Communications for Statistical Applications and Methods
2022, Vol. 29, No. 1, 41–51

https://doi.org/10.29220/CSAM.2022.29.1.041
Print ISSN 2287-7843 / Online ISSN 2383-4757

How to improve oil consumption forecast using google trends from online big data?: the structured regularization methods for large vector autoregressive model

Ji-Eun Choi[a], Dong Wan Shin[1, b]

[a]Department of Statistics, Pukyong National University;
[b]Department of Statistics, Ewha Womans University

Abstract

We forecast the US oil consumption level taking advantage of google trends. The google trends are the search volumes of the specific search terms that people search on google. We focus on whether proper selection of google trend terms leads to an improvement in forecast performance for oil consumption. As the forecast models, we consider the least absolute shrinkage and selection operator (LASSO) regression and the structured regularization method for large vector autoregressive (VAR-L) model of Nicholson *et al.* (2017), which select automatically the google trend terms and the lags of the predictors. An out-of-sample forecast comparison reveals that reducing the high dimensional google trend data set to a low-dimensional data set by the LASSO and the VAR-L models produces better forecast performance for oil consumption compared to the frequently-used forecast models such as the autoregressive model, the autoregressive distributed lag model and the vector error correction model.

Keywords: oil consumption forecast, google trends, online big data, dimension reduction, the least absolute shrinkage and selection operator (LASSO)

1. Introduction

Oil is the primary source of energy that is consumed the most and oil consumption has had a significant impact on global economic system. Therefore, forecasting oil market has been important issue for investors, company executives, government policy-makers, financial risk managers and many others. It makes many studies have been conducted for forecasting oil markets (Zhang *et al.*, 2015; Baumeister and Kilian, 2015; Wen *et al.*, 2016; Zhao *et al.*, 2017; many others).

Recently, as one of the ways to improve forecast performance for the oil market, forecasts using google trend data have been conducted. Online big data, google trend data is the search interest relative to the highest interest for the specific search term, that people search on google, in the specific region for one month. Many studies demonstrated that the google trend data improves forecast performance not only for oil markets but also for other economic or financial data: Guo and Ji, 2013; Fantazzi and Fomicher, 2014; Li *et al.* (2015) for oil market; Carriere-Swallow and Labbe (2013) for now casting in emerging market; Bulut (2018) for exchange rate; Niesert *et al.* (2020) for unemployment and many others.

The authors are grateful of the reviewers whose comments clarify several points. This study was supported by a grant from the National Research Foundation of Korea (2019R1A2C1004679) and by the Pukyong National University Research Fund in 2021 (C-D-2021-0982).

[1] Corresponding author: Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: shindw@ewha.ac.kr

Published 31 January 2022 / journal homepage: http://csam.or.kr
© 2022 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

Yu *et al.* (2019) also found that some google trends improve forecast performance for global oil consumption level and grasped a cointegration relation between the oil consumption level (Y) and some of google trend terms (X). However, Yu *et al.* (2019) forecasted oil consumption using only three google trend terms among many google trend terms. While addressing the issue of cointegration, they neglected the issues of predictor selection and lag selection.

Note that there are many candidates for google trends for possible predictors of oil consumption level, for example "Fossil fuels", "Oil production" and many others. The number of such keywords can be larger than 10. To select proper google trends or to select proper lags of the predictors in a joint model for oil consumption level and the google trends, we consider the large vector autoregressive (VAR-L) model of Nicholson *et al.* (2017) based on least absolute shrinkage and selection operator (LASSO) regularization. The VAR-L model is suitable if either data dimension is large or lag order is large. The VAR-L models select automatically predictors and lags of the predictors, allowing us to avoid the problem of overparameterization for the VAR-L model. This regularization makes forecast performance be improved over the forecasts based on usual un-regularized model, as demonstrated by Smeekes and Wijler (2018) for macroeconomic forecasting, Sagaert *et al.* (2018) for tactical sales forecasting and Cepni *et al.* (2019) for GDP forecasting in emerging market economies. As additional forecast models, we consider an autoregressive distributed lag model (ADL-L) based on LASSO regularization of Li and Chen (2014) and a vector error correction model (VEC-L) extended from the VAR-L model to consider explicitly the cointegration relation.

