Lomonosov Moscow State University
Moscow School of Economics

as a manuscript

Dean Fantazzini

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

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

JEL: C1,C2,C3,C5,E2,E4,G3,I3,L6,Q4

Moscow - 2020

Contents

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.

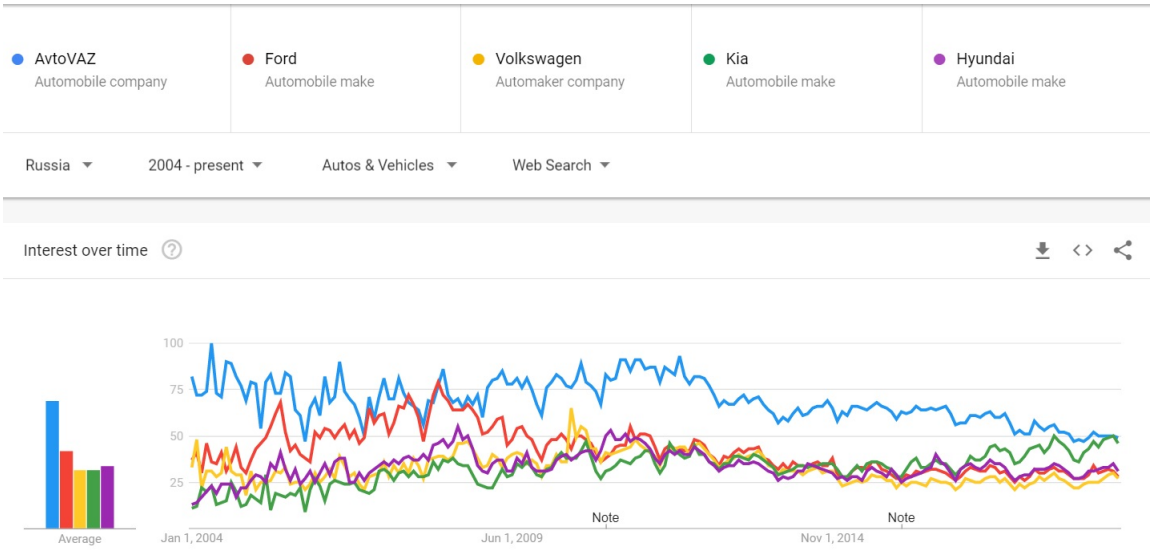

Figure 1: Google Trends search data for Russia - Autos and Vehicles category.

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 $X_1, \dots X_n$ 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 $X_1 \dots X_n$ 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 $i = 1, 2, \dots, n$.
3. $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$, 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, $\mathbf{D}_t = diag(\sigma_{1,t}, \ldots, \sigma_{n,t})$ 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. If we use copula theory, the joint distribution H_t of the vector of standardized errors $\mathbf{z}_t$ can be expressed as follows:

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

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, Dalla Valle and Giudici (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

[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.

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 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***

3 Main results: detailed discussion

This section presents more details of the main results of my dissertation. More specifically, the results of my 20 publications are grouped according to thematic areas and are presented in general terms to make the discussion as self-contained as possible. The full details and the proofs can be found in the original publications.

3.1 Copula theory

3.1.1 Operational Risks

Fantazzini, Dalla Valle and Giudici (2008) proposed copulas for modeling high dimensional operational risks in a more flexible way able to takes (partial) dependence into account. In the first step, they employed the so-called actuarial approach to model the operational risk marginal losses. This approach employs two types of distributions: one distribution to describe the frequency of the risky events and a second one to describe 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 period, operational losses can be defined as a sum S_i of the random number n_i of the losses X_{ij}:

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

The actuarial model assumes that the probability distribution of S_i can 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 that the losses are random variables, independent and identically distributed (i.i.d.), and the distribution of the frequency n_i is independent of the distribution of the severity X_{ij}.

In a second step, Fantazzini, Dalla Valle and Giudici (2008) used the Sklar's theorem (1959) to show that the joint distribution H of a vector of losses S_i (with $i = 1 \ldots R$) is the copula of the cumulative distribution functions of the losses' marginals :

$$H(S_1, \ldots, S_R) = C(F(S_1), \ldots, F(S_R))$$

Finally they proposed the following procedure to compute the required total capital for operational risk:

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 by using simulation methods: first, simulate a multivariate random vector from a specified copula C with marginals uniformly distributed in the unit interval [0,1]. Subsequently, 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, use the *generalized inverse* of the functions F_i, given by $F_i^{-1}(u) = inf\{x : F_i(x) \geq u\}$. Then, sum the losses S_i for each intersection i, obtaining a global loss scenario. Finally, repeat the previous steps a great number of times, and compute the desired risk measure (the Value-at-Risk or the expected shortfall). See Fantazzini, Dalla Valle and Giudici (2008) for the full detailed procedure.

Fantazzini, Dalla Valle and Giudici (2008) compared different marginals distributions to model the losses frequency and severity, and different copula models to estimate the VaR and the ES for different confidence levels. Their empirical analysis showed 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, they noted that the Poisson is much more easier to estimate, especially with small samples. Fantazzini, Dalla Valle and Giudici (2008) further showed that, 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.

3.1.2 Asymptotic results for semi-parametric estimators

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. I now briefly review this semi-parametric method since it constitutes the building block for the following analysis.

Let consider a multivariate random sample represented by the time series $X = (x_{1t} \dots x_{nt})$, $t = 1 \dots T$, where n stands for the number of underlying assets included and T represents the number of observations available. Let f_h be the density of the joint distribution H:

$$f_h(x_i; \alpha_1, ..., \cdot\alpha_n, \gamma) = c(F_1(x_{1;}\alpha_1), ..., F_1(x_{n;}\alpha_n); \gamma) \cdot \prod_{i=1}^{n} f_i(x_i; \alpha_i) \tag{4}$$

where f_i is the univariate density of the marginal distribution F_i and c is the density of the copula. We suppose to have a set of T empirical data of n financial asset log-returns, for example, and let $\theta = (\alpha_1, \dots, \alpha_{n;}\gamma)$ be the parameter vector to estimate, where α_i, i = 1, ...,n is the vector of parameters of the marginal distribution F_i and γ is the vector of the copula parameters. The estimation process is performed in two steps:

Definition 3.1 (CML copula estimation)**.** *1. Transform the dataset* $(x_{1t}, x_{2t} \dots x_{nt}), t = 1 \dots T$ *into approximately uniform variates* $(\hat{u}_{1t}, \hat{u}_{2t}, \dots, \hat{u}_{nt})$ *using the empirical distributions* $F_{iT}(\cdot)$ *defined as follows:*

$$\hat{u}_{it} = F_{iT}(\cdot) = \frac{1}{T+1} \sum_{t=1}^{T} \mathbb{1}_{\{x_{it} \leq x_i)}, \ i = 1 \dots n \tag{5}$$

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

2. *Estimate the copula parameters by maximizing the log-likelihood:*

$$\hat{\gamma}_{CML} = \arg\max \sum_{t=1}^{T} \log(c(F_{1T}(x_{1t}), \ldots, F_{nT}(x_{nt})); \gamma) \tag{6}$$

Genest et al. (1995) showed in Proposition A.1 that under certain regularity conditions, the semiparametric estimator $\hat{\gamma}_{CML}$ is consistent and asymptotically normally distributed.

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. 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 a 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.

More specifically, Bouye et al. (2001), Fantazzini (2010b) and McNeil et al. (2015), have suggested the following estimation procedure[3]:

Definition 3.2 (Three-stage KME - CML copula estimation)**.**

1. *Transform the dataset* $(x_{1t}, x_{2t}, \ldots, x_{nt})$ *into normalized ranks* $(F_{1T}(x_{1t}),\ F_{2T}(x_{2t}),\ \ldots, F_{nT}(x_{nt}))$*, using the empirical distribution function, and which are approximately uniform variates.*

2. *Collect all pairwise estimates of the sample Kendall's tau given by*

$$\binom{T}{2}^{-1} \sum_{1 \leq t < s \leq T} sign\left((x_{1,t} - \tilde{x}_{1,s})(x_{2,t} - \tilde{x}_{2,s})\right) \tag{7}$$

in an empirical Kendall's tau matrix $\hat{R}^{\tau}$ *defined by* $\hat{R}^{\tau}_{jk} = \tau[F_{jT}(X_j),\ F_{kT}(X_k)]$*, 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* $q = n \cdot (n-1)/2$ *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,* $\hat{\mathbf{\Sigma}}$ *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:*

$$\hat{\nu}_{CML} = \arg\max \sum_{t=1}^{T} \log c_{T-copula}(F_{1T}(x_{1t}), \ldots, F_{nT}(x_{nT}); \hat{\Sigma}\ , \nu) \tag{8}$$

[3]We follow here the classical notation so that n refers to the cross-sectional dimension, while T to the temporal dimensional. We remark that in Fantazzini (2010b) -which I follow here closely-, they are instead denoted as d and n, respectively.

