Forecasting with Big Data: A Review*

Hossein Hassani[1,2] and Emmanuel Sirimal Silva[1]

[1] *Statistical Research Centre, The Business School, Bournemouth University, UK*

[2] *Institute for International Energy Studies (IIES), 1967743711, Tehran, I.R., Iran*

* The final version of this article has been published in Annals of Data Science, http://dx.doi.org/10.1007/s40745-015-0029-9

Abstract:

Big Data is a revolutionary phenomenon which is one of the most frequently discussed topics in the modern age, and is expected to remain so in the foreseeable future. In this paper we present a comprehensive review on the use of Big Data for forecasting by identifying and reviewing the problems, potential, challenges and most importantly the related applications. Skills, hardware and software, algorithm architecture, statistical significance, the signal to noise ratio and the nature of Big Data itself are identified as the major challenges which are hindering the process of obtaining meaningful forecasts from Big Data. The review finds that at present, the fields of Economics, Energy and Population Dynamics have been the major exploiters of Big Data forecasting whilst Factor models, Bayesian models and Neural Networks are the most common tools adopted for forecasting with Big Data.

Keywords: Big Data; forecasting; technique; statistics.

1 Introduction

The Big Data phenomenon has revolutionized the modern world, and is now the hottest Data Mining topic according to polls conducted by kdnuggets.com, with the current trend expected to continue into the foreseeable future. At present there is no unified definition of Big Data, however Shi (2014) presented two definitions for Big Data. For academics, Big Data is “a collection of data with complexity, diversity, heterogeneity, and high potential value that are difficult to process and analyze in reasonable time”, whilst for policy makers, Big Data is “a new type of strategic resource in the digital era and the key factor to drive innovation, which is changing the way of humans’ current production and living” (Shi 2014, p. 6). As Varian (2014, p. 24) accurately asserts, “Big data will only get bigger”. This increased availability of data, which has been further escalated through the evolution of Big Data is now a major concern for a large number of industries (Hassani et al., 2014). It is not entirely surprising that this increasing availability of data is causing anxiety, and this is evident through the sound example presented by Smolan and Erwitt (2012) where the authors state that in the modern age, we generate 70 times the information stored in the library of congress simply within the first 24 hours of a new born baby's life. In another report, it is noted that South Korea is currently upgrading its data storage capabilities associated with its national weather information system by increasing the capacity to 9.3 petabytes (Hamm, 2013) and these examples provide an indication of the rate at which Big Data continues to grow. The times have truly 'changed' and we now live in an age where Big Data is identified as the leading edge for innovation, competition and productivity (Manyika et al., 2011). For example, digital data was expected to grow from 161 exabytes in 2006 to 2837 extabytes in 2012 and is now forecasted to reach 40 trillion gigabytes in 2020 (Gantz and Reinsel, 2012). Moreover, in the year 2008 alone the world had produced 14.7 exabytes of new data (Bounie, 2012).

The emergence of Big Data is now history. What is of importance is how organizations develop the tools and means necessary for reacting to, and exploiting the increasingly available Big Data for their advantage. In line with this, Varian (2014) notes that there is a need for the adoption of powerful tools such as Data Mining techniques which can aid in modelling the complex relationships that are inherent in Big Data. Moreover, it is noteworthy that the recent financial crisis has seen an increase in the prolific importance of risk management in organizations, and as Silva (2013) states, companies are now seeking to use risk management as a tool for maximising their opportunities whilst minimising the associated threats. Herein lays the opportunity, as Big Data forecasting has the ability to improve organizational performance whilst enabling better risk management (Brown et al., 2011). As Bernstein (2013) states, Big Data and predictive analysis goes hand in hand in the modern age with companies focussing on obtaining real time forecasts using the increasingly available data.

However, not all authors agree that Big Data is a revolutionary phenomenon. Poynter (2013) states Big Data will be more insightful in simply connecting the dots as opposed to painting a whole new picture. For Walker (2014), 2013 was the year for getting accustomed to Big Data and 2014 is the year for truly exploiting Big Data towards lucrative gains. We share and subscribe to Walker's (2014) perception on Big Data. Accordingly, we present this review paper which aims to: provide an informative review of the forecasting techniques utilized for forecasting with Big Data; provide a concise summary of the contributions of yesteryear; and identify challenges which need be overcome as the world gears to embrace and live in the presence of Big Data. In the process, we are successful in reviewing a wide range of forecasting models which have been adopted for forecasting with Big Data. In order to enable the reader to have a clear understanding of the history, we present the review of applications differentiated by the relevant field (i.e., economics, finance, and energy among others) and topic where appropriate. Those interested in tools which can be used for manipulating Big Data are referred to Varian (2014, Table 1, p. 5)

The remainder of this paper is organized as follows. In the following section we discuss the problems and potential behind Big Data forecasts whilst the associated challenges are discussed in Section 3. Section 4 provides a review of statistical and data mining techniques that have been evaluated for the purpose of forecasting with Big Data and the paper ends with some conclusions in Section 5.

2 The 'Problem' and 'Potential' of Big Data Forecasting

There exists a widespread belief that Big Data can aid in improving forecasts provided that we can analyse and discover hidden patterns, and Richards and King (2013) agree that predictions can be improved through data driven decision making. Tucker (2013) believes Big Data will soon be predicting our every move, and according to Einav and Levin (2013), Big Data is most commonly sought after for building predictive models in a world where forecasting continues to remain a vital statistical problem (Hand 2009). We then come to the question, what is the problem behind forecasting with Big Data? The simplest explanation is that the traditional forecasting tools cannot handle the size, speed and complexity inherent in Big Data (Madden 2012). According to Arribas-Bel (2013) this is owing to the lack of a structure in these data sets and the size. As a result, traditional techniques are seldom preferred for tackling Big Data (Arribas-Bel, 2013). Therefore, forecasting Big Data poses a challenge to organizations, and this is further highlighted by the European Central Bank which is conducting an entire workshop dedicated to using Big Data for forecasting[1]. Laney (2001) is first to discuss the importance of data volume, velocity and variety in the context of Big Data. A decade later, Dumbill (2012), Press (2013), and Shi (2014) identifies these '3V's' as three concepts which define the dimensions of Big Data.

