IEEE.

Hu, Albert, Christine A Parlour, and Uday Rajan. 2018. "Cryptocurrencies: Stylized Facts on a

New Investible Instrument." Haas School of Business, UC Berkeley.

Hwang, Soosung, and Pedro L Valls Pereira. 2006. "Small Sample Properties of Garch Estimates and Persistence." *The European Journal of Finance* 12 (6-7): 473–94.

Joe, Harry. 1997. *Multivariate Models and Multivariate Dependence Concepts*. CRC Press.

Jorion, Philippe. 2006. "Value at Risk: The New Benchmark for Managing Financial Risk." *NY: McGraw-Hill Professional*.

———. 2007. *Financial Risk Manager Handbook*. Vol. 406. Wiley.

Katsiampa, Paraskevi. 2017. "Volatility Estimation for Bitcoin: A Comparison of Garch Models." *Economics Letters* 158: 3–6.

Kohavi, R, and F Provost. 1998. "Glossary of Terms." *Machine Learning* 30: 271–74.

Kostovetsky, Leonard, and Hugo Benedetti. 2018. "Digital Tulips? Returns to Investors in Initial Coin Offerings." Boston College working paper.

Kratz, Marie, Yen H Lok, and Alexander J McNeil. 2018. "Multinomial Var Backtests: A Simple Implicit Approach to Backtesting Expected Shortfall." *Journal of Banking & Finance* 88: 393–407.

Kristoufek, Ladislav. 2013. "Can Google Trends Search Queries Contribute to Risk Diversification?" *Scientific Reports* 3. Nature Publishing Group.

Kupiec, Paul. 1995. "Techniques for Verifying the Accuracy of Risk Measurement Models." *The Journal of Derivatives*, 3(2).

Lansky, Jan. 2018. "Possible State Approaches to Cryptocurrencies." *Journal of Systems Integration* 9 (1): 19–31.

Li, Lili, Jun Yang, and Xin Zou. 2016. "A Study of Credit Risk of Chinese Listed Companies: ZPP Versus Kmv." *Applied Economics* 48 (29): 2697–2710.

Linetsky, Vadim. 2006. "Pricing Equity Derivatives Subject to Bankruptcy." *Mathematical Finance* 16 (2): 255–82.

Liu, Ruiping, Zhichao Shao, Guodong Wei, and Wei Wang. 2017. "GARCH Model with Fat-Tailed Distributions and Bitcoin Exchange Rate Returns." *Journal of Accounting, Business and Finance Research* 1 (1): 71–75.

MacDonell, Alec. 2014. "Popping the Bitcoin Bubble: An Application of Log-Periodic Power Law Modeling to Digital Currency." *University of Notre Dame Working Paper*.

McNeil, Alexander J, and Rudiger Frey. 2000. "Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: An Extreme Value Approach." *Journal of Empirical Finance* 7 (3-4). Elsevier: 271–300.

McNeil, Alexander J, Rudiger Frey, and Paul Embrechts. 2015. *Quantitative Risk Management:*

Concepts, Techniques and Tools. Princeton U.P.

Mensi, Walid, Khamis Hamed Al-Yahyaee, and Sang Hoon Kang. 2018. "Structural Breaks and Double Long Memory of Cryptocurrency Prices: A Comparative Analysis from Bitcoin and Ethereum." *Finance Research Letters.*

Mester, Loretta J. 1997. "What's the Point of Credit Scoring?" *Business Review* 3 (Sep/Oct): 3–16.

Metz, Charles E, and Helen B Kronman. 1980. "Statistical Significance Tests for Binormal Roc Curves." *Journal of Mathematical Psychology* 22 (3): 218–43.

Mitsuru, I, Y Kitamura, and M Tsutomu. 2014. "Is Bitcoin the Only Cryptocurrency in the Town?" Discussion Paper Series A No.602, Institute of Economic Research Hitotsubashi University, Tokyo.

Moore, Tyler, and Nicolas Christin. 2013. "Beware the Middleman: Empirical Analysis of Bitcoin-Exchange Risk." In *International Conference on Financial Cryptography and Data Security*, 25–33. Springer.

Moore, Tyler, Nicolas Christin, and Janos Szurdi. 2018. "Revisiting the Risks of Bitcoin Currency Exchange Closure." *ACM Transactions on Internet Technology* 18.

