11th International Institute of Forecasters' Workshop "Using Big Data for Forecasting and Statistics"

The 11th International Institute of Forecasters' Workshop took place in Frankfurt am Main on the 7th and 8th of April of 2014. It was hosted and sponsored by the European Central Bank[1]. The workshop attracted about 100 participants. The aim of the workshop was to gather big data research experts to discuss the latest discoveries in the field of big data and statistics. The workshop brought together experts with varying backgrounds: statisticians dealing with the use of big data in official statistics, econometricians developing forecasting models capable of exploiting the information contained in big data, experts in semantic analysis and construction of sentiment indices, etc.

Aurel Schubert, Director General Statistics at ECB, welcomed the participants and thanked the IIF and the organizers Chiara Osbat (ECB, DG-E), Gianni Amisano (ECB, DG-R), Per Nymand-Andersen (ECB, DG-S) and Juri Marcucci (Bank of Italy), for bringing such an impressive number of presenters and discussants on such a timely topic. Aurel offered some introductory remarks on the importance of applying new sources of financial and economic data such as big data as these may help central banks to take timely policy actions and to extract information on the impact of monetary policy actions within the financial system and the real economy alike. While well acknowledging that big data will not give us all the answers, Aurel remarked that, like any other type of data, they should be held to strict statistical quality standards and data quality frameworks. He recalled the importance of the workshop and encouraged the audience to address the challenges and quality issues that big data research might come up with. Aurel also emphasized the importance of transparency in data sources and in communicating statistics to the general public.

The workshop was opened by Hal Varian from Google Inc., who delivered the first Keynote Address. Hal presented several methods and results on how to use Google Trends data as a rich source for nowcasting economic series such as unemployment, car and property sales, and consumer sentiment. In particular, he showed some results obtained using Bayesian structural time series models, averaging across various specifications including different sets of Google search results. In two applications he showed how, by adding progressively the top best search terms picked by the algorithm, the nowcast errors shrink. The Google Trends data are now available to the public – via Google Trends. Varian also presented the new Google survey tool, which is available for corporates and public authorities wanting to conduct surveys using the internet.

The first session of the first day, Session 1, focused on "Big data: new sources and opportunities for central banking purposes". The aim of the three papers of this session was to present some of the new "big data" sources and the opportunities they offer to central bankers. These new data sources include Google search data, social media data, online news, postal services data and payment and transactional data.

Amel Aouadi, from Université d'Auvergne (France), presented "Can information demand help to predict stock market liquidity? Google it!". The paper looks at the impact of information demand on liquidity in the French stock market. Information demand, a concept that is being increasingly applied in various studies using big data, is proxied by the intensity of Google web searches. The paper finds that adding search volume to a model of turnover in the French stock

[1] http://www.ecb.europa.eu/events/conferences/html/20140407_workshop_on_using_big_data.en.html

market improves out-of-sample forecast performance. Furthermore, higher search volume is positively and significantly correlated with liquidity.

Piet Daas, from Statistics Netherlands, presented the paper “Social Media Sentiment and Consumer Confidence”. The paper takes the official statistics producer’s point of view and recounts the effort by Statistics Netherlands to use data collected on all social media messages written in Dutch to construct an index of consumer confidence using the “bag of words” approach. The authors find that the correlation with the official consumer confidence survey series depends on the social media platform on which the data are collected: the highest correlation is with messages harvested on Facebook (where not only correlation, but also co-integration is found). Confidence distilled from Twitter messages is also highly correlated with the survey indicator, while confidence constructed on LinkedIn or YouTube, for instance, is negatively correlated. The advantage of the web-based sentiment indicator is that it can be produced every month with a timeliness of about 2 weeks and could even be produced on a weekly basis, though it would tend to be very volatile. Future studies will focus on understanding the representativeness of the sample of harvested web messages.