The second step in the previous definition 3.2 corresponds to a method-of-moments estimation based on q moments and Kendall tau rank correlations estimated with empirical distribution functions (also known as Kendall's tau moment estimator, KME): we stress again that the estimation of the margins is not necessary to estimate the correlation matrix, since the Kendall's tau is unaffected by the rank transformation. We can therefore build a $q \times 1$ moments vector ψ for the parameter vector $\theta_0 = [\rho_1, \ldots, \rho_q]'$ as reported below:

$$\psi\left(F_1(X_1), \ldots, F_n(X_n); \theta_0\right) = \begin{pmatrix} \mathrm{E}\left[\psi_1\left(F_1(X_1), F_2(X_2);\ \rho_1\right)\right] \\ \vdots \\ \mathrm{E}\left[\psi_q\left(F_{n-1}(X_{n-1}), F_n(X_n);\ \rho_q\right)\right] \end{pmatrix} = 0. \tag{9}$$

Then, *Fantazzini (2010b) proved the following theorems*:

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

(i) the parameter space Θ is a compact subset of $\mathbb{R}^q$,

(ii) the q-variate moment vector $\psi\left(F_1(X_1), \ldots, F_n(X_n); \theta_0\right)$ is continuous in θ_0 for all X_j,

(iii) $\psi\left(F_1(X_1), \ldots, F_n(X_n); \theta\right)$ is measurable in X_j for all θ in Θ,

(iv) $\mathrm{E}\left[\psi\left(F_1(X_1), \ldots, F_n(X_n); \theta\right)\right] \neq \mathbf{0}$ for all $\theta \neq \theta_0$ in Θ,

(v) $\mathrm{E}\left[\sup_{\theta \in \Theta} \|\psi\left(F_1(X_1), \ldots, F_n(X_n); \theta\right)\|\right] < \infty$,

Then $\hat{\theta} \xrightarrow{p} \theta_0$ as $T \to \infty$.

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 $\hat{\nu}_{CML} \xrightarrow{p} \nu_0$ as $T \to \infty$.*

The asymptotic normality is not straightforward, since we use a 3-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 3^{rd} 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:

$$\frac{\partial l(\cdot; \nu)}{\partial \nu} = \sum_{t=1}^{T} l_\nu \left(F_{1T}(x_{1t}), \ldots, F_{nT}(x_{nt}); \hat{\mathbf{\Sigma}}\ , \hat{\nu}\right) = 0. \tag{10}$$

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

$$\frac{1}{T}\sum_{t=1}^{T} l_\nu \left(F_{1T}(x_{1t}), \ldots, F_{nT}(x_{nt}); \hat{\mathbf{\Sigma}}\ , \hat{\nu}\right) = \frac{1}{T}\sum_{t=1}^{T} \psi_\nu (F_{1T}(x_{1t}), \ldots, F_{nT}(x_{nt}); \hat{\mathbf{\Sigma}}\ , \hat{\nu}) = 0$$

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. I 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 $i = 1, \ldots, T$, is the solution to the problem

$$\frac{1}{T}\sum_{t=1}^{T} \psi(Z_t, \hat{\theta}) = E[\psi(Z_t, \theta_0)] = 0$$

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

$$\Psi_{\ KME-CML}\left(F_{1T}(x_{1t}), \dots, F_{nT}(x_{nt}); \hat{\boldsymbol{\Xi}}\right) = \\ = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} \psi_1\left(F_{1T}(x_{i1}), F_{2T}(x_{i2});\ \hat{\rho}_1\right) \\ \vdots \\ \frac{1}{T}\sum_{t=1}^{T} \psi_q\left(F_{(n-1),T}(x_{(n-1)t}), F_{nT}(x_{nt});\ \hat{\rho}_q\right) \\ \frac{1}{T}\sum_{t=1}^{T} \psi_\nu\left(F_{1T}(x_{1t}), \dots, F_{nT}(x_{nt}); \hat{\boldsymbol{\Sigma}}, \hat{\nu}\right) \end{pmatrix} = 0$$

Let 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):

$$\Delta_0 = \begin{pmatrix} \psi_1\left(F_1(X_1), F_2(X_2);\ \rho_1\right) \\ \vdots \\ \psi_q\left(F_{n-1}(X_{n-1}), F_n(X_n);\ \rho_q\right) \\ \psi_\nu\left(F_1(X_1), \dots, F_n(X_n); \boldsymbol{\Sigma_0}, \nu_0\right) + \sum_{j=1}^{n} W_{j,\nu}(X_j) \end{pmatrix} = 0 \tag{11}$$

$$\boldsymbol{\Upsilon}_0 \equiv var\left[\boldsymbol{\Delta}_0\right] = \mathrm{E}\left[\boldsymbol{\Delta}_{\ KME-CML}\, \boldsymbol{\Delta}_{\ KME-CML}{}'\right] \tag{12}$$

where

$$W_{j,\nu}(X_j) \quad = \quad \int \mathbb{1}_{F_j(X_j)\leq u_j}\ \frac{\partial^2}{\partial\nu\partial u_j} \log c(u_1, \dots, u_n) dC(u_1, \dots, u_n) \tag{13}$$

Note that the population moments used to estimate the correlations are not affected by the marginals empirical distribution functions, since the Kendall's tau is invariant under strictly increasing marginal transformations. Then, *Fantazzini (2010b) proved this theorem*:

Theorem 3.3 (Asymptotic Distribution 3-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:*

$$\sqrt{T}(\hat{\boldsymbol{\Xi}} - \boldsymbol{\Xi}_0) \xrightarrow{d} \mathrm{N}\left(0, \mathrm{E}\left[\frac{\partial \Psi_{\ KME-CML}}{\partial \boldsymbol{\Xi}'}\right]^{-1} \boldsymbol{\Upsilon}_0\, \mathrm{E}\left[\frac{\partial \Psi_{\ KME-CML}}{\partial \boldsymbol{\Xi}'}\right]^{-1\prime}\right) \tag{14}$$

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. *Fantazzini (2010b) also proved this theorem*:

Theorem 3.4. (Asymptotic Distribution 3-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 (14).*

3.1.3 Finite-sample properties and computational aspects

Since the previous properties of the proposed semiparametric method are asymptotic, *a large-scale simulation study was carried out in Fantazzini (2010b) to examine its finite-sample behavior*. He found out that 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.

Fantazzini (2010b) 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.

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 $\nu > 20$.

Finally, note that *one of the consequences of my works with copulas was the publication of a large review of copula theory and methods in Fantazzini (2011a,b,c)*: the first part introduces copula functions, the main theorems of copula theory, and discusses in detail elliptical and Archimedean copulas. The second part presents pair-copula constructions, measures of dependence, and estimation procedures (parametric, semi-parametric and non-parametric methods). The last part of the review deals with copula selections methods and copula evaluation using goodness-of-fit tests.

3.2 Market risk measurement

3.2.1 Simulation studies

Fantazzini (2009) 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. Fantazzini (2009) 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 the Data Generating Process (DGP). He 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, he 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 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.

Fantazzini (2009) 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. He 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. Ane and Kharoubi (2003) and Junker and May (2005). Finally, Fantazzini (2009) performed an empirical analysis with ten trivariate portfolios, where he quantified the risk of the portfolio under different joint distribution assumptions.

3.2.2 Russian markets

Fantazzini and Shangina (2019) examined 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, they considered the RTS index future for their analysis.

Fantazzini and Shangina (2019) considered four class of models: the Threshold-GARCH model by Glosten et al. (1993) and Zakoian (1994), the HAR model by Corsi (2009), the 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), and the Model Confidence Set by Hansen et al. (2011) associated with the asymmetric quantile loss (QL) function proposed by Gonzalez-Rivera et al. (2004).

Fantazzini and Shangina (2019) first evaluated how the contribution of both online search intensity and options-based implied volatility to the modelling of the volatility of the Russian RTS index future, changed over almost two decades. They found that both the sign and the size of their 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. They then performed a backtesting analy-

sis 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. Fantazzini and Shangina (2019) found that only TGARCH models were able to pass the Kupiec and Christoffersen's tests for most quantiles, and these models also showed 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. They 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, except for TGARCH models, their 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.