Muller, Fernanda Maria, and Marcelo Brutti Righi. 2018. "Numerical Comparison of Multivariate Models to Forecasting Risk Measures." *Risk Management* 20 (1): 29–50.

Naimy, Viviane Y, and Marianne R Hayek. 2018. "Modelling and Predicting the Bitcoin Volatility Using Garch Models." *International Journal of Mathematical Modelling and Numerical Optimisation* 8 (3): 197–215.

Nakamoto, Satoshi. 2008. "Bitcoin: A Peer-to-Peer Electronic Cash System."

Narayanan, Arvind, Joseph Bonneau, Edward Felten, Andrew Miller, and Steven Goldfeder. 2016. *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction.* Princeton University Press.

Nelsen, Roger B. 1999. *An Introduction to Copulas.* Vol. 139. Springer-Verlag, New York.

Osterrieder, Joerg. 2016. "The Statistics of Bitcoin and Cryptocurrencies." Advanced Risk; Portfolio Management Paper.

Osterrieder, Joerg, and Julian Lorenz. 2017. "A Statistical Risk Assessment of Bitcoin and Its Extreme Tail Behavior." *Annals of Financial Economics* 12 (01): 1750003.

Patton, Andrew. 2013. "Copula Methods for Forecasting Multivariate Time Series." In *Handbook of Economic Forecasting*, 2:899–960. Elsevier.

Patton, Andrew J. 2006a. "Estimation of Multivariate Models for Time Series of Possibly Different Lengths." *Journal of Applied Econometrics* 21 (2): 147–73.

———. 2006b. "Modelling Asymmetric Exchange Rate Dependence." *International Economic Review* 47 (2): 527–56.

———. 2009. "Copula–based Models for Financial Time Series." In *Handbook of Financial Time*

Series, 767–85.

Pesaran, M Hashem, and Allan Timmermann. 2007. "Selection of Estimation Window in the Presence of Breaks." *Journal of Econometrics* 137 (1): 134–61.

Pichl, Lukas, and Taisei Kaizoji. 2017. "Volatility Analysis of Bitcoin." *Quantitative Finance and Economics* 1: 474–85.

Reilly, Frank K, and Keith C Brown. 2002. *Investment Analysis and Portfolio Management.* Cengage Learning.

Rodriguez, Arnulfo, and Pedro N Rodriguez. 2006. "Understanding and Predicting Sovereign Debt Rescheduling: A Comparison of the Areas Under Receiver Operating Characteristic Curves." *Journal of Forecasting* 25 (7): 459–79.

Sammut, Claude, and Geoffrey I Webb. 2011. *Encyclopedia of Machine Learning* Science & Business Media.

Satchell, Stephen, and John Knight. 2011. *Forecasting Volatility in the Financial Markets.*

Sharpe, William F. 1992. "Asset Allocation: Management Style and Performance Measurement." *Journal of Portfolio Management* 18 (2): 7–19.

Sid. 2018. "How Peng Coin Will Surge 8-12x These Coming Weeks." Medium.

Sklar, M. 1959. "Fonctions de Repartition an Dimensions et Leurs Marges." *Publ. Inst. Statist. Univ. Paris* 8: 229–31.

Smith, Chris, and Mark Koning. 2017. *Decision Trees and Random Forests: A Visual Introduction for Beginners.* Independently published.

Stavroyiannis, Stavros. 2018. "Value-at-Risk and Related Measures for the Bitcoin." *The Journal of Risk Finance* 19 (2). Emerald Publishing Limited: 127–36.

Su, En-Der, and Shih-Ming Huang. 2010. "Comparing Firm Failure Predictions Between Logit, Kmv, and Zpp Models: Evidence from Taiwan's Electronics Industry." *Asia-Pacific Financial Markets* 17 (3). Springer: 209–39.

Su, Lujing, and Marc O Rieger. 2009. "How Likely Is It to Hit a Barrier? Theoretical and Empirical Estimates." *Technical Report, Working Paper No. 594, National Centre of Competence in Research, Financial Valuation and Risk Management.*

Thies, Sven, and Peter Molnar. 2018. "Bayesian Change Point Analysis of Bitcoin Returns." *Finance Research Letters* 27: 223–27.