José Anson, from the Universal Postal Union (UPU), presented the paper “A Short-Run Analysis of Exchange Rates and International Trade”. The paper describes the data collected by the UPU on near real-time tracking of parcels. The database contains information on the sender’s address, receiver’s address, and product sent (at HS-6 code level). It is well known that distance and language barriers significantly reduce trade intensity between two countries. One finding of the authors’ empirical application is that, while still significant, the negative coefficient of distance is more than halved while that of language barriers more than doubles. The author stresses that the rising predictive power of data from international postal networks might provide rich macroeconomic insights for policy makers in real time.

All the three papers of the first session were discussed by Menno Middeldorp, from the Bank of England, who summarized the opportunities presented to central bankers from big data, from nowcasting to the construction of confidence indicators using semantic analysis, but also highlighted the potential pitfalls. One such pitfall is the risk of falling into the trap of believing that “N=all”, as suggested by some widely cited literature on big data. The other is “information confusion”: the danger of not being able to disentangle the effect of private and public information (e.g. on financial returns), and also of potentially modifying behavior based on cycles in the agreed interpretations of signals (e.g. herding behavior may invalidate patterns, in a fashion akin to the mechanism outlined in the Lucas critique).

The second session of the first day, Session 2, focused on “Big data: a quality framework for big data”. In fact, the sheer size of newly available data is not enough to provide information that is usable and useful. When developing analytical studies based on big data, it is also important to address quality issues. The two papers presented in this session address this issue.

Paola Cerchiello and Paolo Giudici, from the University of Pavia (Italy), presented the paper “How to measure the quality of financial tweets**”**. The authors focus on financial tweets and propose a statistical method to assess their quality, leading to a ranking of different Twitter sources by “reliability”. To do this, they use the h-index introduced by Hirsch (2005) in the context of scientific citations to measure academic journal impact factors. They develop a statistical theory for this index, specifying a parametric distribution for the tweet production mechanism and one for re-tweet patterns. This allows them to use stochastic confidence intervals to compare the impact factor of various financial twitters.

Ric Clarke, from the Australian Bureau of Statistics (ABS), presented the next paper “Small steps towards Big Data – Some Initiatives by the Australian Bureau of Statistics**”**. The author outlines a number of considerations for the official statistician when deciding whether to embrace a particular big data source in the regular production of official statistics, introduces a framework for statistical inference from big data and provides a snapshot of current big data initiatives at the ABS. One illustration refers to using remote sensing to construct agricultural statistics in Australia.

Niels Ploug, from Statistics Denmark, discussed the two papers. In his discussion, Niels reviewed the statistical business process model and the statistical quality principles and framework. In that context, he highlighted some challenges in the use of big data for official statistics, related to data access (most big data are owned by private companies), data analysis (the difficulty of discerning patterns), data usage (fitting with the classification systems used by official statistics), data quality (in particular, the existence and richness of metadata) and privacy issues.

