International Journal of Energy and Statistics
Vol. 4, No. 4 (2016) 1650020 (11 pages)

DOI: 10.1142/S2335680416500204

Forecasting energy data with a time lag into the future and Google trends

Hossein Hassani*,‡ and Emmanuel Sirimal Silva†,§

*Institute for International Energy Studies
Tehran, Iran

†Fashion Business School, London College
of Fashion, University of the Arts London, UK
‡hassani.stat@gmail.com
§e.silva@fashion.arts.ac.uk

Received 22 October 2016
Revised 25 November 2016
Accepted 27 November 2016
Published 31 December 2016

This paper presents a new idea for a forecasting approach which seeks to exploit the information contained within US EIA energy forecasts and related Google trends data for generating a new and improved forecast. The novel forecasting approach can be exploited by using a multivariate system which can consider data with different series lengths and a time lag into the future. Using real historical data, an official forecast for the same variable, and Google Trends search data, we illustrate the possibility of generating a comparatively more accurate forecast for an energy-related variable. The accuracy of the newly generated forecasts are evaluated by comparing with the actual observations and the official forecast itself. We find that the novel forecasting idea can generate promising results which call for further in-depth research into developing and improving this multivariate forecasting approach.

Keywords: Forecastability; energy forecasts; official forecasts; Google trends; future time lagged data.

Nomenclature

EIA : Energy Information Administration.
$STEO$: Short Term Energy Outlook.
$MSSA$: Multivariate Singular Spectrum Analysis.
SVD : Singular Value Decomposition.
L : Window Length.
LRF : Linear Recurrent Formula.
US : United States.
HS : Hassani-Silva test.

$RMSE$: Root Mean Square Error.
$RRMSE$: Ratio of the RMSE.
DC : Direction of Change.

1. Introduction

Forecasting is an art which continues to be of primary importance for resource allocation and planning within the global energy sector. The importance of energy forecasts for planning and resource allocations are clearly evident in the variety of energy-related forecasts which are published by the US EIA via their STEO reports [1]. Whilst the practice of publishing monthly energy-related forecasts have been in existence over a long period, more recently, Google provided users with access to real-time data on online search queries in the form of Google Trends.[a] Google Trends are often found to be correlated with economic indicators and may be useful for short term economic prediction [2]. Moreover, Google Trends are recognized as excellent indicators of public concern and has the potential of being a useful quantitative measure of energy related events [3]. It is noteworthy that the emergence of Big Data such as Google Trends has resulted in an increased availability of real-time information which can be extremely useful for improving the accuracy of forecasts [4].

This paper looks at exploiting both the availability of forecasts for energy-related variables and the access to Google Trends data as we introduce a new forecasting approach within the broad field of time series analysis and forecasting. Accordingly, there are several key contributions which are noteworthy. First, the new multivariate forecasting approach that is introduced herewith is different to the forecast averaging and forecast combination approaches in time series. This is because here we seek to model and extract information from data with a time lag into the future. Second, the approach itself can be exploited by any multivariate forecasting technique which can model data; with different series lengths and a time lag into the future. Third, to the best of our knowledge the application which follows presents the results from the initial attempt at using the chosen multivariate forecasting tool in combination with Google Trends.

Energy-related data are usually affected by many factors including demand, supply, economics, policy, technology, and weather, in addition to noise levels, volatility and nonlinear patterns. Therefore, we have opted for a nonparametric multivariate signal processing technique that can capture the nonlinear pattern of noisy volatile data with different series lengths as the tool for forecasting with forecasts and Google Trends. The chosen multivariate forecasting technique is complemented with auxiliary information in the form of official forecasts (which effectively represents data with a time lag into the future) and Google Trends. The accuracy of the

[a]https://www.google.co.uk/trends/.

forecast achieved via the newly proposed approach is evaluated by comparing with actual data and the official forecast over the same period.