Rey and Wells (2013) believe Data Mining techniques can be exploited to help forecasting with Big Data, a view supported by Varian (2014). However, it should be noted that in the past, Data Mining techniques have mainly been used on static data as opposed to time series (see for example Berry (2000); Pyle (2003); Krugan and Musilek (2006); Han et al. (2012)). Interestingly, Cukier (2010) finds fault with Big Data for the recent financial crisis as he believes the financial models adopted were unable to handle the huge amounts of data that was being inputted into the systems, and thereby resulted in inaccurate forecasts.

The opportunities for gains through forecasting with Big Data are diverse. At present, there is increased research into using Big Data for obtaining accurate

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

weather forecasts and the initial results suggests that Big Data will benefit weather forecasts immensely (Knapp 2013; Hamm 2013). In fact, weather forecasting has been one of the main beneficiaries of Big Data, but the forecasts are still inaccurate beyond a week (Silver 2013). The fashion industry too is exploiting Big Data forecasts with companies such as EDITD (http://editd.com/) using Big Data for forecasting the fashion future by collecting data from social media (Kansara, 2011). According to Bacon (2013), the airline industry is yet another field where Big Data forecasting is crucial. An interesting success story behind forecasting with Big Data is Netflix and its use of Big Data forecasts for decision making prior to commencing production of its own TV show ‘House of Cards’, and this resulted in increased revenue for the company (KRWG, 2013). The potential underlying Big Data forecasts is truly astonishing and at times ‘scary’ as was evident in the experience of an individual in a story narrated by Duhigg (2012) where an irate customer walks into a ‘Target’ store in Minneapolis to complain about the store sending coupons relating to pregnancy products to his high school daughter. A few weeks later the same customer apologizes to the manager as following a discussion with his daughter it was revealed that she was in fact pregnant (Duhigg, 2012).

3 Challenges for Forecasting with Big Data

In this section we focus mainly on the challenges which need be overcome when forecasting with Big Data. It is imperative to note that the availability of Big Data alone does not constitute the end of problems (Bacon, 2013). A good example is the existence of a vast amount of data on earthquakes, but the lack of a reliable model that can accurately predict earthquakes (Silver, 2013). Some existing challenges are related to hypothesis, testing and models utilized for Big Data forecasting (Silver (2013); Poynter (2013)) whilst West (2013) identifies as an added concern, the lack of theory to complement Big Data. Besides these, we have identified the following varied challenges associated with forecasting Big Data that needs to be given due consideration.

3.1 Skills

The skills required for tackling the problem of forecasting with Big Data, and the availability of personnel skilled for this particular task is one of the foremost challenges. As Arribas-Bel (2013) states, the advanced skills required to handle Big Data is a major challenge whilst Poynter (2013) notes there is a short supply of data scientists equipped with the skills required to tackle Big Data. Thornton (2013) also agrees that there is a shortage of people who can understand Big Data. In a world where academics, researchers and statisticians are highly experienced in using traditional statistical techniques for over fifty years to obtain accurate forecasts, the availability of Big Data in itself is challenging. Skupin and Agarwal (2007) states that the inaptness of traditional statistical techniques which are meant for obtaining forecasts from traditional data are hindering the effectiveness and application of forecasts from Big Data and Arribas-Bel (2013) shares this same concern. As majority of statisticians are experienced in these traditional techniques, Einav and Levin (2013) points out that it is a challenge to develop the required skills for Big Data forecasting. In order to overcome this issue, it is important that Higher Educational Institutes around the globe give due consideration to upgrading the educational syllabuses to incorporate the skills necessary for understanding, analysing, evaluating and forecasting with Big Data so that the next generation of statisticians will be well equipped with the mandatory skills.

3.2 Signal and Noise

A more technical, but extremely important challenge in Big Data forecasting is identified by Silver (2013). He suggests that noise is distorting the signal in Big Data, and that there is an increasing noise to signal ratio visible in Big Data. Silver's (2013) notion is further confirmed via Bańbura and Modugno (2014) who points out that with large data sets, extracting the signal is made more complex. A majority of traditional forecasting techniques forecast both the noise and signal, and whilst they perform relatively well in the case of traditional data sets, the increasing noise to signal ratio seen in Big Data is more likely to distort the accuracy of forecasts. This suggests that there is a need for employing and evaluating the use of forecasting techniques which can filter the noise in Big Data and forecast the signal alone. A sound example of a filtering technique is Singular Spectrum Analysis (SSA) which seeks to filter the noise from a given time series, reconstruct a new series which is less noisy, and then use this newly reconstructed series for forecasting future data points. The superiority of the methodology of SSA over traditional techniques has been proven recently in a variety of fields where the data sets have comparatively smaller signal to noise ratios in relation to the considerably higher signal to noise ratios expected in Big Data (see for example, Hassani et al. (2009;2013;2015), Beneki and Silva (2013), Silva (2013) and Silva and Hassani (2015)). Future research should concentrate on evaluating the applicability of such techniques for filtering the noise in Big Data to enable accurate and meaningful forecasts.