We adjourned for the morning with a great lunch which was served at the facilities of the European Central Bank. During lunch we had the first poster session entitled “Big data – new sources and new methods” which included five papers. Yigitcan Karabulut, from Goethe Universität, presented his paper “Can Facebook predict stock market activity?” where the author uses an index of investor sentiment based on Facebook’s Gross National Happiness (GNH) measure, which is calculated using textual analysis of emotion words posted on Facebook. In a VAR model, the author finds that Facebook’s GNH predicts changes in both daily returns and trading volume in the US stock market. For instance, an increase of one standard deviation in GNH is associated with an increase of 11.23 basis points in market returns over the next day. Nikolaus Askitas, from IZA, presented a paper titled “Detecting mortgage delinquencies with Google trends” in which he uses Google Trends data for specific keywords such as “hardship letter”, “short sale”, “REO” (Real Estate Owned), and “FHA” (Federal Housing Administration) to sketch a weekly picture of the US housing market in the crisis. The relative intensity of these searches is used to nowcast the quarterly share of delinquent mortgages held by members of MBA, finding that a model using Google Trends data outperforms a basic autoregressive model. Andreas Joseph, from the City University of Hong Kong, presented his paper “Netconomics: novel forecasting techniques from the combination of big data, network science and economics” where the author applies network science to very large economic datasets. Of particular interest was the application to the matrix of world trade, which introduced the concept of gate-keeping potential: a centrality measure that appears correlated to the future growth path of a country. Menno Middledorp, from the Bank of England, presented his paper titled “Measuring changing market expectations of bank resolution regimes using CDS and news flow data” where he looks at the introduction of laws designed to reduce the need for government support when financial institutions become insolvent. These rules imply a greater risk of default and losses for senior bond holders who may see their claims written down or converted to equity. The paper assesses the impact of market expectations of new resolution regimes using the frequency of Bloomberg news articles on resolution-related topics as an indicator of changing information on resolution regimes in a panel regression of 27 globally systemically important financial institutions (G-SIFIs) over eleven countries. The author finds that market expectations of resolution regimes, together with increased sovereign CDS, increased the CDS-implied probability of bank default over the period 2007-12. Dick van Dijk, from Erasmus University Rotterdam, presented his paper titled “Forecasting with many predictors: allowing for non-linearity” where he looks at 1, 3 and 6 month-ahead point forecasts of employment, personal income, industrial production and manufacturing sales using factor models, based on principal components of 130 macro and financial variables, and ridge regression. In both approaches, nonlinearity is allowed for by including squares and first-level interactions of the original explanatory variables. Allowing for nonlinearity reduces the mean squared prediction error (MSPE) by 10% to 15%, especially at longer horizons, thereby improving forecast quality.

After lunch, we attended two sessions. Session 3, entitled “Methods for big data”, included two papers. Massimiliano Marcellino, from Università Commerciale Luigi Bocconi (Italy), presented the paper “Real-Time Nowcasting with a Bayesian Mixed Frequency Model with Stochastic Volatility”. Massimiliano suggests an approach for nowcasting macroeconomic time series that deals with the curse of dimensionality by using Bayesian shrinkage methods as an alternative to factor models. The contributions of the paper are threefold. First, estimation with Bayesian shrinkage is a viable alternative to factor model methods. Second, incorporating stochastic volatility is useful for both point and density forecasts. Third, direct multi-step methods of forecasting are shown to be at least as accurate as iterated methods. Massimiliano presents results for quarterly GDP growth forecasts based on a range of monthly economic and financial indicators.

The second paper in the session “Mining Big Data Using Parsimonious Factor and Shrinkage Methods” was presented by Hyun Hak Kim, from the Bank of Korea. Kim also refers to the literature on common factors, recalling the shortcomings of Principal Component Analysis (PCA) and proposing two innovative methods: Independent Component Analysis (ICA) and Sparse Principal Component Analysis (SPCA). SPCA aids in interpreting PCA models by placing structure on the factor loadings in a large-size system. This dimensionality reduction is performed using alternative methods, such as bagging, boosting, ridge regression, least angle regression, elastic net and non-negative garotte. The empirical results provide new evidence of the usefulness of factor-based forecasting, especially when combined with shrinkage methods.

These two papers were discussed by Marek Jarocinski, from the European Central Bank. Marek highlighted the simplicity of the approach used by Marcellino and co-authors to deal with ragged edge data, but also its price, namely that the model used for forecasting in each given month of the year is different. However, the predictive performance evaluations suggest that this price is worth paying. He also offered an interesting unifying perspective on both papers by referring to Bayesian theory and suggesting that priors be formalised on the higher (even) moments of the distribution of the coefficients to the factors (or to the variables included in the model). This would make it easier to identify patterns among the many model specifications.