It should be noted that is not our intention or aim to claim that Google Trends can predict the future, and we subscribe to the views expressed in [2] where the authors suggest that Google Trends are more useful for contemporaneous forecasting or nowcasting. However, here we simply seek to present a new idea which can show the benefits of incorporating Google Trends in a multivariate process which considers data with a time lag into the future. Even though we consider an energy-related data set as an example, it is not our intention to indicate that we can forecast tomorrow's energy price. Instead, we seek to introduce this new concept for improving the accuracy of forecasts by exploiting the forecastability of official forecasts, and show that the inclusion of Google Trends data within such a framework can result in a forecast that is more accurate than an existing official forecast. The addition of more related information into the multivariate system can help improve the forecast further.

The remainder of this paper is organized such that Sec. 2 presents a summary of the proposed forecasting approach which is followed by an introduction to the data, metrics, and results following application in Sec. 3. The paper concludes in Sec. 4.

2. Forecasting with Official Forecasts and Google Trends

Let us begin by inputting data into the multivariate system as shown via Fig. 1. Here, the observations represented via $(y_1^{(1)}, \ldots, y_N^{(1)}) \equiv (y_1^{(3)}, \ldots, y_N^{(3)})$ as these represent the historical data for the variable of interest. Then, the observations within $(y_{N+1}^{(3)}, \ldots, y_{N+h}^{(3)})$ represents the official forecast and thus data which has a time lag into the future. The observations in $(y_1^{(2)}, \ldots, y_N^{(2)})$ in this example will represent data from Google Trends, but in general it could represent any auxiliary information and the multivariate system can incorporate additional variables in the modelling process.

In what follows, we seek to summarize the entire forecasting with official forecasts and Google Trends process concisely via the following steps.

(1) Consider three time series $Y_N^{(1)}$, $Y_N^{(2)}$ and $Y_{N+h}^{(3)}$ such that $Y_N^{(1)} = (y_1^{(1)}, \ldots, y_N^{(1)})$, $Y_N^{(2)} = (y_1^{(2)}, \ldots, y_N^{(2)})$, and $Y_{N+h}^{(3)} = (y_1^{(3)}, \ldots, y_{N+h}^{(3)})$ with an identical frequency. Here, $Y_N^{(1)}$ and $Y_N^{(2)}$ represents historical data for a variable of interest and Google Trends data respectively.

    Note: $Y_{N+h}^{(3)}$ represents $Y_N^{(1)}$ plus the h-step ahead official forecast for that same variable. The h-step ahead official forecast represents data with a time lag into the future.

(2) Call upon a multivariate system which can consider data with different series lengths and a time lag into the future, for forecasting with official forecasts and

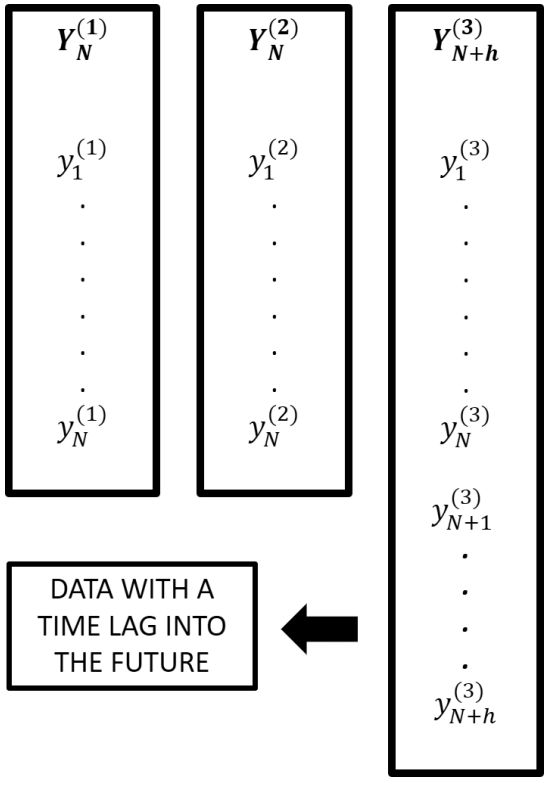

Fig. 1. A graphical illustration of the data input into the system.