Fantazzini and Shangina (2019) also evaluated the backtesting performance of the competing models using a rolling out-of-sample of 250 days: they 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 (see Fig.2)[4]. 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, Fantazzini and Shangina (2019) also performed a robustness check to verify how their previous results changed with the hierarchical-VAR (HVAR) model with LASSO proposed by Nicholson et al. (2018), which is able to both accommodate a large number of regressors and to improve the model estimation and its forecasting performances. Assuming a vector autoregression with 22 lags containing the daily returns, the daily realized volatility, the implied volatility and the Google data, the HVAR approach proposed by Nicholson et al. (2018)) adds structured convex penalties to the least squares VAR problem, so that the optimization is given by

$$\min_{\nu,\Phi} \sum_{t=1}^{T} \left\| Y_t - \nu - \sum_{l=1}^{22} \Phi^l Y_{t-l} \right\|_F^2 + \lambda \left(\mathcal{P}_Y \left(\Phi \right) \right),$$

where $\boldsymbol{\nu}$ is an intercept vector, Φ^l are the usual coefficient matrices, $\|A\|_F$ 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

[4] A better quality image can be found at:
https://sites.google.com/site/deanfantazzini/publications/recursive_VaR.png

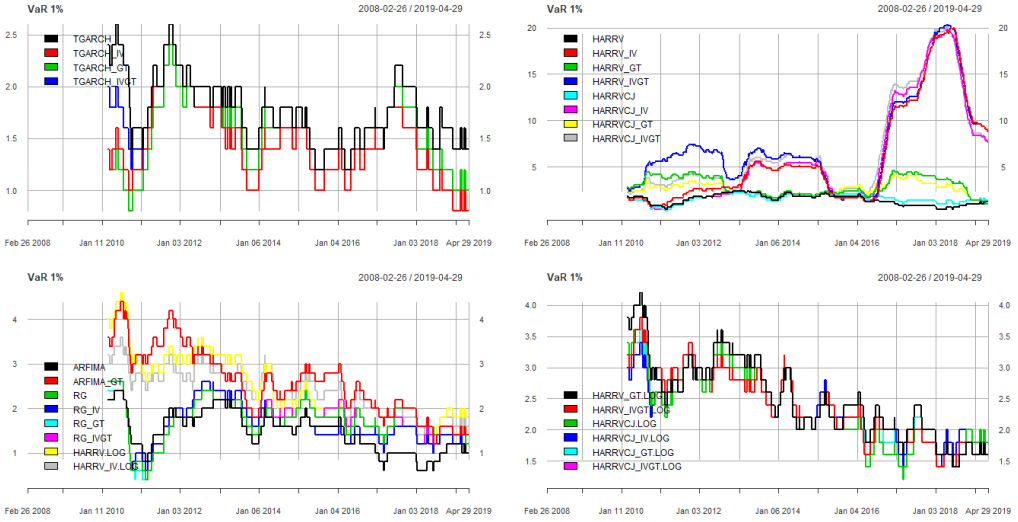

Figure 2: Violations in % of the VaR 1% using a rolling out-of-sample of 250 days.

including the lag order into hierarchical group LASSO penalties, which induce sparsity and a low maximum lag order. Fantazzini and Shangina (2019) 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. When the paper by Fantazzini and Shangina (2019) was published, this was the first application of an HVAR model for mark risk measurement purposes. 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.

3.3 Cryptocurrencies

Bitcoin is an online decentralized currency that allows users to buy goods and services and execute transactions, without involving third parties. It was launched in 2009 by a person or (more likely) by a group of people operating under the name of Satoshi Nakamoto. Bitcoin belongs to the large family of "cryptocurrencies", which are based on cryptographic methods of protection. The main characteristic of these currencies is their decentralized structure: there is no central authority which issues and regulates the currency, and transactions are executed using a peer-to-peer crypto-currency protocol without intermediaries. One of the best introduction to bitcoin and cryptocurrency technologies which requires only a minimal background in informatics is Narayanan et al. (2016), while I refer to Antonopoulos (2014) and Antonopoulos (2018) for a discussion at the advanced level.

3.3.1 Financial modelling of cryptocurrencies

The papers by 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. At the time these two papers were published, there were no similar reviews in the financial literature.

More specifically, Fantazzini et al. (2016) reviewed the methods employed to determine the main characteristics of bitcoin users, finding that the majority of users seem to be computer programming enthusiasts and people possibly engaged in illegal activity, whereas only a small part seem to be driven by political reasons or by investment motives. Nevertheless, these analyses are plagued by several limitations, like the possibility that the samples examined may not be representative of the full population of users and the speed with which bitcoin markets and users change over time, so that all analyses may be quickly out of date. Fantazzini et al. (2016) then examined the main models proposed to assess the bitcoin fundamental value, ranging from market sizing -which is more suitable for the medium/long term-, to the bitcoin marginal cost of production based on electricity consumption, which represents a lower bound in the short term.

Fantazzini et al. (2017) described several econometric approaches suggested to model bitcoin price dynamics, starting with cross-sectional regression models involving the majority of traded digital currencies and then moving to univariate and multivariate time series models, till models in the frequency domain. In general, all these methods confirmed that the main drivers of bitcoin price dynamics are still mainly of speculative nature (usually proxied with Google Trends and other social media data), followed by traditional supply and demand related variables, while global macro-financial variables play no role. Fantazzini et al. (2017) then reviewed the tests employed for detecting the existence of financial bubbles in bitcoin prices and which can be broadly classified into two large families, depending on whether they are intended to detect a single bubble, or (potentially) multiple bubbles. Most of these tests examined the months before the price crash that started in December 2013, while one analysis looked for multiple bubbles over the sample 2011-2014, finding evidence of explosive behavior in the bitcoin/USD exchange rates during August-October 2012 and November 2013 - February 2014. Finally, Fantazzini et al. (2017) examined a recent study dealing with the price discovery process in the Bitcoin market, which is of great importance for both short-term traders and long-term investors who want to know which exchange reacts most quickly to new information, thus reflecting the value of Bitcoin most precisely and efficiently.

3.3.2 Market and credit risk for cryptocurrencies

Fantazzini and Zimin (2020) proposed 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.

The ZPP was firstly introduced in Fantazzini, De Giuli and Maggi (2008) and relies on the fact that the probability of default (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.

Fantazzini and Zimin (2020) are the first to provide a large discussion about credit risk

and market risk for cryptocurrencies and how these risks can be defined with these assets. More specifically, 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. Table 1 in Fantazzini and Zimin (2020) summarizes the main aspects of credit and market risk for cryptocurrencies and it is reported below[5]:

	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

More specifically, Fantazzini and Zimin (2020) employ VAR-DCC and VAR-Copula-GARCH models with different specifications for the error terms to estimate the market risk for a portfolio of cryptocurrencies by using the Value-at-Risk and the Expected Shortfall, and simultaneously to estimate also their credit risk, which is given by the probability of default/death for the single coins and which is computed using the ZPP approach. In this regard, Fantazzini and Zimin (2020) also developed two 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. Assuming that the

[5]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.

coin price differences $X_t = P_t - P_{t-1}$ are normally distributed, the closed-form solutions for the ZPP in these two special cases are briefly described below:

- *Random walk with drift*: if $X_t = \mu + \varepsilon_t$, with $\varepsilon_t \sim NID(0, \sigma^2)$, and we use the results by Su and Rieger (2009) for a geometric Brownian motion with drift together with a barrier $H = 0$, it is straightforward to show that the 1-yead ahead ZPP can be computed at time t as follows:

$$ZPP_{RW} = P_{RW}[P_\tau \leq 0] \approx 2\Phi\left(\frac{P_t - \mu T}{\sigma\sqrt{T}}\right), \quad \text{with } t < \tau \leq t + T \tag{15}$$

  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

$$ZPP_{GARCH11} = P_{GARCH11}[P_\tau \leq 0] \approx 2\Phi\left(\frac{P_t - \mu T}{\sigma_T}\right), \text{with } t < \tau \leq t + T \tag{16}$$

  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[6]

$$\sigma_T^2 = \omega \cdot [A \times B] + \frac{[1 - (\alpha + \beta)^T]}{[1 - (\alpha + \beta)]} \cdot \hat{\sigma}_{t+1|t}^2 \tag{17}$$

  where $\hat{\sigma}_{t+1|t}^2$ can be either the forecasted variance at time $t+1$ conditional on information at time t, or the unconditional variance forecast[7], while A and B are two vectors defined below:

$$\underbrace{A}_{1\times(T-1)} = [(T-1) \quad (T-2) \quad \ldots \quad 2 \quad 1], \qquad \underbrace{B}_{(T-1)\times 1} = \begin{bmatrix} (\alpha+\beta)^0 \\ (\alpha+\beta)^1 \\ \vdots \\ (\alpha+\beta)^{T-3} \\ (\alpha+\beta)^{T-2} \end{bmatrix}$$

Fantazzini and Zimin (2020) performed a backtesting exercise for market risk modelling using two datasets of 5 and 15 coins: they 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 (2009) showed that, when small samples are considered and the data are skewed and leptokurtic, the biases

[6] Compute the recursive forecasts of the conditional variance from time $t+1$ till time $t+T$; then collect all the common components and use the property of the geometric series. The result will be given by (17).