After the coffee break we had Session 4 titled “Nowcasting the macroeconomy using big data” in which two papers were presented. John Galbraith, from McGill University, presented the paper “Nowcasting GDP: Electronic Payments, Data Vintages and the Timing of Data Releases”. John shows results on a real-time forecasting exercise of Canadian GDP using model combination and a big data source: payment card and check data, available with high timeliness in Canada. This study uses vintage data for each quarterly forecast, compiling a new database of monthly GDP revisions, and produces evaluations of the nowcast of first-release and current-vintage GDP data for each quarter in the pseudo-out-of-sample period. The main findings are that i) nowcast quality improves markedly between months 1 and 5 of the observation period; ii) that model averaging produces noteworthy gains at all months; and iii) that using payments system variables appears to produce some gains at dates immediately preceding the release of national accounts.

The second paper entitled “Macroeconomic Nowcasting Using Google Probabilities” was presented by Luca Onorante, from Central Bank of Ireland. Luca presents a model that uses Google trends searches innovatively in forecasting macroeconomic variables: rather than using the search volume as a variable, Onorante and his co-author add to the literature by nowcasting using dynamic model averaging (DMA) methods which allow for model switching between time-varying parameter regression models. They allow for model switching to be controlled by the Google search intensity through “Google probabilities” which determine which nowcasting model should be used at each point in time.

In an exercise involving nine major monthly US macroeconomic variables, this approach provides large improvements in nowcasting performance.

Marco Lombardi, from the BIS, discussed both papers in this session. In his discussion Marco suggested that models which allow for time-varying coefficients could possibly further improve the results on using payment data to forecast GDP by capturing the different role of payments at different stages of the cycle. Similarly, he proposed looking more closely at the time-varying weights of alternative model specifications over the business cycle.

We adjourned for the day, which was full of thought-provoking presentations and discussions, by going to dinner to the "Emma Metzler" restaurant where Peter Praet, Member of the Executive Board of the ECB, gave a dinner speech indicating the policy maker's view on the most pressing issues where the analysis of big data can help and all the participants enjoyed a wonderful dinner.

The second day of the workshop was opened by Alberto Cavallo, from the Massachusetts Institute of Technology, who delivered an invited lecture in which he presented the current status of the Billion Prices Project @ MIT. This innovative database is based on web-scraping, i.e. collecting data from the websites of hundreds of retailers. Online data can be collected in near real-time, which enables information on price changes and inflation to be published at a daily frequency. In addition to using daily indices to try to track turning points in inflationary trends, the Billion Prices Project team also uses the disaggregated data to gauge phenomena such as exchange rate pass-through, the speed of price convergence after entering a monetary union, and frequency of price changes across many goods and across countries.

The first session of the day, Session 5, entitled "Catching animal spirits" included two papers. Ellen Tobback, from Antwerp University, presented a paper titled "Belgian economic policy uncertainty index: improvement through text mining". Ellen discusses the shortcomings of the widely quoted and used "Economic Policy Uncertainty index" (EPU) proposed by Baker, Bloom and Davis (2013) and proposes some improvements based on text mining algorithms. Using an example based on Flemish news articles from six sources, she shows that more advanced text mining methods, such as modality annotation and support vector machines, have a higher explanatory power for a financial market variable that is very sensitive to economic policy uncertainty: the 10-year OLO-Bund spread. She highlights the promising result in her paper, which points to the feasibility of automated tools for real-time nowcasting of macroeconomic variables based on online news.

Rickard Nyman, from University College London, presented a paper entitled "News and narratives in financial systems: exploiting big data for systemic risk assessment". Rickard begins by recalling the importance of narratives and emotions in driving financial market activity. Similarly to Ellen's paper, Nyman and co-authors use machine learning techniques to address two main research questions: measuring shifts in relative sentiment and homogenization, and the formation of "consensus narratives" via shifts in the distribution of narratives. Using the Reuters News archive, broker reports and internal Bank of England Market commentary, and focusing the semantic analysis on the poles of excitement and anxiety, they provide evidence that increasing narrative consensus that is high in excitement and low in anxiety leading up to the crisis can be a warning sign. Methodologically, they reduce the dimensionality of the space of words analyzed using principal components and they measure consensus by the reciprocal of the entropy, which is meant to offer an intuitive representation of the belief in a given paradigm in the financial markets.