3.3 Hardware and Software

Arribas-Bel (2013) was of the view that current statistical software is not able to tackle Big Data forecasting whilst Needham (2013) notes the possible need for supercomputers to handle Big Data forecasts. Recently, Hydman and Athanasopoulos (2013) have developed automatic forecasting techniques which can provide output within a matter of seconds. However, their reliability in the face of Big Data is yet to be tested. Another issue directly related to hardware and software is that personally, we have experienced statistical programs crashing in the face of few thousand observations owing to deficiencies in random access memory (RAM) or the associated software. As such it is prudent to agree that computing capabilities and the structure underlying statistical software will require enhancements in order to be able to successfully handle the increased data input.

3.4 Statistical Significance

Lohr (2012) suggests there is an increased threat of making false discoveries from Big Data. This is because even though obtaining forecasts using an appropriate technique appears to be the major challenge, it is not quite so. Given the sheer quantity of data that needs to be processed and forecasted, with Big Data there is an increased complexity in differentiating between randomness and statistically significant outcomes (Efron 2010). As such there is an increased probability of reporting a chance occurrence as a statistically significant outcome and misleading the stakeholders interested in the forecast.

3.5 Architecture of Algorithms

Data Mining techniques are suggested as important methods which could be used for forecasting with Big Data. However, these techniques have been designed to handle data of comparatively smaller sizes as opposed to the size of Big Data. Therefore, Data Mining algorithms are often unable to work with data that is not loaded on to its main memory, and thus requires the movement of Big Data between locations which can incur increased network communication costs (Jadhav 2013). The architecture of the analytics needs to be redesigned so that it could handle both historical and real time data (Jadhav 2013), and the Lambda architecture proposed in Marz and Warren (2013) is a good example of research currently seeking to overcome this issue. A detailed evaluation of challenges associated with the application of Data Mining techniques to Big Data (explained in the context of official statistics) can be found in Hassani et al. (2014).

3.6 Big Data

Big Data itself is a challenge for forecasting as a result of its inherent characteristics. Firstly, Big Data evolves and changes in real time, and as such it is important that the techniques used to forecast Big Data are able to transform unstructured data into structured data (Shi 2014), accurately capture these dynamic changes and detect change points in advance. Secondly, there are challenges stemming from Big Data’s highly complex structure and as Einav and Levin (2013) point out, it is a challenge to build forecasting models that do not result in poor out-of-sample forecasts owing to the ‘over use’ of potential predictors. Factor modelling which is discussed in the following section is a potential cure for this challenge, but more devoted research is needed to overcome the issue completely.

4 Applications of Statistical and Data Mining Techniques for Big Data Forecasting

In this section we identify existing applications of statistical and Data Mining techniques for forecasting with Big Data. We have summarized these based on the related field and topic (where relevant) in order to provide the reader with a more rewarding experience. At the outset, it is noteworthy to mention that Forni et al. (2000), and Stock and Watson (2002) are closely associated with the developments of econometric techniques for analysis and forecasting with Big Data.

4.1 Forecasting with Big Data in Economics

Researchers in the field of Economics have been major exploiters of Big Data for forecasting various economic variables. Camacho and Sancho (2003) used a Dynamic Factor Model (DFM) based on the methodology presented in Stock and Watson (2002) to forecast a large dataset involving Spanish diffusion indexes which they describe as an exhaustive description of the Spanish economy. DFM models are an extension of Stock and Watson's (2002) factor models and are frequently used for forecasting with Big Data. However, Diebold (2003) asserted that the use of DFM for macroeconomic Big Data forecasting is flawed as it is based on linear models, and also as Big Data is more likely to be nonlinear. Over time through the work of Stock and Watson (2003), Forni et al. (2005) and Kapetanios (2009), the DFM technique was improved, thus enabling it to handle Big Data more appropriately.

The application of Maximum Likelihood estimates of Factor Models for Big Data forecasting has been evaluated by Doz et al. (2012) via a simulation study where the authors find this approach to be effective and efficient. A seasonal AR model was used to show how Big Data from the Google search engine can be used to predict economic indicators in Choi and Varian (2012). Gupta et al. (2013) used a multivariate factor-augmented Bayesian shrinkage model on Big Data comprising of 143 monthly time series to forecast employment in eight sectors of the U.S. economy. Big Data relating to various exchange rates are used to forecast the Euro, British Pound and Japanese Yen in Banerjee et al. (2013) where the authors find their proposed Factor-augmented Error Correction Model (FECM) outperforming a Factor-augmented VAR (FAVAR) model at accurately predicting the three major bilateral exchange rates. Bańbura et al. (2014) proposed an algorithm based on Kalman filtering for large VAR and DFM models to enable obtaining conditional forecasts, and providing a scenario analysis for the European economy using 26 macroeconomic and financial indicators for the Euro area.

In what follows, we further group the applications of Data Mining and statistical techniques for forecasting with Big Data in the field of Economics into topics based on the use of Big Data to forecast economic variables.