Noting that there is a large number of candidates for google trends, for example 17 as tabulated in Table 1 below, we consider two VAR-L models. The first one is a 18-dimensional VAR-L model with all 17 google trends which select predictors and lags of predictors by regularization. The second one is a low-dimensional VAR-L model with google trends whose dimension is reduced by a LASSO regularization. The LASSO regularization is widely used as a variable selection method (Messner and Pinson (2019), Tarassow (2019) and many others).

Out-of-sample forecast comparison reveals that the VAR-L model has better forecast performance than the VAR, ADL-L and VEC-L models. The VAR-L model with reduced google trends has better forecasts than the VAR-L with all google trends. We observe that explicit consideration of the cointegration relationship does not help improve forecast performance. The good VAR-L forecast performance of oil consumption level is achieved by clever pre-selection of important google trend terms by LASSO prior to VAR-L fitting and proper lags for the differenced data by the VAR-L regularization.

The remaining of the paper is organized as follows. Section 2 gives a data description. Section 3 describes the forecast methods. Section 4 contains the results of out-of-sample comparisons. Section 5 provides a conclusion.

2. A data description

We forecast the monthly US oil consumption level using google trends as predictors. Google trend data sets are online big data set including all search terms that people search on google. The values of google trend data represent the search interest in the specific region for one month. The values are 100 for the most frequent searches and the value is 0, if there is not enough data for that query. The search interest is measured based on the search volumes of the search queries in google search across various regions and languages. The google trend data and the US oil consumption data are downloaded from Google trend website (https://trends.google.com/trends/?geo=US) and from the US EIA website (www.eia.gov), respectively.

Table 1: P-values of the ADF test for the oil consumption and google trend data sets

Non-stationary series		Stationary series	
	ADF p-value		ADF p-value
US oil consumption (O_t)	0.47		
Biodiesel (N_{1t})	0.78	Cost oil (S_{1t})	< 0.01
Oil (N_{2t})	0.33	Crude oil (S_{2t})	< 0.01
Oil consumption (N_{3t})	0.07	Gulf Mexico (S_{3t})	< 0.01
OPEC (N_{4t})	0.17	Iran oil (S_{4t})	< 0.01
Oil production (N_{5t})	0.09	Oil price (S_{5t})	0.02
Oil shock (N_{6t})	0.11	Fossil fuels (S_{6t})	0.04
Saudi arabia (N_{7t})	0.32	Middel east (S_{7t})	0.01
Venezuela oil (N_{8t})	0.32	Production gas (S_{8t})	< 0.01
		Shale oil (S_{9t})	< 0.01

Table 2: Test statistics of cointegration tests

	Residual ADF test	Johanson test
Biodiesel (N_{1t})	−2.09	19.09*
Oil (N_{2t})	−1.56	9.46
Oil consumption (N_{3t})	−1.98	27.65**
OPEC (N_{4t})	−1.72	35.03**
Oil production (N_{5t})	−3.83*	26.36**
Oil shock (N_{6t})	−1.59	43.29**
Saudi arabia (N_{7t})	−1.69	6.03
Venezuela oil (N_{8t})	−1.88	60.95**

Note: Significances are denoted by ** and * for 1% and 5% levels, respectively.

For forecasting oil consumption level, among many google search terms related to oil consumption, we consider 17 key search terms which are some of the most frequently searched terms (top popular search terms) or the terms having the most significant growth in volume (top rising search terms) searched together with 'Oil consumption' search term in the same search session as of December 3, 2019. The considered google search terms are 'Biodiesel', 'Cost oil', 'Crude oil', 'Fossil fuels', 'Gulf Mexico', 'Iran oil', 'Middle east', 'Oil', 'Oil consumption', 'Oil price', 'Oil production', 'Oil shock', 'OPEC', 'Production gas', 'Saudi arabia', 'Shale oil', 'Venezuela oil'. The period of the oil consumption and google trend data sets is considered as 01/01/2004-09/01/2019, $T = 189$ months. The variables are given Table 1.