The discussant, Johan Bollen from Indiana University, reviewed the literature on algorithmic semantic analysis and recalled some recent applications to forecasting in the social sciences, from predicting flus and elections to emotional contagion in social networks and financial market forecasting. His main comment on the approach chosen to improve the economic policy uncertainty index was that, in the supervised machine learning algorithm proposed by the authors, the most discriminating words were events rather than nouns or adjectives. He suggested an alternative approach based on psycho-social models of human emotion and the associated N-grams, or recurrent sets of words appearing together when a certain mood dominates. On the narratives approach, he commented that it is difficult to operationalize the concept of narrative with a bag-of-words approach, as narratives are semantics that evolve over time, requiring some form of knowledge modeling. He also stressed the importance of modeling social network effects when operationalizing the concept of narratives.

The second plenary session of the second day, Session 6, was entitled "Financial market sentiment" and two papers were presented. Huina Mao, from Indiana University, presented her paper titled "Quantifying the Effects of Online Bullishness on International Financial Markets". Huina reviews various investor sentiment measures, recalling that the relevance of such measures is in capturing the irrationality of noise traders, which can temporarily drive asset prices away from their fundamental values. She also introduces a new measure that extracts investor mood from Twitter data. The Twitter mood indicator is based on two terms: "bullish" and "bearish". This is compared and contrasted with a similar indicator extracted from Google searches of these two terms. Both indicators are correlated, at various leads and lags, with survey measures of investor sentiment, and the Twitter-based index is able to predict daily stock returns.

The second paper of the session was presented by Qingwei Wang from Bangor Business School. His paper, entitled "Investor Attention and FX Market Volatility", looks at the use of Google searches from a different angle, using them to measure investor attention. Attention can be modeled as a scarce resource, and this scarcity can affect the volatility of prices and portfolios. In particular, the theory predicts that market volatility increases with investor attention. In this paper, investor attention in the FX market is measured "actively" by the intensity of Google searches of abbreviations for exchange rate pairs. This proxy is added to GARCH models and is found to be significant in the variance equation of most currency pairs (with the exception of the GBP/USD) and, in the case of the JPY/USD, also in the mean equation.

Peter Reinhard Hansen, from the European University Institute, discussed both papers in the session. He noted that it would be interesting to know if the measure of Twitter bullishness is correlated with market volatility and suggested that the econometrics would have to be adjusted accordingly. Further, he thought it would be interesting to see whether bullishness predicts risk-adjusted returns. On the Google-based investor attention index, Peter suggested applying a different GARCH specification to which a realized volatility measure is added, as the standard GARCH tends to adjust very slowly to changes in volatility.

We adjourned for the morning of the second day with a delicious lunch which was served at the facilities of the European Central Bank. During lunch we enjoyed the second poster session entitled "Text mining" which included three papers. Qingyu Yuan, from the Graduate University of the Chinese Academy of Sciences, presented the paper entitled "Preprocessing method of internet search data for prediction improvement: application to the Chinese stock market", where he suggests a method for Internet search data pre-processing based on the Baidu keywords index. He and his co-authors find that adding this index to a prediction model for the Chinese stock market can reduce the mean absolute prediction error from 3.8% to 1.4%. Michela Nardo, from the European Commission, presented her paper titled "Differences in opinion make a market: web-based inference of stock prices and volumes for a subset of systemically important banks". She presented the results based on the Europe Media Monitor, which automatically scans the World Wide Web (every 10 minutes, 24h/day) and retrieves news containing the keywords in each alert (more than 70 languages). Using this source, the authors look at the relationship between web buzz and stock prices and volumes for a set of 6 banks (Barclays, HSBC, Deutsche Bank, BNP, Crédit Agricole, Royal Bank of Scotland). Each piece of news is scanned for tonality, and from the text metadata they derive 12 variables of web buzz, matching them with stock and volume data from various stock exchanges (New York, London, Frankfurt, Paris). They compare estimates from two autoregressive models with stocks with and without the web variable. They find that the web anticipates stock prices and, for some banks, also volumes and volatility. The reduction in average prediction error is between 7% and 16%. The third paper entitled "Big Data and Economic Forecasting: A Top-Down Approach Using Directed Algorithmic Text Analysis" was presented by Paul Ormerod, from the University College London. Paul and his co-authors look at the evolution of "relative sentiment" and "shifts" (using excitement and anxiety measures), using text analysis directed by a theory of individual behavior in the face of uncertainty to extract relative sentiment time series from broker research reports. Looking at the difference between two readings of the Michigan Confidence Index, they find that the forecasts using their model and the relative sentiment from broker reports are better than consensus forecasts published in Reuters at capturing the sign.