Google Trends. Our aim is to obtain a multivariate h-step ahead forecast for $Y_N^{(1)} = (y_1^{(1)}, \ldots, y_N^{(1)})$ as represented by $\hat{y}_{N+1}^{(1)}, \ldots, \hat{y}_{N+h}^{(1)}$ in Fig. 2.

(3) Exploit a multivariate system's filtering and signal extraction capabilities for modelling and extracting information in $Y_{N+h}^{(3)}$ (which represents data with a time lag into the future) and $Y_N^{(2)}$, for generating a new and improved forecast for the variable in $Y_N^{(1)}$. Figure 2 summarises the process in graphical format.

(4) In this paper we exploit Multivariate Singular Spectrum Analysis (MSSA) as the multivariate system. Those interested in a detailed description of the theory underlying this nonparametric technique are referred to [5]. However, below we provide a very brief introduction into the MSSA process.

The MSSA technique consists of two stages known as decomposition and reconstruction. The decomposition stage has two steps known as embedding and SVD. Initially, through embedding we create the trajectory matrices $\mathbf{X}^{(i)}$ $(i = 1, 2, 3)$ of the one-dimensional time series $Y_N^{(1)}$, $Y_N^{(2)}$ and $Y_{N+h}^{(3)}$ respectively. Accordingly, we will have 3 different $L_i \times K_i$ trajectory matrices $\mathbf{X}^{(i)}$ $(i = 1, 2, 3)$, where $\mathbf{X}^{(1)}$ will

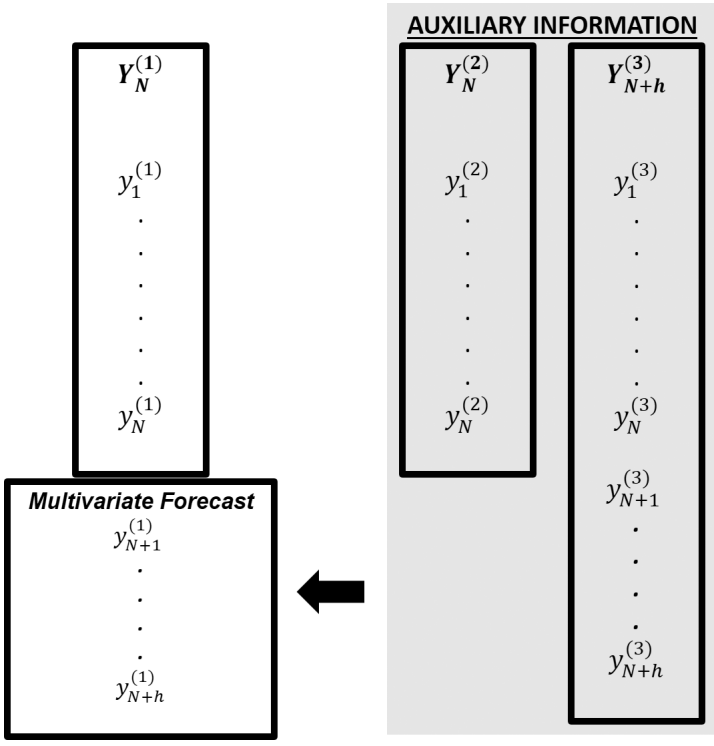

Fig. 2. A graphical illustration of the process and objective.

take the form:

$$\mathbf{X}^{(1)} = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & \cdots & y_K \\ y_2 & y_3 & \cdots & y_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & \cdots & y_N \end{pmatrix}. \tag{1}$$