An unit root analysis is made for the variables by the augmented Dickey Fuller (ADF) test as given in Table 1. The ADF test is conducted with Akaike information criterion (AIC) order. The monthly oil consumption level, say O_t, is a nonstationary series in the sense of having ADF p-value > 0.05. On the other hand, some google trends, denoted by N_{it}, $i = 1, \ldots, 8$, are nonstationary having p-value > 0.05 and the other google trends, denoted by S_{it}, $i = 1, \ldots, 9$ are stationary. In the VAR-L modeling, if no cointegration exists among O_t and N_{it}, $i = 1, \ldots, 8$, we need differencing for oil consumption level and for nonstationary google trend series $N_{1t}, \ldots, N_{8t}$; otherwise, we need to consider a modification of the VAR-L model to a vector error correction model, VEC-L say.

In order to see cointegration relations, the Johanson test and the Engle-Granger tests are employed for O_t and each element of $N_{1t}, \ldots, N_{8t}$. Test results are reported in Table 2. Both tests indicate cointegration between the google trend N_{5t} of 'Oil production' and the oil consumption level at 5% level. Accordingly, we will identify whether addressing cointegration relationship in (O_t, N_{5t}) brings an improvement in the forecast performance or not in Section 4.

3. Forecast methods

We forecast oil consumption level using various time series forecast models composed of some selected google trend terms. As the forecast models, we discuss three VAR-L, ADL-L, VEC-L models based on least absolute shrinkage and selection operator (LASSO) to specify their lags for which predictor selection methods are first discussed.

3.1. Predictor selection

For the VAR-L model in Section 3.2 below, we consider two methods for predictor selection. The first method is the automatic selection by the 18-dimensional VAR-L model which select the variables (and lags) automatically by the regularization. The second method is a two-step method in which predictor variables are the first selected by the method below of this subsection and then a lower-dimensional VAR-L model with selected variables is considered for forecasting. A forecast comparison for the two VAR-L models is provided in Section 4.2. For the ADL-L model in Section 3.3 below, predictors are selected by the second regularization method. Both of the VAR-L model and ADL-L model are applied to differenced data as indicated by the unit root analysis in Table 1, which ignores possible cointegration relations among oil consumption level and google trends. Such cointegration is addressed by a VEC-L model in Section 3.4 below in which predictor N_{5t} is selected by the cointegration analysis as in Table 3.

In the remaining of this subsection, we discuss the second two-step method. The predictor variables for the VAR-L or ADL-L models are chosen by LASSO regression. The unit root analysis in Table 1 leads us to consider difference for nonstationary series. Let the data set $\{(O_t, N_{1t}, \ldots, N_{8t}, S_{1t}, \ldots, S_{9t})', \ t = 1, \ldots, T\}$ be given. Recall that all elements of O_t, N_t are nonstationary. Let $S_{10t} = \Delta N_{1t}, \ldots, S_{17t} = \Delta N_{8t}$. Let $S_t = (S_{1t}, \ldots, S_{19t})'$. Note that all elements of S_t are stationary. Now, the VAR-L or ADL-L models are applied to the differenciated data set $\{(\Delta O_t, S_t), t = 2, \ldots, T\}$.

Let h be a given forecast step from 1, 3, 6. We discuss variable selection for VAR-L and ADL-L models for h-step ahead forecasts based on the differenced data set $\{(\Delta_h O_{t+h}, S_t), t = 1, \ldots, T-h\}$, $\Delta_h O_{t+h} = O_{t+h} - O_t$. The google trend terms are selected by LASSO regression which minimizes the sum of squares of h-step forecast errors and a LASSO penalty,

$$\frac{1}{T-h}\sum_{t=1}^{T-h}\left(\Delta_h O_{t+h} - \beta_0 - \sum_{i=1}^{K}\beta_i S_{it}\right)^2 + \lambda\sum_{i=1}^{K}|\beta_i|, \ K = 17. \tag{3.1}$$

For λ, we consider the value of λ minimizing the Bayesian information criterion (BIC) of LASSO (3.1). The other criterions such as AIC, AICc, cross-validation method based on rolling window sample can be considered but these criterions select many google trend terms, which leads to poor forecasting performances in an unreported analysis of our data. Only one term, 'Fossil fuel' is selected for all $h = 1, 3, 6$. We denote the selected predictor S_{6t} by $\tilde{S}_t$.