After lunch, we had the last session of the second day, session 7, entitled "Network modeling for big data" where two papers were presented. Christian Brownlees, from Universitat Pompeu Fabra, presented his paper entitled "NETS: Network Estimation for Time Series". Christian and his co-author suggest new network estimation techniques for high-dimension sparse multivariate time series using graph theory to characterize the correlation structure. The main contribution of the paper is that it proposes a model that can also be applied in the presence of serial dependence, using long-run partial correlations. The NETS network estimation algorithm is based on a two-step LASSO procedure. The paper also establishes conditions for consistent estimation and illustrates the procedure on a financial application, using a panel of monthly equity returns. The resulting "network of idiosyncratic returns" is then analyzed using typical network analysis tools, such as similarity, centrality and clustering ("small-world effects").

Andreea Minca, from Cornell University, presented the second paper of the session entitled "Networks of Common Asset Holdings: Aggregation and Measures of Vulnerability". Andreea presents measures of financial distress that originate in illiquidity, where contagion is mediated by price feedback effects. The analysis uses a weighted network in which portfolios are nodes and the weight of links between them is based on a model of the effect of liquidation of one fund on the price of the other. The proposed measure of vulnerability is the fraction by which the value of one portfolio will decrease if all its neighbors liquidate their portfolios by a given factor. Using mutual fund data, she finds that the vulnerability index is useful in predicting returns in periods of mass liquidations, when one can identify vulnerable funds based on asset holdings and the liquidity characteristics of the stock.

Galo Nuno, from ECB and Banco de España, discussed both papers in the session. In his discussion**,** Galo highlighted the complementarity of the two papers: in the NETS approach, less information is needed to estimate the network (which is stationary) and traditional indicators such as clustering and network centrality can be used to characterize vulnerability, while in the second approach more information is needed, the network varies in time and it is used to construct novel indicators of vulnerability. The complementarity lies especially in the kind of question that each approach tries to answer: individually, which institutions or portfolios or assets are more vulnerable, and at the aggregate level, how systemic vulnerability evolves over time.

The second day was adjourned in the afternoon after a panel discussion entitled “Big data initiatives – challenges and opportunities for central bankers”. In this panel discussion five panelists discussed their views of the opportunities and challenges of using big data from different angles: Tobias Preis from Warwick Business School, Kenneth Cukier from The Economist, Jojy Mathew from Deloitte Touche Tohmatsu Limited, Michail Skaliotis from Eurostat and Per Nymand-Andersen from the ECB. Tobias presented the usefulness of big data for predicting stock market movements using Google search volume data and indicated that certain keywords may be significantly correlated with stock market returns. Kenneth focused on the usefulness and innovations of big data and indicated that big data will give rise not only to new economic indicators but also to multiple gadgets supporting consumers using “location” information combined with an array of public sources and pattern data. Michail asked what will change in terms of official statistics. He also brought up the Fundamental Principles of Official Statistics and encouraged the audience to consider the use of big data if it helps to provide better statistics. Per elaborated on the core fundamental and methodological points related to the challenges, opportunities and policy measures needed for using big data for central banking purposes.

Chiara Osbat
Gianni Amisano
Per Nymand-Andersen
Juri Marcucci