[7] This second option is preferable in case of risk management.

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.

Fantazzini and Zimin (2020) also performed a backtesting exercise for credit risk modelling using a dataset of 42 coins. Their 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. Interestingly, as the Editors of the Journal of Industrial and Business Economics noted in the editorial accompanying the journal issue containing the paper by Fantazzini and Zimin (2020), "*the traditional credit scoring models based on the previous month trading volume, the one-year trading volume and the average yearly Google search volume work remarkably well, suggesting indeed a similarity between the newly defined credit risk for cryptocurrencies and the one traditionally used for other asset classes*[8]."

Finally, Fantazzini and Zimin (2020) performed a set of robustness checks to verify that our results also hold under different forecasting setups. They 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.

3.4 Financial bubbles modelling and testing

3.4.1 Bubbles

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), 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

[8] https://link.springer.com/article/10.1007/s40812-019-00138-6

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), Sornette (2003b) and Sornette (2003a). 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 (2018). 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.

Geraskin and Fantazzini (2013) presented an easy-to-use and self-contained guide for modeling and detecting financial bubbles with LPPLs, which contains the sufficient steps to derive the main models and discusses the important aspects for practitioners and researchers and proposed several extensions. They reviewed the original JLS model and they discussed early criticism to this approach and recent generalizations proposed to answer these remarks. Moreover, they described three different estimation methodologies which can be employed to estimate LPPL models. They 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 and diagnostic tests based on rational expectation models with stochastic meanreverting termination times, as well as graphical tools useful for capturing bubble development and for understanding whether a crash is in sight or not. Moreover, they first proposed in the financial literature the use of the Google search volume index to get some insights as to when a bubble started: the choice of the starting point for a bubble is of critical importance when modelling and testing such hypothesis. They 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. Figure 3 reports the Gold price and Google search volume index: it is evident that a massive interest around gold started to build during the year 2008, just before the price minima in November 2008, so that a possible LPPL model can be fitted using data starting from

the year 2008.

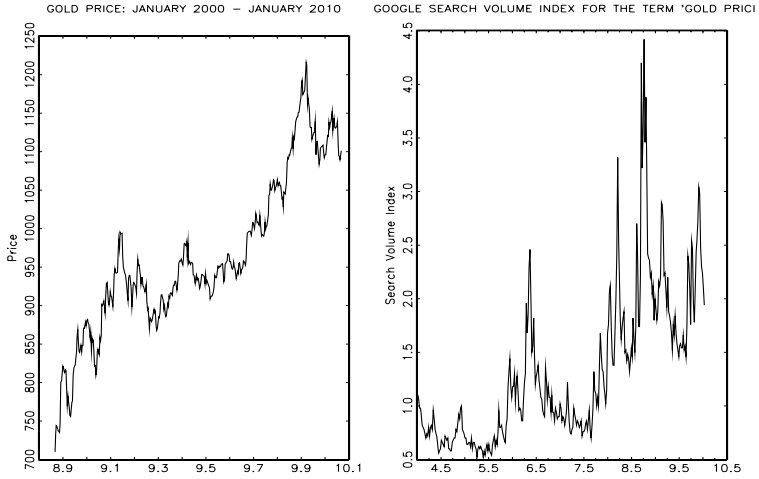

Figure 3: Gold price and `Google` Search Volume Index. Time t converted in units of 1-year

3.4.2 Anti-Bubbles

Fantazzini (2010a) proposed the use of log-periodic power law models to model and forecast the global financial crisis of 2007-2009 as a special case of "anti-bubble". 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.

By observing that world stock markets peaked in October 2007, Fantazzini (2010a) presented an econometric investigation of the log-periodic models, which tackled potential inferential problems arising from autocorrelation and heteroskedasticity. He considered the AR(1)- GARCH(1,1) log-periodic power law model 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. He then compared the previous set of log-periodic power law models with standard times series models in terms of long-term out-ofsample forecasting performances. Fantazzini (2010a) performed forecasting exercises with the American SP500 stock index, considering 120-step ahead and 180-step ahead forecasts: he 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 LPPL model showed mixed results, probably due to the misspecifications resulting from not considering the short-term dynamics. The main implication that Fantazzini (2010a) drew was 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. Figure 4 displays the plot of the SP500 stock index in the years 2007-2009 and the fitted curve according to the log-periodic power-law specification:

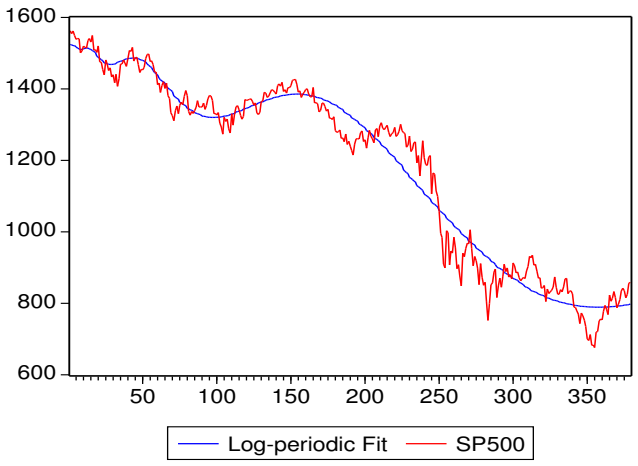

Figure 4: SP500 (10/10/2007 - 13/04/2009) and the log-periodic fit.

Interestingly, Fantazzini (2010a) concluded the paper by computing an ex-ante long-term out-of-sample forecast produced by the AR(1)-GARCH(1,1) LPPL model from the 14/04/2009 till the 09/10/2010, together with 95% bootstrap confidence bands. This ex-ante forecast of the SP500 index which was originally submitted to the Economics Bulletin on the 15/05/2009 was later analyzed by Fantazzini (2011d). Figure 5 below shows 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, as well as a vertical line highlighting the day of the FED Chairman's speech at Jackson Hole (Wyo., USA):

The previous figure (which is Figure 1 in Fantazzini (2011d)) shows 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, he 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 which announced the second round of Quantitative Easing (QE), the stock market index never returned inside the forecast confidence bands: had additional rounds of QE 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. Fantazzini (2011d) also provides additional evidence 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.

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.

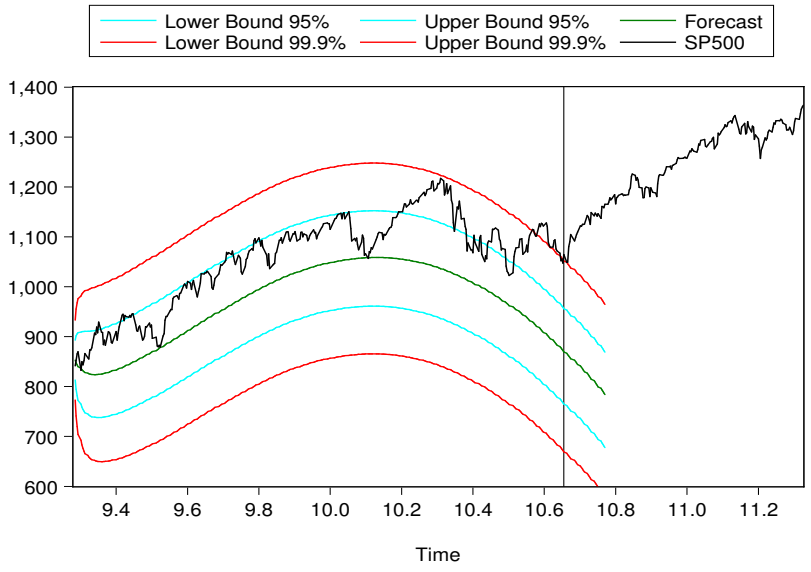

Figure 5: *SP500 ex-ante forecast, SP500 realized values, 95% and 99.9% confidence bands over the time sample* ***14/04/2009 - 29/04/2011****. The vertical black line on the 27/08/2010 signals the day of the Chairman's speech at Jackson Hole (Wyo., USA). Time* `t` *converted in units of one year (0 is set at Jan. 1st 2000).*

Fantazzini (2016) employed two sets of bubble detection strategies to corroborate this proposition: the first set consisted of econometric tests for financial bubbles proposed by Phillips et al. (2015) and Phillips and Shi (2018). 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 for negative financial bubbles developed by Yan et al. (2012). This model adapts the 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. Moreover, following Geraskin and Fantazzini (2013), Fantazzini (2016) used Google Trends to select the optimal time sample for the model estimation: the Google search indexes for the topics "*West Texas Intermediate*" and "*Brent Crude*" showed that a large interest in these oil prices started to build at the beginning of 2014, so that a time sample from January 2013 to April 2015 was chosen. Note that a large 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), Shiller (2002), Dyck and Zingales (2003), Case and Shiller (2003), Veldkamp (2006), Bhattacharya et al. (2009).