A similar trajectory matrix as in Eq. (1) can be constructed for $\mathbf{X}^{(2)}$ to represent the data in $Y_N^{(2)}$. Finally, the trajectory matrix $\mathbf{X}^{(3)}$ which incorporates the official forecast can be constructed as:

$$\mathbf{X}^{(3)} = (x_{ij})_{i,j=1}^{L,K+h} = \begin{pmatrix} y_1 & y_2 & \cdots & y_K & y_{K+1} & \cdots & \omega_{K+h} \\ y_2 & y_3 & \cdots & y_{K+1} & y_{K+2} & \cdots & \omega_{K+h+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & \cdots & y_N & \omega_{N+1} & \cdots & \omega_{N+h} \end{pmatrix}, \tag{2}$$

where $\omega_1, \ldots, \omega_h$ represents the official forecast.

Thereafter, a new block Hankel trajectory matrix, $\mathbf{X}_H$ is constructed. Assume $L_1 = L_2 = \cdots = L_M = L$ where M is the number of time series. Therefore, we have different values of K_i ($K_i = N_i - L_i + 1$) and series length N_i, but similar L_i. The result of this step is:

$$\mathbf{X}_H = [\mathbf{X}^{(1)} : \mathbf{X}^{(2)} : \cdots : \mathbf{X}^{(M)}].$$

Hence, the structure of the matrix $\mathbf{X}_H\mathbf{X}_H^T$ is as follows:

$$\mathbf{X}_H\mathbf{X}_H^T = \mathbf{X}^{(1)}\mathbf{X}^{(1)^T} + \cdots + \mathbf{X}^{(M)}\mathbf{X}^{(M)^T}. \tag{3}$$

As it appears from the structure of the matrix $\mathbf{X}_H\mathbf{X}_H^T$ in MSSA, we do not have any cross-product between Hankel matrices $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$. Moreover, in this format, the sum of $\mathbf{X}^{(i)}\mathbf{X}^{(i)^T}$ provides the new block Hankel matrix. Note also that performing the SVD of $\mathbf{X}_H$ in MSSA yields L eigenvalues as in univariate SSA.

Thereafter, in the reconstruction stage we are faced with two steps known as grouping and diagonal averaging. Initially, we group the eigenvalues from the SVD process as either signal or noise (there are several approaches for grouping, see [5, 6]) and then perform diagonal averaging on the signal components to reconstruct a new, less noisy time series which can be used for forecasting. A more detailed description of the theory underlying decomposition and reconstruction with MSSA and the forecasting process can be found in [5] and is therefore not reproduced here.

Finally, we present the MSSA forecasting algorithm used in this paper, and in doing so we mainly follow [5].

(1) For a fixed value of L, construct the trajectory matrix $\mathbf{X}^{(i)} = [X_1^{(i)}, \ldots, X_K^{(i)}] = (x_{mn})_{m,n=1}^{L,K_i}$ for each single series $Y_{N_i}^{(i)}$ $(i = 1, \ldots, M)$ separately.
(2) Construct the block trajectory matrix $\mathbf{X}_H$ as follows:

$$\mathbf{X}_H = [\mathbf{X}^{(1)} : \mathbf{X}^{(2)} : \cdots : \mathbf{X}^{(M)}].$$

(3) Let vector $U_{H_j} = (u_{1j}, \ldots, u_{Lj})^T$, with length L, be the j^{th} eigenvector of $\mathbf{X}_H\mathbf{X}_H^T$.
(4) Consider $\widehat{\mathbf{X}}_H = \sum_{i=1}^{r} U_{H_i}U_{H_i}^T\mathbf{X}_H$ as the reconstructed matrix obtained using r eigentriples:

$$\mathbf{X}_H = \widehat{\mathbf{X}}^{(1)} : \widehat{\mathbf{X}}^{(2)} : \cdots : \widehat{\mathbf{X}}^{(M)}].$$