- ***Gross Domestic Product (GDP)***

Schumacher (2007) evaluated the forecasting performance of Factor models using Static and Dynamic Principal Components and Subspace algorithms for State Space models. He finds Factor models outperforming AR models at forecasting a large panel of quarterly time series relating to German GDP. Moreover, the Subspace Factor Model and Dynamic Principal Component model is seen outperforming the Static Factor Model, but this ranking depends greatly on the correct specification of the model parameters (Schumacher 2007). A large factor model which uses an Expectation Maximization (EM) algorithm combined with Principal Components is adopted in Schumacher and Breitung (2008) to forecast a large dataset comprising of German real-time GDP. They find the Mixed Frequency Factor model performing better than simple benchmark models but find meagre differences in forecast accuracy between the Factor models themselves. Biau and D'Elia (2009) apply the ensemble machine learning technique of Random Forests to forecast European Union GDP using large survey datasets. They find Random Forests outperforming the AR model and forecasts from the 'Euro zone economic outlook'. Biau and D'Elia (2009) also note that Random Forests are popular for its ability of not over-fitting when handling a large number of inputs. Altissimo et al. (2010) use monthly accumulated Big Data along with a modified DFM to forecast medium-to long run GDP growth in the Euro area and finds their model is able to perform better than Bandpass Filter in terms of fitting and change point detection. Carriero et al. (2012a) adopted a Bayesian Mixed Frequency model in combination with stochastic volatility for nowcasting with Big Data to obtain real time GDP predictions for the United States. Banerjee et al. (2013) used 90 monthly time series for the German economy and showed that a FECM can outperform a FAVAR model at forecasting real GDP in Germany. Kopoin et al. (2013) used factor models with national and international Big Data for improving the accuracy of GDP forecasts for Canadian provinces below the one year ahead mark. Beyond the one year ahead horizon they find that relying on the provincial data alone optimizes the forecasts. Bańbura and Modugno (2014) use factor models with maximum likelihood estimation on over 101 series for nowcasting GDP in the Euro area. They find that sectoral information is not mandatory for obtaining accurate GDP predictions in the Euro area.

- ***Monetary Policy***

In Bernanke et al. (2005), a Factor-augmented Vector Autoregressive (FAVAR) model was used for Big Data forecasting and structural analysis in order to accurately identify the monetary policy transmission mechanism so that the exact impact of monetary policy on the economy could be ascertained. They find the proposed FAVAR model outperforming the Structural VAR model by exploiting far more informative content for assessing the monetary policy transmission mechanism. De Mol et al. (2008) use a Bayesian regression model with macroeconomic Big Data which includes real and nominal variables, asset prices, surveys and yield curves for forecasting the industrial production and consumer price indices. They find the results from the Bayesian regression are highly correlated with forecasts from principal components. Alessi et al. (2009) exploits a monthly panel dataset comprising of 130 U.S. macroeconomic time series and four price indexes (PCE, PCE core, CPI and CPI core) for forecasting inflation and its volatility using a DFM in combination with multivariate GARCH models (DF-GARCH). They find the DF-GARCH model outperforming GARCH, AR*(p)* and AR*(p)*-GARCH(1,1) and other univariate and classical factor models. The work of De Mol et al. (2008) was extended in Bańbura et al. (2010) to show that combining VAR with Bayesian Shrinkage can improve forecasts. The authors conclude that Bayesian VAR models are appropriate for Big panel Data. Bordoloi et al. (2010) developed a DFM to forecast India's industrial production and price level, and cited the DFM model's ability to handle many variables found in Big Data as the reason behind its selection. Here, they find the DFM model outperforming an Ordinary Least Squares (OLS) model. Figueiredo (2010) exploits 368 monthly time series which include a variety of economic variables alongside a Factor model with Targeted Predictors (FTP) for forecasting Brazilian inflation. They find the FTP outperforming VAR, Bayesian VAR and a Principal Component based Factor model at forecasting Brazilian inflation. Carriero et al. (2011) considered Big Data relating to 52 U.S. macroeconomic time series taken from Stock and Watson (2006) along with a Bayesian Reduced Rank multivariate model for forecasting industrial production and consumer price indices and the federal funds rate. Their results are compared against models based on Rank Reduction, which include Bayesian VAR models, Multivariate Boosting and the Factor model from Stock and Watson (2002). They find that combining Rank Reduction with Shrinkage can improve forecasts attained when applied to Big Data. Giovanelli (2012) proposes the use of Kernel Principal Component Analysis (PCA) (as this enables factors to take a nonlinear relationship to the input variables) and an Artificial Neural Networks (ANN) model on Big Data containing 259 predictors for the Euro area, and 131 predictors for the U.S. economy for forecasting the industrial production and consumer prices indices. The author finds that using the Kernel PCA approach for predicting nonlinear factors yield results of better quality in comparison to the linear method, and that the ANN method reports a similar forecast to that obtained via the Factor Augmented Linear forecasting equation. In Carriero et al. (2012b) a large BVAR model coupled with optimised shrinkage towards univariate AR models are used to forecast interest rates. The authors find the BVAR model showing small gains over random walk forecasts. Banerjee et al. (2013) used a FECM and showed that FECM can outperform a FAVAR model at forecasting; U.S. inflation using a large panel of 132 U.S. macroeconomic variables, and German inflation and interest rate using 90 monthly series for the German economy. Ouysse (2013) compared Bayesian Model Averaging (BMA) and Principal Component Regression (PCR) on a large panel data set for forecasting U.S. inflation and industrial production. Based on the Root Mean Squared Error the author concludes that in general, PCR can provide more accurate forecasts than BMA. Koop (2013) exploits the U.S. macroeconomic data set found in Stock and Watson (2008) which includes 168 variables along with BVAR model for forecasting inflation and interest rates. They find that BVAR models can provide better forecasts than those attainable via Factor methods. Using the U.S. Treasury zero coupon yield curve estimates, Banerjee et al. (2013) showed that a FECM model can outperform a FAVAR model at forecasting interest rates at different maturities.

4.2 Forecasting with Big Data in Finance

Alessi et al. (2009) use their DF-GARCH model for forecasting financial asset returns using Big Data relating to transaction prices of stocks traded on the London Stock Exchange after cleaning the data for outliers. They find the DF-GARCH model outperforming a GARCH (1,1) model and that the full BEKK specification (Engle and Kroner 1995) provides better forecasts in comparison to the DCC specification (Engle 2002).