Despite the methodological differences between the two sets of 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 these 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 the work by Fantazzini (2016) 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, see Fantazzini (2016) and references therein. 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.

3.5 Energy markets

3.5.1 Coal-fired power plants

Chronologically speaking, the paper by Fantazzini (2016) was not the first paper of this dissertation using Google data for analyzing an important energy market.

The increase of oil and natural gas prices since the year 2000 stimulated the planning and construction of almost 150 coal-fired electricity generating plants by 2007 (of Energy (2007)). Several coal-to-liquids (CTL hereafter) plants were also proposed (see Hook et al. (2014)). 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[9] in April 2012. Meanwhile, the construction cost for coal plants has increased considerably but US coal prices remain relatively low (see Freese et al. (2011) and Fleischman et al. (2013) for recent reviews). Since 2011 the US Environmental Protection Agency (EPA) has began regulating greenhouse gases from mobile and stationary sources

[9] Million British thermal units.

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 cancelled or abandoned by 2013. This estimate is based on the Sierra Club database Club (2014) and the CoalSwarm - Center for Media and Democracy (2014) database. The Energy Information Administration (EIA) expects that very few new coal plants will be built through 2040 EIA (2014). *Given this background, Fantazzini an Maggi (2015) analyzed 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 (CoalSwarm - Center for Media and Democracy; 2014), and a large set of binary models.*

The work by Fantazzini and Maggi (2015) was the first study that analyzed the variables influencing the viability of a coal plant project after the advent of US shale gas and the global economic crisis in 2008-2009. In this regard, 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 Ansolabehere and Konisky (2009) and Garrone and Groppi (2012) for extensive reviews). Therefore, beside common industrial explanatory variables (size, input, output and labour costs, substitute costs, infrastructure), Fantazzini and Maggi (2015) also considered measures of social and environmental awareness such as Google search data to measure public attitudes towards coal plants. Table 2 illustrates the regressors that Fantazzini and Maggi (2015) used to explain the status of a coal plant project.

After controlling for collinearity, stationarity and robustness, Fantazzini and Maggi (2015) performed an extensive model specification, comparison and selection: they 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, they 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.

Fantazzini and Maggi (2015) tested the predictive power of their binary data models by means of an out-of-sample comparison and their previous findings were confirmed. These results also held 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 developed by the political scientists Poole and Rosenthal in the early 1980s to analyze the legislative roll-call voting behavior in the US congress, see Poole and Rosenthal (1985) and Poole (2005).

Variables	Description	Sources
	Externalities costs	
CO2(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 photo-voltaics system ($/W)	US Department of Energy (DOE) / Lawrence Berkeley National Laboratory
NG PRICE	US Henry Hub natural gas price ($/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 2: Regressors: description and source

3.6 Social welfare and social well-being

The previous successful use of Google data to measure social attitudes stimulated additional research work in the fields of social welfare and social well being. In general, social welfare systems help individuals and families through programs such as unemployment benefits or food stamps, just to name the most well known programs. 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 6 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*

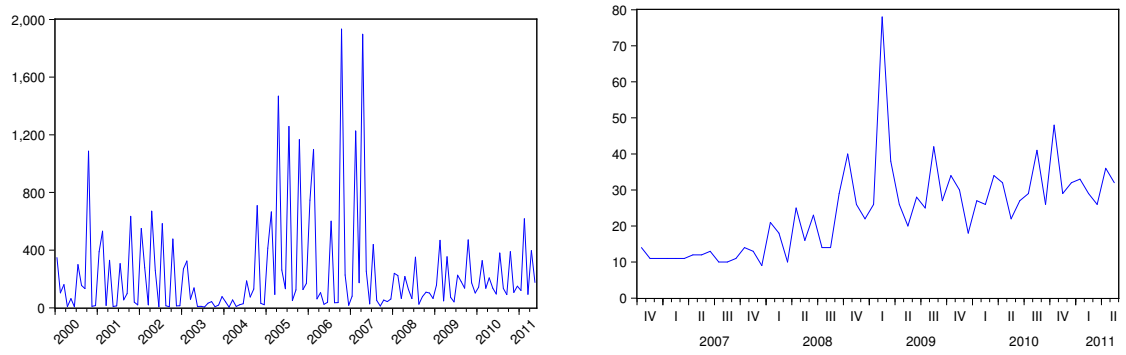

Figure 6: **Bloomberg story-count for "food stamps" worldwide (left plot); Google standardized volume of news related to "food stamps" worldwide (right plot)**.

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"[10]. 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*",The-Economist (2011). 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

[10]The original note by Mark Doms was posted at http://www.esa.doc.gov/Blog/2011/08/26/no-smoke-and-mirrors-gdp-note-bea's-recent-revisions . In 2019, this page and its cached version are no longer available.

in this regard is very limited: the main paper dealing with food stamps forecasting is in fact the one by Dynaski et al. (1991) for the USDA in 1991. Despite an extensive modelling effort, Dynaski et al. (1991) 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 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 Klerman and Danielson (2011), the review by Wilde (2013) and references therein for a discussion and an overview of this literature

Fantazzini (2014) proposed to use Google search data for nowcasting and forecasting the monthly number of food stamps recipients: he justify this choice because the administrative burden for enrolling and remaining enrolled in the food stamps program is nontrivial, see e.g. Bartlett et al. (2004), Office (1999) and Klerman and Danielson (2011), 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 by Fantazzini (2014) is a detailed analysis of the main determinants of food stamps dynamics using the structural relationship identification methodology discussed by Sa-ngasoongsong et al. (2012) and Greenslade et al. (2002), 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, he 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, Fantazzini (2014) 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 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. These results held 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 proposed by D'Amuri and Marcucci (2017). Similar results were found when considering the directional accuracy of the models' forecasts and when forecasting at the state-level. To get an intuitive idea of why Google based models forecasted so well food stamps 12 steps and 24 steps ahead, I report in Figure 7 the yearly changes of the number of food stamps recipients and the Google index for "food stamps" (this is Figure 5 in Fantazzini (2014)).

The yearly changes of the GI for the keywords "food stamps" show a similar pattern to the yearly changes of the number of food stamps, but the former always anticipate the turning points of the latter: from a minimum of 3 months in advance in 2006, up to 16 months in 2008 and 14 months in 2010.

There has been a growing interest in measuring the people well-being because tradi-

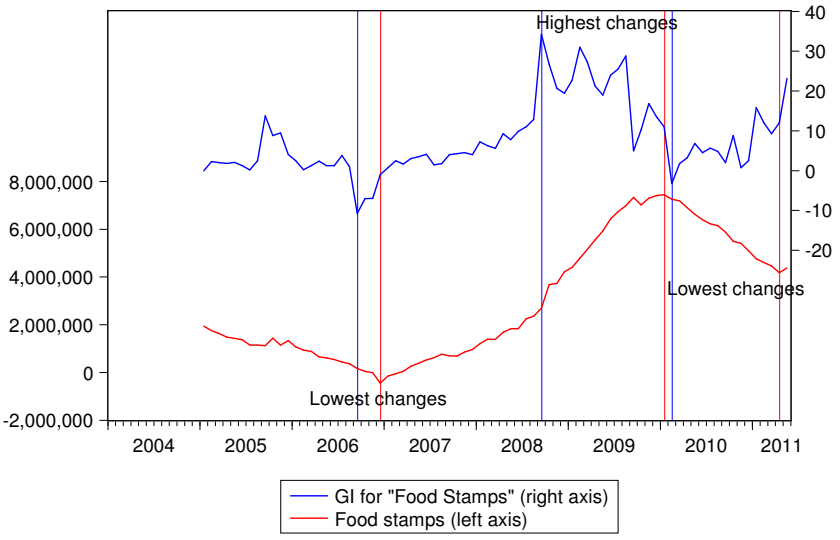

Figure 7: **Yearly changes for the food stamps data and the Google index for "food stamps"**. Sample: 2004M1 - 2011M5. The turning points for each series is highlighted by a vertical line of the same color.