3.2. Large vector autoregressive (VAR-L) model

The VAR-L model is a general penalized multivariate regression framework for large vector autoregressive (VAR) model proposed by Nicholson *et al.* (2017). Note that the number 17 of google trend terms is too large to allow us a practical forecast VAR model. In order to resolve this problem, we consider two methods. The first method is a direct application of 18 dimensional VAR-L model with all 18 variables. The second method is a two-step method in which the predictor $\tilde{S}_t = S_{6t}$ is first

selected by the method in Section 3.1 and next a dimension reduced VAR-L model with the selected predictor $\tilde{S}_t$ is employed for forecasting.
Then a VAR model is

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix} - \Phi_0 - \sum_{l=1}^{p} \Phi_l \begin{pmatrix} Y_{t-l} \\ X_{t-l}, \end{pmatrix} = a_t, \tag{3.2}$$

where $(Y_t, X_t) = (\Delta O_t, S_t), (\Delta O_t, \tilde{S}_t)$ for models with all google trends and the reduced google trend, respectively. Then, the VAR-L model is the VAR model (3.2) whose coefficients are obtained by minimizing

$$\sum_{t=1}^{T} \|a_t\|_2 + \lambda \Psi(\Phi), \tag{3.3}$$

where $\|\cdot\|_2$ is L_2-norm, $\Psi(\Phi)$ is a penalty function and λ is a penalty parameter. The value of λ is, for example as in Section 4, chosen by minimizing rolling cross-validation error squared sum computed from 1-step forecast for last 33% of train data, (Nicholson *et al.*, 2017). Some discussions on the penalty functions are given in Section 3.5.

3.3. ADL-L model

Note that VAR-L is a multi-variable forecast model in which multi-step forecasts can be constructed recursively. On the other hand, the ADL-L model is a single equation forecast model which cannot be iterated for multi-step forecasts. For multi-step, say h-step, forecast, we therefore specify an ADL-L model for O_{t+h}. The ADL-L model is considered only with the reduced predictor $\tilde{S}_t = S_{6t}$. Let a value $h \in \{1, 3, 6\}$ be given for forecast step. The ADL model for h-step ahead forecast is

$$\Delta_h O_{t-h+1} - \phi_0 - \sum_{j=1}^{p} \phi_j \Delta O_{t-j+1} - \sum_{j=1}^{q} \delta_j \tilde{S}_{t-j+1} = a_t, \tag{3.4}$$

where $\Delta_h O_{t+h} = O_{t+h} - O_t$. The ADL-L model of Li and Chen (2014) is the ADL model, whose coefficients are obtained by minimizing

$$\frac{1}{T-h} \sum_{t=1}^{T-h} a_t^2 + \lambda \left(\sum_{j=1}^{p} \left|\phi_j\right| + \sum_{j=1}^{q} \left|\delta_j\right| \right), \tag{3.5}$$

where λ is set by the same method as in VAR-L model (3.3). Note that an AR model for ΔO_t is augmented by lags of $\tilde{S}_t = S_{it}$ to become ADL and ADL-L models in (3.4)-(3.5).

3.4. VEC-L model

In order to address explicitly the cointegration relation between the google trend N_{5t} of 'Oil production' and the oil consumption level, we extend the VAR-L model of Nicholson *et al.* (2017) to a VEC-L model. Forecasts are made by a VEC model

$$\begin{pmatrix} \Delta O_t \\ \Delta N_{5t} \end{pmatrix} - \Phi_0 - \sum_{j=1}^{p} \Phi_j \begin{pmatrix} \Delta O_{t-j} \\ \Delta N_{5,t-j} \end{pmatrix} - \Xi z_{t-1} = a_t, \tag{3.6}$$

Table 3: VAR-L, VEC-L penalty functions

	$\Psi(\Phi)$	$\Psi^*(\Xi)$
Basic	$\|\Phi\|_1$	$\|\Xi\|_1$
Lag	$\sqrt{k^2}\sum_{j=1}^{p}\|\Phi_j\|_F$	$\sqrt{k}\|\Xi\|_F$
Own/other	$\sqrt{k}\sum_{j=1}^{p}\|\Phi_j^{on}\|_F + \sqrt{k(k-1)}\sum_{j=1}^{p}\|\Phi_j^{off}\|_F$	$\sqrt{k}\|\Xi\|_F$
Sparse lag	$(1-\alpha)\sqrt{k^2}\sum_{j=1}^{p}\|\Phi_j\|_F + \alpha\|\Phi\|_1$	$(1-\alpha)\sqrt{k}\|\Xi\|_F + \alpha\|\Xi\|_1$
Sparse own/other	$(1-\alpha)(\sqrt{k}\sum_{j=1}^{p}\|\Phi_j^{on}\|_F + \sqrt{k(k-1)}\sum_{j=1}^{p}\|\Phi_j^{off}\|_F) + \alpha\|\Phi\|_1$	$(1-\alpha)\sqrt{k}\|\Xi\|_F + \alpha\|\Xi\|_1$