Trucios, Carlos. 2018. "Forecasting Bitcoin Risk Measures: A Robust Approach."

Tse, Yiu K, and Albert K C Tsui. 2002. "A Multivariate Generalized Autoregressive Conditional Heteroscedasticity Model with Time-Varying Correlations." *Journal of Business & Economic Statistics*

20 (3): 351–62.

Weiss, Gregor NF. 2013. "Copula-Garch Versus Dynamic Conditional Correlation: An Empirical Study on Var and Es Forecasting Accuracy." *Review of Quantitative Finance and Accounting* 41 (2): 179–202.

Weiss, Gregor NF. 2011. "Are Copula-Gof-Tests of Any Practical Use? Empirical Evidence for Stocks, Commodities and Fx Futures." *The Quarterly Review of Economics and Finance* 51 (2): 173–88.

Woo, D., I. Gordon, and V. Iaralov. 2013. "Bitcoin: A First Assessment." *FX and Rates* December 2013. Bank of America Merrill Lynch.

Yelowitz, Aaron, and Matthew Wilson. 2015. "Characteristics of Bitcoin Users: An Analysis of Google Search Data." *Applied Economics Letters* 22 (13): 1030–6.

Appendix A: a review of the history and the financial literature devoted to cryptocurrencies

Brief historical overview: Bitcoin & sons

Bitcoin was not the first cryptocurrency proposed in the IT literature, see e.g. the works by Chaum (1983), Chaum and Brands (1997) and Back (2002). The idea of Bitcoin was presented in a paper published in 2008 by a group of anonymous authors under the pseudonym of Satoshi Nakamoto (Nakamoto (2008)). Since then, Bitcoin has become the most popular online decentralized currency, which allows users to make transactions without any third party by using a peer-to-peer protocol. The system is based on cryptography algorithms which provide security for all operations, see Antonopoulos (2014), Narayanan et al. (2016) for more details. The Bitcoin market capitalization was close to $ 70 bn at the end of 2018 and represented almost 50% of all cryptocurrencies total capitalization. The price for a single bitcoin decreased to approximately $ 4000, after reaching a top of almost $ 20000 at the end of 2017. Most cryptocurrencies are *hard forks* of the Bitcoin protocol, that is they introduced new rules to create blocks which are not considered valid by the older (Bitcoin) software. A description of the history of bitcoin (and cryptocurrencies in general) can be found in Burniske and Tatar (2017).

A common way to raise funds to create a new coin, app, or service with cryptocurrencies is to launch an *Initial Coin Offering* (ICO). An ICO is a type of crowdfunding, where funds are collected by selling a fixed number of new coins to investors. It is similar to an Initial Public Offering (IPO) of a company, but there are some important differences: ICOs may fall outside current regulations and can be prone to scams and securities law violations. Moreover, holding coins not necessarily gives dividends, but services and goods manufactured by the company. According to the CoinDesk State of Blockchain Q1 2018[18], the number of ICOs is still increasing (number of ICOs in 2017 was twice larger than in 2016) and its funding attained $ 12 bn. Only in the first quarter of 2018, there were 202 ICOs with $ 6.3 bn of total funding. However, investing in cryptocurrencies can be an extremely risky process and the final return depends on the strategy adopted: Kostovetsky and Benedetti (2018) shows that approximately 56 percent of crypto startups that raise money through ICOs die within four months of their initial coin offerings. However, they also show that the representative ICO investor earns 82% and the safest strategy is to acquire coins in an ICO and then sell them on the first day -if individual investors can participate in ICOs-, or to sell them in the first six months, otherwise.