(5) Consider matrix $\widetilde{\mathbf{X}}^{(i)} = \mathcal{H}\widehat{\mathbf{X}}^{(i)}$ $(i = 1, \ldots, M)$ as the result of the Hankelization procedure of the matrix $\widehat{\mathbf{X}}^{(i)}$ obtained from the previous step.
(6) Let $U_{H_j}^{\triangledown}$ denote the vector of the first $L-1$ coordinates of the eigenvectors U_{H_j}, and π_{H_j} indicate the last coordinate of the eigenvectors U_{H_j} $(j = 1, \ldots, r)$.
(7) Define $\upsilon^2 = \sum_{j=1}^{r} \pi_{H_j}^2$.
(8) Denote the linear coefficients vector $\mathcal{R}$ as follows:

$$\mathcal{R} = \frac{1}{1-\upsilon^2}\sum_{j=1}^{r} \pi_{Hj}U_{Hj}^{\triangledown}. \tag{4}$$

(9) If $\upsilon^2 < 1$, then the h-step ahead MSSA forecasts exist and is calculated by the following formula:

$$\left[\hat{y}_{j_1}^{(1)}, \ldots, \hat{y}_{j_M}^{(M)}\right]^T = \begin{cases} [\tilde{y}_{j_1}^{(1)}, \ldots, \tilde{y}_{j_M}^{(M)}], & j_i = 1, \ldots, N_i, \\ \mathcal{R}^T \mathbf{Z}_h, & j_i = N_i + 1, \ldots, N_i + h, \end{cases} \tag{5}$$

where, $\mathbf{Z}_h = [Z_h^{(1)}, \ldots, Z_h^{(M)}]^T$ and $Z_h^{(i)} = [\hat{y}_{N_i - L + h + 1}^{(i)}, \ldots, \hat{y}_{N_i + h - 1}^{(i)}]$ $(i = 1, \ldots, M)$.

Note that Eq. (5) indicates that the h-step ahead forecasts of each series are achieved by the same LRF generated considering all series in a multivariate system.

3. Data, Metrics and Application

3.1. *Data*

This paper considers historical monthly data from January 2011–November 2015 as in-sample for model training and testing whilst the forecasting performance is evaluated over the period from December 2015–November 2016 (12 observations) such that it is a one-year ahead forecast. The variable being forecasted is the Henry Hub Spot price for Natural Gas (dollars per million Btu). The official monthly forecasts for the Henry Hub Spot price between December 2015–November 2016 as provided via the US EIA, and monthly Google Trends data for the search term 'Natural Gas Price' (January 2011–November 2015) are used as auxiliary information. The actual values corresponding to the EIA forecasts considered in this paper have been published in the December 2016 STEO report [1].

3.2. *Metrics*

3.2.1. *RMSE and RRMSE*

We rely on the RMSE and RRMSE criterions for evaluating forecast accuracy. Both these criterions are popular and frequently cited loss functions, see for example [8, 9].

$$\text{RMSE} = \left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right)^{\frac{1}{2}}, \tag{6}$$

where, Y_i is the actual value, $\hat{Y}_i$ refers to a forecast from a given model, and n is the number of the forecasts. Likewise, the ratio of the RMSE can be easily calculated as:

$$\text{RRMSE} = \frac{\text{RMSE}_{\text{Proposed Approach}}}{\text{RMSE}_{\text{Benchmark}}}, \tag{7}$$

where $\text{RMSE}_{\text{Proposed Approach}}$ refers to the RMSE from the proposed multivariate system and $\text{RMSE}_{\text{Benchmark}}$ refers to the RMSE for the official forecast. Then, if the RRMSE is less than 1 the proposed multivariate approach outperforms the benchmark model by 1-RRMSE%.

3.2.2. *Direction of change probability*

The application which follows considers a forecast for the gas price. When predicting such variables it is important that forecasts are not only able to report a low error as measured by the RMSE, but also successful in capturing or predicting the actual direction of change and movement in future prices. As such, we consider the DC criterion as a metric (see, [10, 11] for examples of previous applications) and present its calculation by following [10]. Let Z_{Y_i} take the value 1 if the forecast correctly predicts the direction of change and 0 otherwise. Then $\bar{Z_Y} = \sum_{i=1}^{n} Z_{Yi}/n$ shows the proportion of forecasts that correctly predict the direction of the series. For example, if DC is 0.70 this implies that 70% of the actual movement has been captured by the forecasting method.