Note: $\Phi = (\Phi_1, \ldots, \Phi_p)$, $\|\cdot\|_1$ is the L_1-norm and $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$ is the Frobenius norm of a matrix A and Φ_l^{on} and Φ_l^{off} are the diagonal and the off-diagonal elements of the coefficient matrix Φ_l, respectively.

where z_t is cointegration residual at time t in the linear regression of $\{O_t, t = 1, \ldots, T\}$ on $\{N_{5t}, t = 1, \ldots, T\}$. The VEC-L model is a VEC model (3.6) whose coefficients are estimated by minimizing

$$\sum_{t=1}^{T} \|a_t\|_2 + \lambda(\Psi(\Phi) + \Psi^*(\Xi)), \tag{3.7}$$

where $\|\cdot\|_2$ is the L_2-norm, Ξ is a 2×1 matrix, $\Psi(\Phi)$ and $\Psi^*(\Xi)$ are penalty functions on $\Phi = (\Phi_1, \ldots, \Phi_P)$ and Ξ, and λ is set by the same method as in VAR-L model (3.3). Note that the VEC, VEC-L models do not contain S_{6t} because S_{6t} is a stationary series.

3.5. Penalty functions

Some discussions on the penalty functions follow. Table 3 provides penalty functions $\Psi(\Phi)$ and $\Psi^*(\Xi)$. Each regularization penalty function gives different lag selections. For example, the lag VAR-L model tends to make all elements of estimated coefficient matrix Φ_l, $l = 1, \ldots, p$ be zero for some lag l. The basic VAR-L model tends to make some elements of estimated Φ_l be zero for all l, implying elimination of a variable in a equation. The own/other VAR-L model tends to make each estimated Φ_l be partitioned into separate groups. The own/other VAR-L model is designed to address the feature appearing in the macroeconomic forecast that the diagonal elements of Φ_l, indicating lagged own serial dependence, are more likely to be non-zero than the off-diagonal elements, indicating lagged cross-sectional dependence (Nicholson *et al.*, 2017). Nicholson *et al.* (2017) showed that the own/other VAR-L model has better forecast performance than the VAR-L models with other regularization penalties for US macroeconomic data of Stock and Watson (2005) and Canadian macroeconomic data. We will discuss which VAR-L models have the best forecast performance for oil consumption level in Section 4.

4. Out-of-sample forecast

We compare out-of-sample performances of the forecast models discussed in Section 3. In this section, O_t is set to 1/10000 times the original oil consumption level at time t for scale adjustment. Out-of-sample forecasts are obtained for the expanding window samples starting from $t_0 = 0.85T$, where $T = 189$ is the time length of the data. For each $t = t_0, \ldots, T - h$, the h-step ahead forecast value $\hat{O}_{t+h|t}$, $h = 1, 3, 6$ is obtained from the models fitted to the expanding window sample up to time t. We have chosen the 15% out-of-sample size because it is a very common choice in the forecasting literature (Cho and Shin, 2016; Choi and Shin, 2018; Kim and Shin, 2019; many others). The lag

Table 4: h-step out-of-sample performances of the forecast models

	1-step				3-step				6-step			
	AR	VAR-L	ADL-L	VEC-L	AR	VAR-L	ADL-L	VEC-L	AR	VAR-L	ADL-L	VEC-L
RMSE	0.0422	**0.0339**	0.0352	0.0384	0.0462	**0.0369**	0.0385	0.0448	0.0513	**0.0430**	0.0433	0.0550
MAE	0.0337	**0.0269**	0.0283	0.0321	0.0397	**0.0302**	0.0307	0.0364	0.0414	**0.0340**	0.0344	0.0438
MAPE	1.653	**1.324**	1.392	1.572	1.940	**1.478**	1.505	1.772	2.010	**1.654**	1.674	2.120

selection is made for each time $t \in \{t_0, \ldots, T-h\}$. For forecast performance, we consider

$$\text{MAE} = \frac{1}{m}\sum_{t=0.85T}^{T-h}\left|O_{t+h} - \hat{O}_{t+h|t}\right|,$$