[18] https://www.coindesk.com/research/state-blockchain-q1-2018

Literature review

The Bitcoin phenomenon has attracted significant attention in the academic literature with regard to its fundamental value (Woo, Gordon, and Iaralov (2013), Garcia et al. (2014), Hayes (2015), Hayes (2017)), price dynamics (Buchholz et al. (2012), Kristoufek (2013), Garcia et al. (2014), Garcia and Schweitzer (2015), Glaser et al. (2014), Bouoiyour and Selmi (2015), Bouoiyour, Selmi, and Tiwari (2015), Ciaian, Rajcaniova, and Kancs (2016), Bouri, Azzi, and Dyhrberg (2017)), bubble modelling (MacDonell (2014), Cheah and Fry (2015), Gerlach, Demos, and Sornette (2018)), price discovery (Brandvold et al. (2015)), and more recently about univariate volatility modelling (Dyhrberg (2016a), Dyhrberg (2016b)), Balcilar et al. (2017), Chu et al. (2017), Katsiampa (2017), Liu et al. (2017), Pichl and Kaizoji (2017), Naimy and Hayek (2018) and Catania, Grassi, and Ravazzolo (2018). See Fantazzini et al. (2016) and Fantazzini et al. (2017) for a large survey of the econometric and mathematical tools which have been proposed so far to model the bitcoin price and several related issues, highlighting advantages and limits.

With regards to risk management for cryptocurrencies, the number of works is much more limited: Chu et al. (2017) compared twelve GARCH models with seven popular cryptocurrencies, and their fits were assessed in terms of five in-sample criteria and out-of-sample Value at Risk performances. Chan et al. (2017) analyzed the (unconditional) statistical properties of seven cryptocurrencies, while Osterrieder and Lorenz (2017) examined the tail behavior of bitcoin returns using extreme value distributions but no backtesting was performed. Stavroyiannis (2018) examined a set of market risk measures for the BTC and compared these measures with the SP500 index, the Brent crude oil spot price and the gold spot price by performing a backtesting analysis. Gkillas and Katsiampa (2018) studied the tail behavior of the returns of five major cryptocurrencies by using again extreme value analysis and computing the Value-at-Risk and Expected Shortfall, but no backtesting analysis was implemented. Trucios (2018) compared the one-step-ahead volatility forecast of Bitcoin using several GARCH-type models and also evaluated the performance of several procedures when estimating the Value-at-Risk. As it is possible to notice, all these studies dealt only with univariate models, focused almost exclusively on the Value-at-Risk and volatility forecasting, while only three works performed backtesting analysis.

In general, standard models for market risk tend to work poorly with cryptocurrencies due to the frequent presence of structural breaks, see Bouri et al. (2016), Fantazzini et al. (2016), Fantazzini et al. (2017), Mensi, Al-Yahyaee, and Kang (2018) and Thies and Molnar (2018). To make matters worse, price manipulations and market frauds caused by the lack of financial oversight (Gandal et al. (2018), Griffin and Shams (2018)) and the fact that cryptocurrencies are still mainly used for speculative purposes, make financial bubbles a recurring phenomenon, see Corbet, Lucey, and Yarovaya (2018), Cheah and Fry (2015) and Gerlach, Demos, and Sornette (2018). A potential solution could be to use complex model

specifications able to accommodate structural breaks and extreme price volatility, but this would come at the cost of lower computational tractability and potential model over-fitting. Moreover, this solution would become quickly unfeasible in the multivariate case, due to the well-known curse of dimensionality. Finally, we remark that credit risk modelling and the implications of having invested in dead coins have not been considered so far.

Appendix B: Multivariate time series models

We employed two multivariate models for simultaneously computing the market and credit risk measures of a portfolio of cryptocurrencies: the VAR-DCC and the VAR-Copula-GARCH models. The DCC model was originally proposed by Engle (2002) and Tse and Tsui (2002), while copula-GARCH models were discussed in Cherubini, Luciano, and Vecchiato (2004), A. J. Patton (2006a), A. J. Patton (2006b), Fantazzini (2008) and Fantazzini (2009b). We provide below a brief review of these two approaches, while we refer the interested reader to Bauwens, Hafner, and Laurent (2012) for a more detailed treatment at the textbook level.

These two approaches share similar building blocks for the conditional mean and the conditional variance: for the mean, a Vector Auto-Regression model (VAR) is used, while for the variance a set of GARCH models was employed. Let $\mathbf{Y}_t$ be a vector stochastic process of dimension $n \times 1$, then a conditional model for $\mathbf{Y}_t$ can be expressed as follows:

$$\mathbf{Y}_t = \boldsymbol{\mu}_t + \mathbf{D}_t \mathbf{z}_t$$

where $\boldsymbol{\mu}_t$ is a vector of conditional means, $\mathbf{D}_t = diag(\sigma_{1,t}, \ldots, \sigma_{n,t})$ a diagonal matrix of conditional standard deviations, while $\mathbf{z}_t$ is a vector of standardized errors with a conditional multivariate distribution $H_t(z_{1,t}, \ldots, z_{n,t}; \boldsymbol{\theta})$ and parameter vector $\boldsymbol{\theta}$.