tional macroeconomic indicators cannot measure moral values, friendship, happiness, etc. In this regard, an increasing number of studies has showed that social media can better assess people's opinions and behaviors than traditional surveys where respondents can distort their answers (O'Connor et al. (2010), Ceron et al. (2014), Mavragani and Tsagarakis (2016), Isotalo et al. (2016), Oliveira et al. (2017)). Besides, there is an extensive literature in sociology that examines the problem of respondents not telling the truth, starting from the well-known work by Hyman (1944) to more recent studies, see e.g. Preisendörfer and Wolter (2014) and Whiteley (2016). All these works emphasize that the predictive power of social media and online search data increases as the number of citizens expressing their opinions over the Internet increases over time. Moreover, Cody et al. (2016) found that "*not only [Twitter data] can anticipate survey responses, but we can use individual words within tweets to determine why one time period is happier than another, something that is not possible in traditional polls due to the multiple-choice aspect of most surveys*". Polykalas et al. (2013) argued that "*...web search based predictions may soon rival traditional polls despite the fact that the data released to the public by major search engines contain less demographic information compared to traditional polls*". However, they also highlighted that the effects of social events not related to the topic of research have to be filtered out as noise to get a reliable estimation of the model parameters. Similarly, Beauchamp (2017) highlighted the need to filter the data and correctly model them "*...to measure, extrapolate, and interpolate properly representative polling variation, both across states and over time*". In this regard, Algan et al. (2019) clearly showed that Google Trends data can improve substantially our understanding of people well-being and they have several important advantages over standard methods used to measure well-being. First, they allow us to examine people's behavior which is much more informative than the answers to a selective group of questions: searching for information online is a simple way to reveal our preferences and our priorities. Moreover, by the specific construction of Google data, it is very easy to see the relative importance of different issues and how

it changes over time. Third, Google data are very timely and this is of particular interest to policy-makers. Furthermore, they are available at the local level and can cover an extremely wide range of issues, which is not the case for standard measures of well-being. Last but not least, they are free and for policymakers who struggle with budget constraints, this is definitely a big advantage.

In Russia, the social well-being indices computed by the All-Russian Center for the Study of Public Opinion (VTsIOM) are widely used to assess well-being. These indices reflect the respondents' answers to questions regarding personal and social life and they reflect the population's mood to current and future living conditions in Russia. *Fantazzini et al. (2018) showed that it is possibile to use Google data to explain and predict the dynamics of the Russian social well-being indices computed by VTsIOM*. More specifically, Fantazzini et al. (2018) constructed a set of Google indices using a Google Trends dataset for 2006-2016 containing 512 search queries relative to housing conditions, income, education, etc.. After the time series of the search queries were collected, a data cleaning procedure was employed to deal with the presence of jumps, trends, zero search volumes, and seasonalities. Factor analysis was then applied to the standardized time series to create a small set of well-being indexes, while Bayesian Model Averaging was used to select the indexes mostly associated with the VCIOM indices measuring the social well-being of the Russian population. Additional regression models and forecasting exercises confirmed the validity of this approach.

Google-based indexes of well-being are an interesting additional tool that can be used along with standard indexes: the main determinants of the VTSIOM social well-being indices can be used to identify key areas in socio-economic policies, which can be useful when developing future programs aimed at improving the people well-being and their quality of life. Moreover, the categories of subjective well-being built using the search queries are statistically reliable and can be used for further analyses, such as understanding which factors mainly determine the social well-being of a person during an economic crisis, or after the implementation of specific socio-economic policy measures.

3.7 Sales forecasting

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 Bruhl 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 and the increasing popularity of Google as a search engine for obtaining information about cars, 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. 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 Stats; 2014).

Fantazzini and Toktamysova (2015) considered multivariate models for both deseasonalized data and for raw data, and performed a forecasting exercise for ten car brands in Germany, computing out-of-sample forecasts ranging from 1 month to 24 months ahead. Their 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, Fantazzini and Toktamysova (2015) performed a set of robustness checks to verify that their results also hold under different forecasting setups. They 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. These 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. Fantazzini and Toktamysova (2015) 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. Again, to give a simple idea why Google data can help medium- and long-term forecasting, the plots of the monthly car sales (right vertical axis) and the Google searches for the car brands (left vertical axis) are reported in Figure 8.

It is interesting to note that the turning points in the google data anticipate those in the car sales for all car brands.

4 Conclusions

Google Trends has become an important source for big data research, and several researchers in the economic and financial fields have begun to use Google search data for nowcasting and forecasting purposes, see Jun et al. (2018) for a review. The papers of my dissertation strongly contributed to the understanding of how Google data explain consumer and investor decision-making processes in several economic and financial fields, from risk management to crypto-currencies, from bubble modelling and testing to energy markets, from social welfare to social well-being, highlighting both advantages and disadvantages.

Copula models can deal with departures from the assumption of normality and they are computationally tractable even with large datasets, thus solving two important issues when dealing with big data. Particularly in finance, the use of copulas represented a major step forward in modelling complex multivariate dynamics. The papers of my dissertation further developed the econometric theory of copulas, focusing on the T-copula which is the most important for risk and portfolio management. Moreover, they improved the understanding of the effects of biased multivariate estimates on risk measurement. Furthermore, I was among the first to propose copulas for modeling high dimensional operational risks in a more flexible way, which is currently one of the main approaches to modeling and measuring operational risks. At the end of 2019, the papers that constitute this dissertation were cited almost 400 times according to Google Scholar. Several results discussed in these papers were included in the monograph titled *"Quantitative finance with R and Cryptocurrencies"* published by Amazon KDP in 2019. This textbook was officially presented at the Russian Central Bank on the 22/10/2019. The papers of this dissertation are often used in the professional field, particularly in risk management and business planning, and several professionals contacted me for comments and suggestions.