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{t=0.85T}^{T-h}\left(O_{t+h} - \hat{O}_{t+h|t}\right)^2},$$

$$\text{MAPE} = \frac{1}{m}\sum_{t=0.85T}^{T-h}\left|\frac{O_{t+h} - \hat{O}_{t+h|t}}{O_t}\right|,$$

where $m = 0.15T - h + 1$.

4.1. Forecast comparison: AR, VAR-L, ADL-L, VEC-L

As a benchmarking model, we consider the autoregressive (AR) model,

$$\Delta O_t = \phi_0 + \sum_{l=1}^{p}\phi_l \Delta O_{t-l} + a_t, \tag{4.1}$$

whose the order p is specified by minimizing the BIC. We compare AR, VAR-L, ADL-L VEC-L forecasts. For the VAR-L and ADL-L models, the predictor is the reduced google trend $\tilde{S}_t = S_{6t}$, which is selected from the first 85% observations as discussed in Section 3.1. A forecast comparison for VAR-L model with all google trends and VAR-L model with reduced google trend is given in Section 4.2. It will be demonstrated that the VAR-L is the best among the forecast models AR, VAR-L, VEC-L and ADL-L. The AR, VAR-L, VEC-L models compute the h-step ahead forecast recursively from models of (3.7), (3.3), (4.1). The ADL-L models compute the h-step ahead forecasts by fitting (3.4) with LASSO regression (3.5) for each h. For ADL-L, VAR-L, VEC-L models, we set $p = q = 12$ in (3.3), (3.5), (3.7) and use the own-other penalty in Table 3.

Table 4 reports h-step out-of-sample forecast performances of the models. The best forecast performance is achieved by the VAR-L. It shows that explicit consideration of the cointegration relation does not help improve forecast performance. From the forecast performance comparison between the AR and ADL-L models, we see that the google trends considerably improve forecast performance for oil consumption. We also find that the VAR-L model shows the best forecasts for all h. Note that the comparison is within the given data set of Section 2 and is not for randomly generated data sets. There are no other random components in the comparison. Therefore, there is no issue of sampling error.

To demonstrate a superiority of the VAR-L model more formally, the model confidence set (MCS) analysis of Hansen *et al.* (2011) is made in Table 5. The MCS analysis is a statistically regorous tool showing relative forecast performance of several forecasting methods. The MCS is a set of one or more good forecast models in terms of the RMSE or the MAE. The MCS analysis gives each model forecast performance rank and p-value, the probability that the model is included in the MCS. The

Table 5: *P*-values (rank) of MCS analysis for the forecast models

	1-step				3-step				6-step			
	AR	VAR-L	ADL-L	VEC-L	AR	VAR-L	ADL-L	VEC-L	AR	VAR-L	ADL-L	VEC-L
RMSE	0.00	**1.00(1)**	0.00	0.00	0.00	**1.00(1)**	0.09(3)	0.19(2)	0.00	**1.00(1)**	0.75(2)	0.00
MAE	0.12(3)	**1.00(1)**	0.12(2)	0.03(4)	0.00	**1.00(1)**	0.59(2)	0.00	0.05(4)	**1.00(1)**	1.00(2)	1.00(3)

Table 6: *h*-step out-of-sample performances of VAR-L forecasts with reduced google trends and all google trends

	1-step		3-step		6-step	
	Reduced	All	Reduced	All	Reduced	All
RMSE	0.0339	**0.0335**	**0.0369**	0.0382	**0.0430**	0.0434
MAE	**0.0269**	0.0275	**0.0302**	0.0326	**0.0340**	0.0353
MAPE	**1.324**	1.349	**1.478**	1.593	**1.654**	1.717

performance of the VAR-L model is the best, being always ranked 1 and having p-value of 1.00 for all 1, 3, 6, steps. The rank 1 of the VAR-L model is not shared by any other models. On the other hand, the ranks of the AR and VEC-L models are mostly poor. This MCS analysis shows that the VAR-L model is statistically better than the other models.