The conditional means are modelled with a VAR(p) model,

$$\boldsymbol{\mu}_t = \mathbf{a}_0 + \sum_{m=1}^{p} \mathbf{A}_m \mathbf{Y}_{t-m}$$

while the conditional variances with GARCH(p,q) models,

$$\sigma_{i,t}^2 = \omega_i + \sum_{m=1}^{p} \alpha_{i,m} (\sigma_{i,t-m} z_{i,t-m})^2 + \sum_{k=1}^{q} \beta_{i,k} \sigma_{i,t-k}^2$$

Other univariate GARCH models, like the Exponential-GARCH, the Threshold-GARCH, etc. can be used. Where the VAR-DCC and the VAR-Copula-GARCH models differ is how they specify the condi-

tional joint distribution H_t.

Dynamic Conditional Correlation (DCC) models

The DCC model by Engle (2002) assumes that H_t is a multivariate Normal or Student's t distribution with correlation matrix $\mathbf{R}_t$,

$$\mathbf{R}_t = (diag\mathbf{Q}_t)^{-1/2}\mathbf{Q}_t(diag\mathbf{Q}_t)^{-1/2} \tag{5}$$

where the $n \times n$ symmetric positive definite matrix $\mathbf{Q}_t$ is given by:

$$\mathbf{Q}_t = \left(1 - \sum_{l=1}^{L}\theta_{1,l} - \sum_{s=1}^{S}\theta_{2,s}\right)\bar{\mathbf{Q}} + \sum_{l=1}^{L}\theta_{1,l}\mathbf{z}_{t-l}\mathbf{z}'_{t-l} + \sum_{s=1}^{S}\theta_{2,s}\mathbf{Q}_{t-s}$$

where $\bar{\mathbf{Q}}$ is the $n \times n$ unconditional variance/covariance matrix of $\mathbf{z}_t$, $\theta_{1,l}(\geq 0)$ and $\theta_{2,s}(\geq 0)$ are scalar parameters satisfying $\sum_{l=1}^{L}\theta_{1,l} + \sum_{s=1}^{S}\theta_{2,s} < 1$ to have $\mathbf{Q}_t > 0$ and $\mathbf{R}_t > 0$. $\mathbf{Q}_t$ is the covariance matrix of $\mathbf{z}_t$, so that $q_{ii,t}$ is not equal to 1 by construction and it is subsequently transformed into a correlation matrix by (5).

In the multivariate normal case, the total likelihood can be decomposed in two parts, so that the models for the conditional means and variances can be estimated in a first stage, while the parameters of the conditional correlation are estimated in a second stage using the parameters from the first step. Instead, in case of a multivariate student's t distribution, the estimation should be performed either in one step (so that the shape parameter is jointly estimated for all volatility models), or in two stages: in this latter case, the first step is based on a Gaussian quasi-maximum likelihood (QML) estimator for the conditional means and variances, while the shape parameter is estimated in a second step, as suggested by Bauwens and Laurent (2005).

The DCC model is one of the most used models in applied finance, thanks to its computational flexibility that allows this model to be computed also with portfolios consisting of up to 100 assets (Engle and Sheppard (2001), Bauwens, Hafner, and Laurent (2012)). Despite known problems (Aielli (2013), Caporin and McAleer (2013)), it still remains an important benchmark when multivariate volatility modelling is of concern, see Bauwens, Laurent, and Rombouts (2006), Satchell and Knight (2011), Caporin and McAleer (2014) and Bali and Zhou (2016).