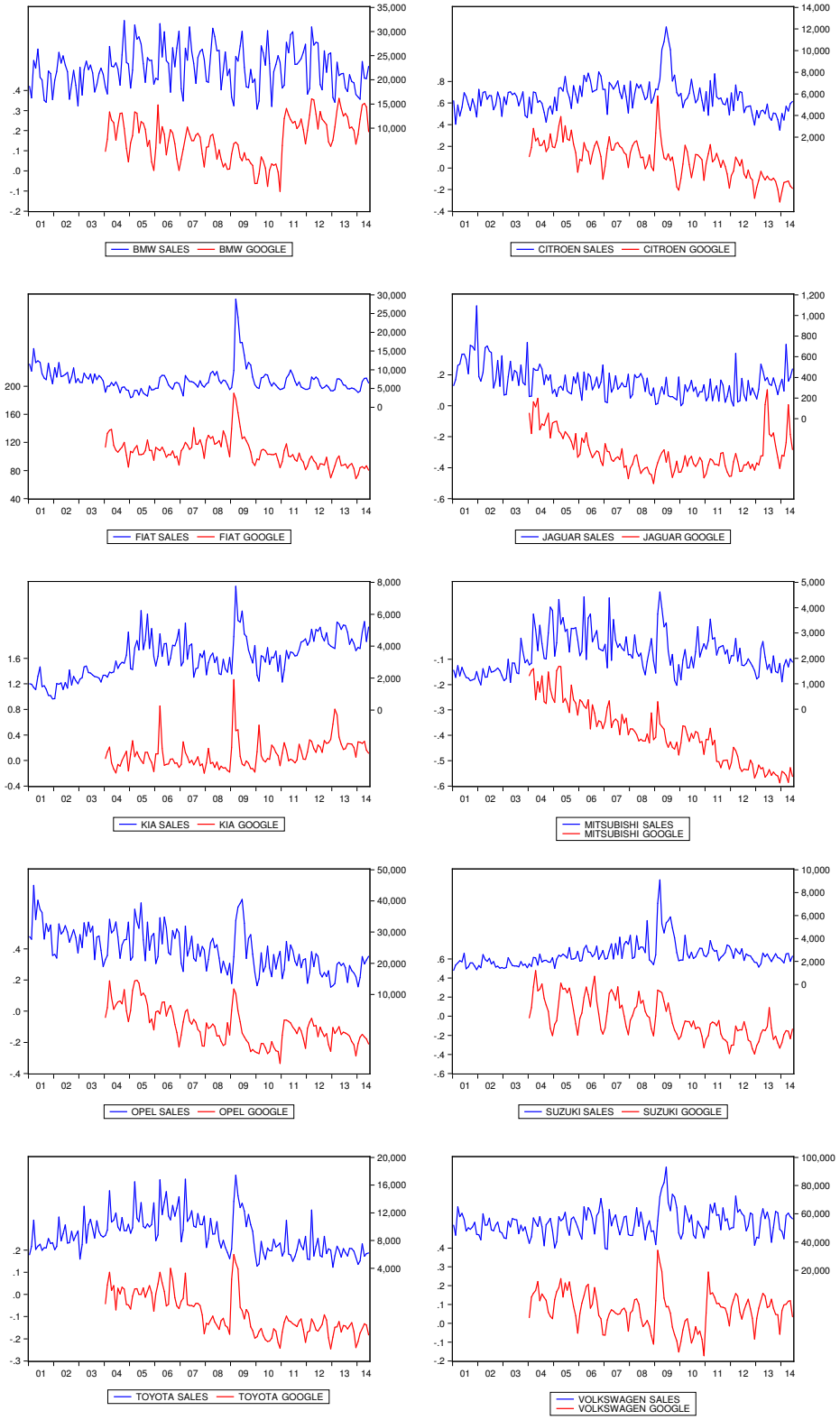

Figure 8: Car sales (right vertical axis) and relative GIs (left vertical axis) - not seasonally adjusted. Sample: 2001M1 - 2014M6.

References

Algan, Y., Murtin, F., Beasley, E., Higa, K. and Senik, C. (2019). Well-being through the lens of the internet, *PloS one* **14**(1): e0209562.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2003). Modeling and forecasting realized volatility, *Econometrica* **71**(2): 579–625.

Ane, T. and Kharoubi, C. (2003). Dependence structure and risk measure, *The journal of business* **76**(3): 411–438.

Ansolabehere, S. and Konisky, D. (2009). Public attitudes toward construction of new power plants, *Public Opinion Quarterly* **doi: 10.1093/poq/nfp041**.

Antonopoulos, A. M. (2014). *Mastering Bitcoin: unlocking digital cryptocurrencies*, " O'Reilly Media, Inc.".

Antonopoulos, A. M. (2018). *Mastering Ethereum: Building Smart Contracts and Dapps*, " O'Reilly Media, Inc.".

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

Bams, D., Blanchard, G. and Lehnert, T. (2017). Volatility measures and value-at-risk, *International Journal of Forecasting* **33**(4): 848–863.

Bartlett, S., Burstein, N. and Hamilton, W. (2004). Food stamp program access study final report., *Technical report*, Washington, DC: U.S. Department of Agriculture, Economic Research Service., Available from www.myfoodstamps.org/pdf files/ProgAccess.pdf. Accessed 2013 January 15.

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.

Beauchamp, N. (2017). Predicting and interpolating state-level polls using twitter textual data, *American Journal of Political Science* **61**(2): 490–503.

Bhattacharya, U., Galpin, N., Ray, R. and Yu, X. (2009). The role of the media in the internet ipo bubble, *Journal of Financial and Quantitative Analysis* **44**(03): 657–682.

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

Bouye, E., Durrleman, V., Nikeghbali, A., Riboulet, G. and Roncalli, T. (2001). Copulas for finance a reading guide and some applications, *Technical report*, Groupe de Recherche Operationnelle, Credit Lyonnais, Working Paper.

Bruhl, B.and Hulsmann, M., Borscheid, D., Friedrich, C. and Reith, D. (2009). A sales forecast model for the german automobile market based on time series analysis and data mining methods., *in* P. Perner (ed.), *Advances in Data Mining Applications and Theoretical Aspects*, Springer, pp. 146–160.

Case, K. E. and Shiller, R. J. (2003). Is there a bubble in the housing market?, *Brookings Papers on Economic Activity* **2003**(2): 299–362.

Ceron, A., Curini, L., Iacus, S. M. and 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 & society* **16**(2): 340–358.

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.

Christoffersen, P. F. (1998). Evaluating interval forecasts, *International economic review* **39**: 841–862.

Club, S. (2014). Proposed coal plant tracker, http://content.sierraclub.org/coal/environmentallaw/plant-tracker.

CoalSwarm - Center for Media and Democracy (2014). Proposed coal plants in the United States, http://www.sourcewatch.org/index.php/Category:Proposed_coal_plants_in_the_United_States.

Cody, E. M., Reagan, A. J., Dodds, P. S. and Danforth, C. M. (2016). Public opinion polling with twitter, *arXiv preprint arXiv:1608.02024* .

Collins, D. (2010). Cftc steps up on limits. futures: news, analysis & strategies for futures, *Option Derivatives Traders* **39**(2): 14.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility, *Journal of Financial Econometrics* **7**(2): 174–196.

Cosgrove, M. (2009). Banishing energy speculators is not the solution, *FOW: Global Derivatives Management* **454**: 39.

Dekimpe, M. G., Hanssens, D. M. and Silva-Risso, J. M. (1998). Long-run effects of price promotions in scanner markets, *Journal of Econometrics* **89**(1): 269–291.

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.

Domanski, D., Kearns, J., Lombardi, M. J. and Shin, H. S. (2015). Oil and debt, *BIS Quarterly Review* **March**: 55–65.

Dutt, H. R. and Harris, L. E. (2005). Position limits for cash-settled derivative contracts, *Journal of Futures Markets* **25**(10): 945–965.

Dyck, A. and Zingales, L. (2003). The bubble and the media, *in* P. Cornelius and B. Kogut (eds), *Corporate governance and capital flows in a global economy*, Oxford University Press, pp. 83–104.

Dynaski, M., Rangarajan, A. and 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.

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.

EIA (2014). EIA Annual Energy Outlook 2014, http://www.eia.gov/forecasts/aeo/.

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., De Giuli, M. and Maggi, M. (2008). A new approach for firm value and default probability estimation beyond merton models, *Computational Economics* **31**(2): 161–180.

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 Toktamysova, Z. (2015). Forecasting german car sales using google data and multivariate models, *International Journal of Production Economics* **170**: 97–135.

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.

Fleischman, L., Cleetus, R., Clemmer, S., Deyette, J. and Frenkel, S. (2013). Ripe for retirement: An economic analysis of the U.S. coal fleet, *The Electricity Journal* **26**(10): 51–63.

Freese, B., Clemmer, S., Martinez, C. and Nogee, A. (2011). A risky proposition – The financial hazards of new investments in coal plants, *Working paper*, Union Of Concerned Scientists.

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.

Garrone, P. and Groppi, A. (2012). Siting locally-unwanted facilities: What can be learnt from the location of Italian power plants, *Energy Policy* **45**: 176–186.

Gazola, L., Fernandes, C., Pizzinga, A. and Riera, R. (2008). The log-periodic-ar (1)-garch (1, 1) model for financial crashes, *The European Physical Journal B* **61**(3): 355–362.

Genest, C., Ghoudi, K. and Rivest, L. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, *Biometrika* **82**(3): 543–552.

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.

Glosten, L. R., Jagannathan, R. and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks, *The journal of finance* **48**(5): 1779–1801.

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.

Goldenfeld, N. (2018). *Lectures on phase transitions and the renormalization group*, CRC Press.

Gonzalez-Rivera, G., Lee, T.-H. and 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.

Greenslade, J. V., Hall, S. G. and Henry, S. B. (2002). On the identification of cointegrated systems in small samples: a modelling strategy with an application to uk wages and prices, *Journal of Economic Dynamics and Control* **26**(9-10): 1517–1537.

GTAI (2014). http://www.gtai.de.

Gürkaynak, R. S. (2008). Econometric tests of asset price bubbles: taking stock, *Journal of Economic surveys* **22**(1): 166–186.

Hansen, P. R., Huang, Z. and Shek, H. H. (2012). Realized garch: a joint model for returns and realized measures of volatility, *Journal of Applied Econometrics* **27**(6): 877–906.

Hansen, P. R., Lunde, A. and Nason, J. M. (2011). The model confidence set, *Econometrica* **79**(2): 453–497.

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

Hook, M., Fantazzini, D., Angelantoni, A. and Snowden, S. (2014). Hydrocarbon liquefaction: viability as a peak oil mitigation strategy, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **372**(2006): 20120319.

Hwang, S. and Valls Pereira, P. L. (2006). Small sample properties of garch estimates and persistence, *The European Journal of Finance* **12**(6-7): 473–494.