3.3. *Application*

Let us now consider the performance of the proposed multivariate system at forecasting 12-months ahead for the Henry Hub Spot price, for which the results are reported in Table 1. All forecasts have been evaluated for statistically significant differences between competing forecasts via the HS test in [7]. We begin by focusing on the RMSE criterion. The first observation is that the MSSA models can provide forecasts with a lower RMSE than the EIA official forecast for the same variable. This implies that the proposed forecasting with official forecasts approach (MSSA | (EIA)) works in practice and can provide useful forecasting accuracy gains. In fact, by using only the official forecast provided by EIA as auxiliary information, we are able to generate a new forecast which is 6% more accurate than the EIA official forecast (as measured by the RRMSE: $1 - \frac{0.51}{0.54}$). Therefore, we can conclude that considering future information as auxiliary information aids in obtaining a better forecast. Moreover, it appears that data with a time lag into the future can help to capture the general pattern of a future forecast when combined with past data.

Next, let us consider using both the EIA forecast and Google Trends data as auxiliary information. Based on the RMSE values it is clear that MSSA | (EIA,GT) is able to outperform the EIA and MSSA | (EIA) forecasts for the Henry Hub Spot price. This in turn implies that the incorporation of Google Trends within the

Table 1. Henry Hub Spot forecasting RMSE and RRMSE results.

	RMSE			RRMSE	
Series	EIA	MSSA | (EIA)	MSSA | (EIA,GT)	$\frac{\text{MSSA} \mid (\text{EIA,GT})}{\text{EIA}}$	$\frac{\text{MSSA} \mid (\text{EIA,GT})}{\text{MSSA} \mid (\text{EIA})}$
Henry Hub Spot	0.54	0.51	0.45	0.83	0.88

Note: MSSA | (EIA) represents the MSSA results with only the official forecast as auxiliary information. MSSA | (EIA,GT) represents the MSSA results with both the official forecast and Google Trends as auxiliary information.

Table 2. Direction of change results for Henry Hub Spot forecasts.

Series	EIA	MSSA | (EIA)
Henry Hub Spot	0.58	0.75

Note: MSSA | (EIA) represents the MSSA model with only the official forecast as auxiliary information.

newly proposed forecasting with official forecasts framework can result in positive outcomes in terms of improved accuracy levels. As the MSSA | (EIA,GT) model reports the lowest RMSE, we consider this as the numerator in calculating the RRMSE values which are also reported via the last two columns in Table 1. Accordingly, we can see that the inclusion of GT within our framework has enabled us to produce a forecast which is 17% more accurate than the EIA official forecast. The MSSA | (EIA,GT) forecast is also 12% more accurate than the MSSA | (EIA) forecast. Our findings are in line with previous research which also indicated that Google Trends can help with improving energy market analysis [3, 12]. However, in this case we do not find any evidence of statistically significant differences between the MSSA | (EIA,GT) forecasts and competing forecasts, and this can be attributed to the low sample size of 12 observations.

Finally, let us now consider how well the official forecast and MSSA | (EIA) forecast perform in terms of capturing the movement in future prices by looking at the direction of change results. These are reported in Table 2. We find that the EIA official forecast reports a 58% accurate DC prediction. However, the MSSA | (EIA) forecast reports a DC prediction at 75% which is a 17% gain in relation to the official forecast. This clearly indicates yet another advantage in being able to exploit data with a time lag into the future as it enables the capturing of dynamic price changes which are likely to happen in future. Accordingly, these results show that the MSSA | (EIA) forecasts are more likely to accurately capture the movements in gas prices in comparison to the official forecast.