Copula-VAR-GARCH models

Copula theory provides an easy way to deal with the (usually) complex multivariate modelling. Using the so-called Sklar (1959) theorem, a joint distribution can be factored into the marginals and a dependence

function called a *copula*. The joint distribution H_t can be expressed as follows:

$$\mathbf{z}_t \sim H_t(z_{1,t}, \ldots, z_{n,t}; \boldsymbol{\theta}) \equiv C_t(F_{1,t}(z_1; \delta_1), \ldots, F_{n,t}(z_{n,t}; \delta_n); \boldsymbol{\gamma}) \tag{6}$$

which means that the joint distribution H_t of a vector of standardized errors $\mathbf{z}_t$ is the copula $C_t(\cdot; \boldsymbol{\gamma})$ of the cumulative distribution functions of the innovation marginals $F_{1,t}(z_1; \delta_1), \ldots, F_{n,t}(z_{n,t}, ; \delta_n)$, where $\boldsymbol{\gamma}, \delta_1, \ldots, \delta_n$ are the copula and marginal parameters, respectively. For more details about copulas, we refer the interested reader to the textbooks by Joe (1997) and Nelsen (1999), while Cherubini, Luciano, and Vecchiato (2004) provide a detailed discussion of copula techniques for financial applications. It follows from (6) that the log-likelihood function for the joint conditional distribution $H_t(\cdot, \boldsymbol{\theta})$ is given by

$$l(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log \Big(c\big(F_{1,t}(z_1; \delta_1), \ldots, F_{n,t}(z_{n,t}; \delta_n); \boldsymbol{\gamma}\big) \Big) + \sum_{t=1}^{T} \sum_{i=1}^{n} \log f_i(z_{i,t}; \delta_i)$$

where c is the copula density function, whereas f_i are the marginal densities. The maximization of the previous log-likelihood with respect to the parameters $(\boldsymbol{\gamma}, \delta_1, \ldots, \delta_n)$ can be made in 1 step, or in several steps by partitioning of the parameter vector into separate parameters for each margin and the parameters for the copula. This multi-step procedure is known as the method of Inference Functions for Margins (IFM), see Joe (1997) for details.

It is rather straightforward to show that the previous DCC model can be represented as a special case within a more general copula framework,

$$\begin{aligned} \mathbf{Y}_t &= \boldsymbol{\mu}_t + \mathbf{D}_t \mathbf{z}_t, \quad \text{where} \quad \mathbf{z}_t \sim H_t(z_{1,t}, \ldots, z_{n,t}; \boldsymbol{\theta}), \quad \text{and} \\ \mathbf{z}_t &\sim H_t \equiv C_t^{Normal}(F_{1,t}^{Normal}(z_1; \delta_1), \ldots, F_{n,t}^{Normal}(z_{n,t}; \delta_n); \mathbf{R}_t) \end{aligned}$$

see e.g. A. J. Patton (2006a), A. J. Patton (2006b), Fantazzini (2008) and Fantazzini (2009b) for more details.

The copula approach allows us to consider more general cases than a multivariate normal DCC model. For example, if we consider skewed-t distributions for the marginals, like those proposed by Hansen (1994) or Fernandez and Steel (1998), then a multivariate model allowing for marginal skewness, kurtosis and normal dependence can be expressed as follows:

$$\begin{aligned} \mathbf{Y}_t &= \boldsymbol{\mu}_t + \mathbf{D}_t \mathbf{z}_t \\ \mathbf{z}_t &\sim H_t \equiv C_t^{Normal}(F_{1,t}^{Skewed-t}(z_{1,t}; \delta_1), \ldots, F_{n,t}^{Skewed-t}(z_{n,t}; \delta_n); \mathbf{R}_t) \end{aligned}$$

where $F_{i,t}^{Skewed-t}$ is the cumulative distribution function of the marginal Skewed-t, and $\mathbf{R}_t$ can be made constant or time-varying, as in the constant conditional correlation (CCC) model or in the DCC model, respectively.

If we suppose that our assets have symmetric tail dependence, we can use a Student's t copula, instead,

$$\begin{aligned}
\mathbf{Y}_t &= \boldsymbol{\mu}_t + \mathbf{D}_t \mathbf{z}_t \\
\mathbf{z}_t &\sim H_t \equiv C_t^{Student'sT}(F_{1,t}^{Skewed-t}(z_1;\delta_1), \ldots, F_{n,t}^{Skewed-t}(z_{n,t};\delta_n); \mathbf{R}_t, \nu)
\end{aligned}$$

where ν denotes the degrees of freedom of the t-copula.