4.3 Forecasting with Big Data in Population Dynamics

An imputation based on Neural Networks model was applied to the Norwegian population census data of 1990 in order to perform a population census by combining administrative data along with data gathered through sample surveys (Nordbotten, 1996). A procedure based on Neural Networks was used by Frutos et al. (2003) to predict trends in Spanish economic indexes per household and censal section by using the Spanish Population and Housing Census, and Family Expenditure Survey. Bayesian regression was used by Paaβ and Kindermann (2003) for predicting long term illness in Stockport UK by using statistics from the 1991 census. Cluster Analysis was used as a method for predicting missing data by analysing the 2007 census donor pool screening in McCarthy et al. (2009). Unlikely representations of farming operations in the initial Census mail list have been predicted using Classification Trees according to Garber (2009). Gilary (2011) reports the US Census Bureau exploited the Decision Trees technique by combining a Stepwise regression with the Classification and Regression Tree (CART) concept for recursive portioning of racial classification cells. Moreover, there is evidence of Decision Trees being used to forecast survey non-respondents through the work of McCarthy et al. (2010).

4.4 Forecasting with Big Data in Crime

Wu et al. (2007) rely on a Kohonen Neural Network Clustering algorithm to find outliers and then forecast fraudulent behaviour in the data intensive Chinese telecom industry after evaluating its performance in comparison to a two-step Clustering algorithm and K-means algorithm.

4.5 Forecasting with Big Data in Energy

Wang (2013) uses Support Vector Machines as an auxiliary method, along with Neural Networks, and 'MapReduce' technology, for forecasting Big Data originating from China's electricity consumption. He finds the developed prediction model is able to provide sound portability and feasibility in terms of processing Big Data relating to electricity. Nguyen and Nabney (2010) evaluate the use of Wavelet Transform (WT) in combination with a variety of models such as GARCH, Linear regressions, Radial Basis functions and Multilayer Perceptrons (MLP) to forecast UK gas price and electricity demand by exploiting Big Data from the British energy markets. They find that the use of WT and adaptive models can provide considerable improvements to forecasting accuracy. The conclusion here is that adaptive models combining WT with either MLP or GARCH are the optimal models for forecasting gas price and electricity demand based on the lowest mean squared error. Fischer et al. (2013) evaluates the use of Exponential Smoothing and ARIMA models in combination with a model configuration advisor to forecast energy demand using Big Data from an energy domain.

4.6 Forecasting with Big Data in Environment

Sigrist et al. (2012) utilizes Stochastic Advection Diffusion Partial Differential Equations (SPDEs) to improve the precipitation forecasts for northern Switzerland using Big Data from a numerical weather prediction model. They find that following the application of SPDE, the forecasts are greatly improved in comparison to the raw forecasts attained via the numerical model.

4.7 Forecasting with Big Data in Biomedical Science

Lutz and Buhlmann (2006) provide theoretical evidence for the applicability of Multivariate Boosting for forecasting with Big Data. They propose a Multivariate L_2 Boosting method to be used with multivariate regression and can also be applied to a VAR series. An application to 795 Arabidopsis thaliana genes is used as an example to show the appropriateness of the proposed method.

4.8 Forecasting with Big Data in Media

Using data on hundreds and thousands of Youtube videos, Gursun et al. (2011) show an ARMA model with Singular Value Decomposition can be used for analyzing and forecasting video access patterns. They find that for rarely accessed videos Hierarchical Clustering can provide the better forecasts whilst for daily accessed videos PCA can provide an efficient forecast.

5 Conclusions

Big Data will continue to grow even bigger in the years to come, and if organizations are not inclined and willing to embrace the challenges, develop and employ the mandatory skills, they will find themselves in dire straits. In this review which is focussed on forecasting with Big Data, we have initially identified several problems and outlined the potential that Big Data has to offer and generate lucrative outcomes provided that we devote sufficient time and effort to overcome the identified issues. Thereafter we note a set of key challenges which at present hinder and impede the accuracy and effectiveness of Big Data forecasts.

In terms of the applications of statistical and Data Mining techniques for forecasting with Big Data, based on past literature it is evident that Factor models are the most common and popular tool currently used for Big Data forecasting whilst Neural Networks and Bayesian models are two other popular choices. The review also finds the field of Economics to be the most popular field in terms of exploitation of Big Data for forecasting variables of interest with the topics of GDP and Monetary Policy being the recipients of majority of the attention. The fields of Population Dynamics and Energy appear to be the second and third most popular based on published research. It is evident that there remains vast scope for research into forecasting with Big Data and that such work has the potential to yield better techniques which can enhance the forecasting accuracy. For example, it would be interesting to consider evaluating the use of a noise filtering technique such as Multivariate Singular Spectrum Analysis for forecasting with Big Data as this could aid in overcoming one of the major challenges at present which is the increased noise distorting the signal in Big Data.

In conclusion we wish to reinforce the necessity and responsibility of higher educational institutes to incorporate modules and courses which develop the skills required to be able to understand, analyze and forecast with Big Data using a variety of novel techniques. We believe that overcoming the constraints imposed by skills should be on top of the list for ensuring the increased application of relevant techniques for the exploitation and attainment of accurate and lucrative forecasts from Big Data in the future.

References:

Alessi, L., Barigozzi, M., and Capasso, M. (2009). Forecasting Large Datasets with Conditionally Heteroskedastic Dynamic Common Factors. *Working Paper No. 1115,* European Central Bank.

Altissimo, F., Cristadoro, R., Forni, M., Lippi, M., and Veronese, G. (2010). New Eurocoin: Tracking Economic Growth in Real Time. *The Review of Economics and Statistics,* **92**(4), pp. 1024-1034.

Arribas-Bel, D. (2013). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography,* forthcoming.