4. Conclusion

This paper presents a new idea for forecasting with official forecasts and improving accuracy levels further via the incorporation of auxiliary information in the form of Google Trends data. The aim being, to use the new forecasting idea to generate a forecast that can outperform the accuracy of the official forecast. The process is unique as it involves the modelling of data with a time lag into the future, which requires a multivariate system that can model data with different series lengths. The general idea is concisely presented to the reader in the most basic form, clearly outlining the nature of the suggested process. Thereafter, the proposed idea is put into action by applying the concept to real world data from the US EIA. We find promising results which not only show the applicability of the newly proposed

forecasting idea in practice, but also shows the positive influence of Google Trends on improving forecast accuracy.

Here we consider monthly data and a single energy variable as an example. However, the proposed forecasting idea is applicable to data from any given frequency and can be applied to forecast any given energy variable. Moreover, the idea itself can be exploited by any multivariate forecasting technique which can model data with different series lengths and a time lag into the future, even though we consider MSSA as a tool in this paper. In addition, it should be noted that users are not restricted to exploiting official forecasts. It is possible to consider and exploit information from forecasts generated via other time series models or professional forecasts as well.

We believe that the favourable results presented here open up a new research avenue in the field of time series analysis and forecasting. Future research should consider developing a more theoretically sound methodology for the proposed forecasting approach, optimization criteria for the tools used, and perform tests on a variety of different data sets. In terms of selecting the most appropriate search terms for matching with a variable of interest, researchers could consider Google Correlate [13] which can identify search patterns which correspond with real-world trends. Finally, there is huge scope and potential to incorporate Big Data within the proposed framework for forecasting with official forecasts and Google Trends — the possibilities are endless as there is no restriction on the number of variables one could input into the multivariate system.

References

[1] STEO Archives (2016). Available via: http://www.eia.gov/outlooks/steo/outlook.cfm. [Accessed: 24.11.2016].

[2] Choi, H. and Vairan, H. (2012). Predicting the Present with Google Trends. *Economic Records*, **88**(s1), 2–9.

[3] Ji, Q. and Guo, J.-F. (2015). Oil price volatility and oil-related events: An Internet concern study perspective. *Applied Energy*, **137**, 256–264.

[4] Hassani, H. and Silva, E. S. (2015). Forecasting with big data: A review. *Annals of Data Sciencie*, **2**(1), 5–19.

[5] Sanei, S. and Hassani, H. (2015). *Singular Spectrum Analysis of Biomedical Signals.* CRC Press.

[6] Hassani, H., Ghodsi, Z. and Silva, E. S. (2016). From nature to maths: Improving forecasting performance in subspace-based methods using genetics Colonial Theory. *Digital Signal Processing*, **51**, 101–109.

[7] Hassani, H. and Silva, E. S. (2015). A Kolmogorov-Smirnov based test for comparing the predictive accuracy of two sets of forecasts. *Econometrics*, **3**(3), 590–609.

[8] 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*, **141**, 34–49.

[9] Altavilla, C. and De Grauwe, P. (2010). Forecasting and combining competing models of exchange rate determination. *Applied Economics*, **42**(27), 3455–3480.

[10] Hassani, H., Heravi, S. and Zhigljavsky, A. (2013). Forecasting UK industrial production with multivariate Singular Spectrum Analysis. *Journal of Forecasting*, **32**(5), 395–408.
[11] Hassani, H., Webster, A., Silva, E. S. and Heravi, S. (2015). Forecasting U.S. Tourist arrivals using optimal Singular Spectrum Analysis. *Tourism Management*, **46**, 322–335.
[12] Li, X., Ma, J., Wang, S. and Zhang, X. (2015). How does Google search affect trader positions and crude oil prices? *Economic Modelling*, **49**, 162–171.
[13] Google Correlate. (2011). Available via: https://www.google.com/trends/correlate. [Accessed: 24.11.2016].