Hyman, H. (1944). Do they tell the truth?, *The Public Opinion Quarterly* **8**(4): 557–559.

Internet World Stats (2014). http://www.internetworldstats.com/stats.htm.

Isotalo, V., Saari, P., Paasivaara, M., Steineker, A. and Gloor, P. A. (2016). Predicting 2016 US Presidential Election Polls with Online and Media Variables, *Designing Networks for Innovation and Improvisation*, Springer, pp. 45–53.

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.

Johansen, A., Ledoit, O. and Sornette, D. (2000). Crashes as critical points, *International Journal of Theoretical and Applied Finance* **3**(02): 219–255.

Johansen, A. and Sornette, D. (1999). Financial" anti-bubbles": Log-periodicity in gold and nikkei collapses, *International Journal of Modern Physics C* **10**(04): 563–575.

Johansen, A., Sornette, D. and Ledoit, O. (1999). Predicting financial crashes using discrete scale invariance, *Journal of Risk* **1**(4): 5–32.

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.

Junker, M. and May, A. (2005). Measurement of aggregate risk with copulas, *The Econometrics Journal* **8**(3): 428–454.

Kaizoji, T. and Sornette, D. (2010). Bubbles and crashes, *Encyclopedia of quantitative finance* .

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

Kilian, L. (2008). The economic effects of energy price shocks, *Journal of Economic Literature* **46**(4): 871–909. **URL:** *http://www.aeaweb.org/articles.php?doi=10.1257/jel.46.4.871*

Kim, G., Silvapulle, M. J. and Silvapulle, P. (2008). Estimating the error distribution in multivariate heteroscedastic time-series models, *Journal of Statistical Planning and Inference* **138**(5): 1442–1458.

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.

Klerman, J. and Danielson, C. (2011). Transformation of the supplemental nutrition assistance program, *J Policy Anal Manag* **30(4)**: 863–888.

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

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models, *The J. of Derivatives* **3**(2): 73–84.

Landwehr, J. R., Labroo, A. A. and Herrmann, A. (2011). Gut liking for the ordinary: incorporating design fluency improves automobile sales forecasts, *Marketing Science* **30**: 416–429.

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.

Lin, Y. and Zhang, W. J. (2004). Aesthetic design for automobile interiors: critical and conceptual framework, *2004 IEEE International Conference on Systems, Man and Cybernetics*, The Hague, pp. 6313–6317.

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*.

Maugeri, L. (2010). *Beyond the age of oil: the myths, realities, and future of fossil fuels and their alternatives*, Praeger.

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.

VDA (2014). http://www.vda.de.

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

Narayanan, A., Bonneau, J., Felten, E., Miller, A. and Goldfeder, S. (2016). *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*, Princeton University Press.

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

Nicholson, W. B., Wilms, I., Bien, J. and Matteson, D. S. (2018). High dimensional forecasting via interpretable vector autoregression, *arXiv preprint arXiv:1412.5250* .

Nijs, V. R., Dekimpe, M. G., Steenkamp, J.-B. E. M. and Hanssens, D. M. (2001). The category-demand effects of price promotions, *Marketing Science* **20**: 1–22.

O'Connor, B., Balasubramanyan, R., Routledge, B. R. and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series, *Fourth International AAAI Conference on Weblogs and Social Media*.

of Energy, T. N. E. T. L. D. (2007). www.netl.doe.gov.

Office, G. A. (1999). Food stamp program: Various factors have led to declining participation., *Available from http://www.gao.gov/archive/1999/rc99185.pdf . accessed 2013 january 15*, GAO/RCED-99-185. Washington, DC.

Oliveira, D. J. S., Bermejo, P. H. d. S. and dos Santos, P. A. (2017). Can social media reveal the preferences of voters? a comparison between sentiment analysis and traditional opinion polls, *Journal of Information Technology & Politics* **14**(1): 34–45.

Pauwels, K., Hanssens, D. M. and Siddarth, S. (2002). The long-term effects of price promotions on category incidence, brand choice, and purchase quantity, *Journal of Marketing Research* **39**: 421–439.

Pauwels, K., Silva-Risso, J., Srinivasan, S. and Hanssens, D. M. (2004). New products, sales promotions, and firm value: the case of the automobile industry, *The Journal of Marketing* **68**: 142–156.

Phillips, P. C. and Shi, S.-P. (2018). Financial bubble implosion and reverse regression, *Econometric Theory* **34**(4): 705–753.

Phillips, P., Shi, S. and Yu, J. (2015). Testing for Multiple Bubbles: Historical Episodes of Exuberance and Collapse in the SP500, *International Economic Review* **56**(4): 1043–1078.

Polykalas, S. E., Prezerakos, G. N. and Konidaris, A. (2013). An algorithm based on google trends' data for future prediction. case study: German elections, *IEEE International Symposium on Signal Processing and Information Technology*, IEEE, pp. 000069–000073.

Poole, K. (2005). *Spatial Models of Parliamentary Voting*, Cambridge University Press.

Poole, K. and Rosenthal, H. (1985). A spatial model for legislative roll call analysis, *American Journal of Political Science* **29**(2): 357–384.

Preisendörfer, P. and 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.

Reinhart, C. M. and Rogoff, K. S. (2009). *This time is different: Eight centuries of financial folly*, Princeton university press.

Rousseeuw, P. J. and Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices, *Communications in Statistics–Theory and Methods* **22**(4): 965–984.

Sa-ngasoongsong, A., Bukkapatnam, S. T., Kim, J., Iyer, P. S. and Suresh, R. (2012). Multi-step sales forecasting in automotive industry based on structural relationship identification, *International Journal of Production Economics* **140**(2): 875–887.

Sa-ngasoongsong, A. and Bukkapatnam, S. T. S. (2011). Willingness-to-pay prediction based on empirical mode decomposition, *in* T. Doolen and E. V. Aken (eds), *Proceedings of the 2011 Industrial Engineering Research Conference*, Reno.

Shahabuddin, S. (2009). Forecasting automobile sales, *Management Research News* **32**: 670–682.

Shiller, R. J. (2000). *Irrational exuberance*, Princeton University Press.

Shiller, R. J. (2002). Bubbles, human judgment, and expert opinion, *Financial Analysts Journal* **58**(3): 18–26.

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.

Sornette, D. (2003a). Critical market crashes, *Physics Reports* **378**(1): 1–98.

Sornette, D. (2003b). *Why stock markets crash: critical events in complex financial systems*, Princeton University Press.

Sornette, D. (2009). Dragon-kings, black swans and the prediction of crises, *arXiv preprint arXiv:0907.4290* .

Sornette, D., Ledoit, O. and Johansen, A. (1999). Critical crashes, *Risk Magazine* **12**(1): 91–95.

Sornette, D. and Woodard, R. (2010). Financial bubbles, real estate bubbles, derivative bubbles, and the financial and economic crisis, *Econophysics approaches to large-scale business data and financial crisis*, Springer, pp. 101–148.

Sornette, D., Woodard, R. and Zhou, W.-X. (2009). The 2006–2008 oil bubble: Evidence of speculation, and prediction, *Physica A: Statistical Mechanics and its Applications* **388**(8): 1571–1576.

Su, L. and Rieger, M. O. (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* .

The-Economist (2011). Food stamps - the struggle to eat.
URL: *http://www.economist.com/node/18958475*

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.

Tokic, D. (2015). The 2014 oil bust: Causes and consequences, *Energy Policy* **85**: 162–169.

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

Veldkamp, L. L. (2006). Media frenzies in markets for financial information, *The American Economic Review* **96**(3): 577–601.

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.

Wang, F., Chang, K. and Tzeng, C. (2011). Using adaptive network-based fuzzy inference system to forecast automobile sales, *Expert Systems with Applications* **38**: 10587–10593.

Whiteley, P. (2016). Why do voters lie to the pollsters?, *Political Insight* **7**(1): 16–19.

Wilde, P. (2013). The new normal: The supplemental nutrition assistance program (snap), *Am J Agr Econ* **95(2)**: 325–331.

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.

Yan, W., Woodard, R. and Sornette, D. (2012). Diagnosis and prediction of rebounds in financial markets, *Physica A: Statistical Mechanics and its Applications* **391**(4): 1361–1380.

Zakoian, J.-M. (1994). Threshold heteroskedastic models, *Journal of Economic Dynamics and control* **18**(5): 931–955.

Zhou, W.-X. and Sornette, D. (2005). Testing the stability of the 2000 US stock market "antibubble", *Physica A: Statistical Mechanics and its Applications* **348**: 428–452.

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.