Bacon, T. (2013). Big Bang? When 'Big Data' gets too Big. Available via: http://www.eyefortravel.com/mobile-and-technology/big-bang-when-%E2%80%98big-data%E2%80%99-gets-too-big

Bańbura, M., Giannone, D., and Lenza, M. (2014). Conditional Forecasts and Scenario Analysis with Vector Autoregressions for Large Cross-Section. Working Papers ECARES ECARES 2014-15, ULB -- Universite Libre de Bruxelles.

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian Vector Autoregressions. *Journal of Applied Econometrics,* **25**(1), pp. 71-92.

Bańbura, M., and Modugno, M. (2014). Maximum Likelihood Estimation of Factor Models on Datasets with Arbitrary Pattern of Missing Data. *Journal of Applied Econometrics,* **29**(1), pp. 133-160.

Banerjee, A., Marcellino, M., and Masten, I. (2013). Forecasting with Factor-augmented Error Correction Models. *International Journal of Forecasting,* In Press.

Beneki, C., and Silva, E. S. (2013). Analysing and Forecasting European Union Energy Data. *International Journal of Energy and Statistics,* **1**(2), pp. 127-141.

Bernanke, B., Boivin, J., and Eliasz, P. S. (2005). Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive Approach. *The Quarterly Journal of Economics,* **120**(1), pp. 387-422.

Bernstein, D. (2013). Big Data's Greatest Power: Predictive Analysis. Available via: http://www.equest.com/cartoons/cartoons-2013/big-datas-greatest-power-predictive-analytics/

Berry, M. (2000). *Data Mining Techniques and Algorithms.* John Wiley and Sons.

Biau, O., and D'Elia, A. (2009). Euro Area GDP Forecasting using Large Survey Datasets. A random forest approach. Availabile via: http://unstats.un.org/unsd/nationalaccount/workshops/2010/moscow/AC223-S73Bk4.PDF

Bordoloi, S., Biswas, D., Singh, S., Manna, U. K., and Saggar, S. (2010). Macroeconomic Forecasting Using Dynamic Factor Models. *Reserve Bank of India Occasional Papers,* **31**(2), pp. 69-83.

Bounie, D. (2012). International Production and Dissemination Information: Results, Methodological Issues, and Statistical Perspectives. *International Journal of Communication,* **6,** pp. 1001-1021.

Boyd, D., and Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication and Society,* **15**(5), 662-679.

Brown, B., Chui, M., and Manyika, J. (2011). Are You Ready for the Era of 'Big Data'? *In: McKinsey Quarterly, October 2011.* Available via: http://www.mckinsey.com/insights/strategy/are_you_ready_for_the_era_of_big_data

Camacho, M., and Sancho, I. (2003). Spanish Diffusion Indexes. *Spanish Economic Review,* **5**(3), pp. 173-203.

Carriero, A., Clark, T. E., and Marcellino, M. (2012a). Real-time Nowcasting with a Bayesian Mixed Frequency Model with Stochastic Volatility. *Working Paper,* No. 1227, Federal Reserve Bank of Cleveland.

Carriero, A., Kapetanios, G., and Marcellino, M. (2012b). Forecasting government bond yields with large Bayesian Vector Autoregressions. *Journal of Banking & Finance,* **36**(7), pp. 2026-2047.

Carriero, A., Kapetanios, G., and Marcellino, M. (2011). Forecasting Large Datasets with Bayesian Reduced Rank Multivariate Models. *Journal of Applied Econometrics,* **26**(5), pp. 735-761.

Choi, H., and Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record,* **88**(s1), pp. 2-9.

Cox, M., and Ellsworth, D. (1997). Application-Controlled Demand Paging for Out-of-Core Visualization. *In: IEEE 8th Conference on Visualization,* 18-24 October, 1997, Phoenix, AZ.

Cukier, K. (2010). Data, data everywhere. *The Economist.* Available via: http://www.economist.com/node/15557443

De Mol, C., Giannone, D., and Reichlin, L. (2008). Forecasting using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components? *Journal of Econometrics,* **146**(2), pp. 318-328.

Diebold, F. X. (2003). 'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting. In M. Dewatripont, L.P. Hansen and S.Turnovsky (Eds.), Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society. Cambridge: Cambridge University Press, 115-122.

Doz, C., Giannone, D., and Reichlin, L. (2012). A Quasi-Maximum Likelihood Approach for Large, Approximate Dynamic Factor Models. *The Review of Economics and Statistics,* **99**(4), pp. 1014-1024.

Dumbill, E. (2012). What is Big Data? An Introduction to the Big Data Landscape. Available via: http://strata.oreilly.com/2012/01/what-is-big-data.html

Duhigg, C. (2012). How Companies Learn Your Secrets. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Efron, B. (2010). *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction.* Cambridge University Press.

Einav, L., and Levin, J. D. (2013). The Data Revolution and Economic Analysis. *Working Paper No. 19035,* National Bureau of Economic Research.

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

Engle, R. F., and Kroner, K. (1995). Multivariate simultaneous GARCH. *Econometric Theory*, **11**(1), pp. 122-150.

Fan, W., and Bifet, A. (2012). Mining Big Data: Current Status and Forecast to the Future. *ACM SIGKDD Explorations,* **14**(2), pp. 1-5.

Figueiredo, F. M. R. (2010). Forecasting Brazilian Inflation using a Large Dataset. *Central Bank of Brazil,* Working Paper No. 228. Available via: http://www.bcb.gov.br/pec/wps/ingl/wps228.pdf

Fischer, U., Schildt, C., Hartmann, C., and Lehner, W. (2013). Forecasting the Data Cube: A Model Configuration Advisor for Multi-Dimensional Data Sets. *In: IEEE 29th International Conference on Data Engineering (ICDE),* 8-12 April 2013, Brisbane.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The Generalized Factor Model: Identification and Estimation. *Review of Economics and Statistics,* **82**(4), pp. 540-554.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting. *Journal of the American Statistical Association,* **100**(471), pp. 830-840.

Frutos, S., Menasalva, E., Montes, C., and Segovia, J. (2003). Calculating Economic Indexes per Household and Censal Section from Official Spanish Databases. *Intelligent Data Analysis,* **7**(6), pp.603-613.

Gantz, J., and Reinsel, D. (2012). The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in Far East. Available via: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

Garber, S.C. (2009). Census Mail List Trimming using SAS Data Mining. *In: RRD Research Report,* 9 May 2009 Washington DC.

Gilary, A. (2011). Recursive Partitioning for Racial Classification Cells. *In: Proceedings of the Survey Research Methods Section, American Statistical Association – Session 628: Survey Analysis and Issues with Data Quality.* 2011 Miami Beach, pp.2706-2720.

Giovanelli, A. (2012). Nonlinear Forecasting using Large Datasets: Evidence on US and Euro Area Economies. *CEIS Tor Vergata,* **10**(13), pp. 1-29. Available via: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2172399

Gupta, R., Kabundi, A., Miller, S., and Uwilingiye, J. (2013). Using Large Datasets to Forecast Sectoral Unemployment. *Statistical Methods & Applications,* forthcoming.

Gursun, G., Crovella, M., and Matta, I (2011). Describing and Forecasting Video Access Patterns. *In: INFOCOM '11: Proceedings of the 30th IEEE International Conference on Computer Communications,* IEEE, 2011. Available via: http://www.cs.bu.edu/techreports/pdf/2010-037-video-access-patterns.pdf

Hamm, S. (2013). How Big Data can Boost Weather Forecasting. Available via: http://readwrite.com/2013/02/28/how-big-data-can-boost-weather-forecasting#awesm=~ou64ZEaKe2HtUu

Han, J., Kamber, M., and Pie, J. (2012). *Data Mining: Concepts and Techniques.* Elsevier, Inc.

Hand, D. J. (2009). Mining the Past to Determine the Future: Problems and Possibilities. *International Journal of Forecasting,* **25**(3), pp. 441-451.

Hassani, H., Heravi, S., and Zhigljavsky, A. (2009). Forecasting European Industrial Production with Singular Spectrum Analysis. *International Journal of Forecasting,* **25**(1), pp. 103-118.

Hassani, H., Heravi, S., and Zhigljavsky, A. (2013). Forecasting UK Industrial Production with Multivariate Singular Spectrum Analysis. *Journal of Forecasting,* **32**(5), pp. 395-408.

Hassani, H., Saporta, G., and Silva, E. S. (2014). Data Mining and Official Statistics: The Past, The Present & The Future. *Big Data,* **2**(1), BD1-BD10.

Hassani, H., Webster, A., Silva, E. S., and Heravi, H. (2015). Forecasting U.S. Tourist Arrivals using Optimal Singular Spectrum Analysis. *Tourism Management,* **46**, 322-335.

Hyndman, R. J. and Athanasopoulos, G. (2013). *Forecasting: Principles and Practice.* Otexts, Australia.

Jadhav, D. K. (2013). Big Data: The New Challenges in Data Mining. *International Journal of Innovative Research in Computer Science & Technology,* **1**(2), pp. 39-42.

Kansara, V. A. (2011). The Long View| How Realtime Data is Reshaping the Fashion Business. Available via: http://www.businessoffashion.com/2011/08/the-long-view-how-realtime-data-is-reshaping-the-fashion-business.html

Kapetanios, G., and Marcellino, M. (2009). A Parametric Estimation Method for Dynamic Factor Models of Large Dimensions. *Journal of Time Series Analysis,* **30**(2), pp. 208-238.

Knapp, A. (2013). Forecasting the Weather with Big Data and the Fourth Dimension. Available via: http://www.forbes.com/sites/alexknapp/2013/06/13/forecasting-the-weather-with-big-data-and-the-fourth-dimension/2/

Koop, G. M. (2013). Forecasting with Medium and Large Bayesian VARs. *Journal of Applied Econometrics,* **28**(2), pp. 177-203.

Kopoin, A., Moran, K., and Paré, J. P. (2013). Forecasting regional GDP with factor models: how useful are national and international data? *Economics Letters*, **121**(2), pp. 267–270.

Kurgan, L., and Musilek, P. (2006). A Survey of Knowledge Discover and Data Mining Process Models. *The Knowledge Engineering Review,* **21**(1), pp. 1-24.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. Available via: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lohr, S. (2013). The Age of Big Data. Available via: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=1&

Lutz, R. W., and Buhlmann, P. (2006). Boosting for High Multivariate Responses in High-Dimensional Linear Regression. *Statistica Sinica,* **16**, pp. 471-494.

Madden, S. (2012). From Databases to Big Data. *IEEE Internet Computing,* **16**(3), pp. 4-6.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). Big data: the next frontier for innovation, competition, and productivity. *McKinsey Global Institute,* Available via: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Marz, N., and Warren, J. (2013). *Big Data: Principles and Best Practices of Scalable Reatime Data Systems.* Manning Publications.

McCarthy, J.S., Jacob, T., and Atkinson, D. (2009). Innovative Uses of Data Mining Techniques in the Production of Official Statistics. *In: UN Statistical Commission Session on Innovations in Official Statistics,* 20 February 2009 New York.

McCarthy, J. Jacob, T. and McCracken, A. (2010). Modeling NASS Survey Non-response using Classification Trees. *In: RDD Research Report,* November 2010 Washington DC, pp.1-28.

Needham, J. (2013). *Disruptive Possibilities: How Big Data Changes Everything.* O'Reilly Media. Available via: http://chimera.labs.oreilly.com/books/1234000000914/index.html

Nordbotten, S. (1996). Neural Network Imputation Applied to the Norwegian 1990 Population Census Data. *Journal of Official Statistics,* **12**(4), pp.385-401.

Nguyen, H. T., and Nabney, I. T. (2010). Short-term Electricity Demand and Gas Price Forecasts using Wavelet Transforms and Adaptive Models. *Energy,* **35**(9), pp. 3674-3685.

Ouysse, R. (2013). Forecasting using a Large Number of Predictors: Bayesian Model Averaging versus Principal Components Regression. *Australian School of Business Research Paper,* No. 2013ECON04, pp. 1-34. Available via: http://research.economics.unsw.edu.au/RePEc/papers/2013-04.pdf

Paaβ, G., and Kindermann, J. (2003). Bayesian Regression Mixtures of Experts for Geo-referenced Data. *Intelligent Data Analysis,* **7**(6), pp.567-582.

Poynter, R. (2013). Big Data Successes and Limitations: What Researchers and Marketers Need to Know. Available via: http://www.visioncritical.com/blog/big-data-successes-and-limitations

Press, G. (2013). A Very Short History of Big Data. Available via: http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/2/

Pyle, D. (2003). *Business Modeling and Data Mining.* Elsevier Science.

Rey, T., and Wells, C. (2013). Integrating Data Mining and Forecasting. *OR/MS Today,* **39**(6). Available via: https://www.informs.org/ORMS-Today/Public-Articles/December-Volume-39-Number-6/Integrating-data-mining-and-forecasting

Richards, N. M., and King, J. H. (2013). Three Paradoxes of Big Data. *Stanford Law Review Online,* **66**(41), pp. 41-46.

Schumacher, C. (2007). Forecasting German GDP using Alternative Factor Models Based on Large Datasets. *Journal of Forecasting,* **26**(4), pp. 271-302.

Schumacher, C., and Breitung, J. (2008). Real-time Forecasting of German GDP based on a Large Factor Model with Monthly and Quarterly Data. *International Journal of Forecasting,* **24**(3), pp. 386-398.

Shi, Y. (2014). Big Data: History, Current Status, and Challenges Going Forward. *The Bridge, The US National Academy of Engineering,* **44**(4), Winter 2014, pp. 6-11.

Sigrist, F., Kunsch, H. R., and Stahel, W. A. (2012). SPDE based modeling of large space-time data sets. Available via: http://arxiv.org/pdf/1204.6118v4.pdf

Silva, E. S. (2013). A Combination Forecast for Energy-related CO_2 Emissions in the United States. *International Journal of Energy and Statistics,* **1**(4), pp. 269-279.

Silva, E. S., Wu, Y., and Ojiako, U. (2013). Developing Risk Management as a Competitive Capability. *Strategic Change,* **22**(5-6), pp. 281-294.

Silva, E. S., and Hassani, H. (2015). On the Use of Singular Spectrum Analysis for Forecasting U.S. Trade Before, During and After the 2008 recession. *International Economics,* http://dx.doi.org/10.1016/j.inteco.2014.11.003.

Silver, N. (2012). *The Signal and the Noise: The Art and Science of Prediction.* Penguin Books, Australia.

Skupin, A., & Agarwal, P. (2007). Introduction: what is a self-organizing map? In P. Agarwal, & A. Skupin (Eds.), *Self-organizing maps: Applications in geographic information science,* Chichester, Sussex: John Wiley.

Smolan, R., and Erwitt, J. (2012). *The Human Face of Big Data.* Sterling Publishing. Available via: http://humanfaceofbigdata.com/

Stock, J. H., and Watson, M. W. (2002). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association,* **97**(460), pp. 1167-1179.

Stock, J. H., and Watson, M. W. (2006). Forecasting with many predictors. In *Handbook of Economic Forecasting*, Elliott, G., Granger, C. W. J., Timmermann, A. (eds). Elsevier: Amsterdam; 517–554.

Stock, J. H., and Watson, M. W. (2008). Forecasting in dynamic factor models subject to structural instability. In *The Methodology and Practice of Econometrics: A Festschrift in Honour of Professor David F. Hendry*, Castle J, Shephard N (eds). Oxford University Press: Oxford; 173–205.

Suthaharan, S. (2013). Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning. *In: Big Data Analytics Workshops in conjunction with ACM Sigmetrics,* 21 June 2013. Available via: http://www.sigmetrics.org/sigmetrics2013/bigdataanalytics/abstracts2013/bdaw2013_submission_4.pdf

Thornton, D. (2013). The Problem with Big Data. Available via: http://moneyweek.com/arria-nlg-the-problem-with-big-data/

Tucker, P. (2013). The Future is Not a Destination. Available via: http://www.slate.com/articles/technology/future_tense/2013/10/futurist_magazine_s_predictions_on_quantum_computing_big_data_and_more.html.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives,* **28**(2), pp. 3-28.

Wang, X. (2013). Electricity Consumption Forecasting in the Age of Big Data. *Telkomnika,* **11**(9), pp. 5262-5266.

Walker, A. (2014). Trends in Big Data: A Forecast for 2014. Available via: http://www.csc.com/big_data/publications/91710/105057-trends_in_big_data_a_forecast_for_2014

West, G. (2013). Big Data Needs a Big Theory to Go with It. Available via: http://www.scientificamerican.com/article/big-data-needs-big-theory/

Wu, S., Kang, N., and Yang, L. (2007). Fraudulent Behaviour Forecast in Telecom Industry Based on Data Mining Technology. *Communications of the IIMA,* **7**(4), pp. 1-6.

Corresponding Author:

Dr. Hossein Hassani,

Associate Professor in Statistics and Econometrics
Statistical Research Centre,
The Business School,
Bournemouth University, UK.

hhassani@bournemouth.ac.uk