Therefore, from Tables 4-5, we can say that proper selection of predictors and lags by the VAR-L method and proper differencing are important factors in improving forecast performance but explicit cointegration consideration between oil consumption level and google trend is not an important factor. From now on, we therefore concentrate on the VAR-L model.

4.2. Comparison of VAR-L forecasts with all google trends and reduced google trends

This subsection compares two VAR-L models. The first one is the 18-dimensional VAR-L model with all 17 google trends S_t as predictors. The second one is a two dimensional VAR-L model whose predictor $\tilde{S}_t = S_{6t}$ is selected by the second method in Section 3.1. Table 6 provides the out-of-sample forecast performances.

The 18-dimensional VAR-L model tends to yield worse forecast performance than the 2-dimensional VAR-L model especially for multi-step forecasts. It may be caused by the estimation problem of over-parametrized model for high dimensional predictor. The 18-dimensional VAR-L model has too many parameters yielding poor parameter estimates for the data set of the total length of 187 times even under L1 regularization. Therefore, although the VAR-L model allows us to give efficient estimation and accurate forecasting for high-dimensional data sets as discussed by Nicholsen *et al.* (2017), it may be better that the VAR-L model is composed parsimoniously including only important predictors.

4.3. Comparisons of VAR vs VAR-L forecasts

We next compare unregularized classical VAR model (3.2) with the regularized VAR-L model with the dimension reduced predictor $\tilde{S}_t = S_{6t}$ to emphasize the necessity of the lag selection. For each $t = t_0, \ldots, T - h$, in computing $\hat{O}_{t+h|t}$ by the VAR model, the model order p is chosen by the BIC applied to the data up to time t. Table 7 reports h-step out of forecast performances. The table reveals that the regularized VAR-L model performs much better than the corresponding unregularized VAR model. The finding means that the lag selection by the LASSO-type regularization should be considered to improve forecast performance even for the reduced 2-dimensional VAR model.

Table 7: h-step out-of-sample forecast performances for VAR and VAR-L with dimension reduced predictor $\tilde{S}_t = S_{6t}$ for differenced data set

	1-step		3-step		6-step	
	VAR	VAR-L	VAR	VAR-L	VAR	VAR-L
RMSE	0.0426	**0.0339**	0.0463	**0.0369**	0.0523	**0.0430**
MAE	0.0342	**0.0269**	0.0398	**0.0302**	0.0420	**0.0340**
MAPE	1.677	**1.324**	1.941	**1.478**	2.043	**1.654**

Table 8: h-step out-of-sample forecast performances of VAR-L models made with different penalties

h-step		Basic	Lag	Own/Other	Sparse Lag	Sparse Own/Other
1-step	RMSE	0.0343	**0.0339**	**0.0339**	0.0341	0.0351
	MAE	0.0271	**0.0266**	0.0269	0.0269	0.0269
	MAPE	1.332	**1.308**	1.324	1.319	1.319
3-step	RMSE	0.0371	**0.0363**	0.0364	0.0369	0.0376
	MAE	0.0304	**0.0298**	0.0302	0.0300	0.0309
	MAPE	1.487	**1.459**	1.478	1.472	1.512
6-step	RMSE	0.0437	**0.0424**	0.0430	0.0426	0.0453
	MAE	0.0348	0.0342	**0.0340**	0.0341	0.0357
	MAPE	1.692	1.664	**1.654**	1.658	1.739

4.4. Comparison of VAR-L forecasts with different penalty functions

Since own/other group VAR-L model with the dimension reduced predictor $\tilde{S}_t = S_{6t}$ has superior forecasting performance to the other forecast models in Tables 4, 5, we consider the VAR-L model in more detail with other penalty functions. Table 8 provides h-step ahead out-of-sample forecast performances of the VAR-L models having different penalties $\Psi(\Phi)$, $\Psi^*(\Xi)$ in Table 3. For h =1, 3 steps, the lag group VAR-L model performs better than the other VAR-L models. For 6 step, both the lag group and the own/other group VAR-L models have better performance than the other VAR-L models. However, the difference in the forecast performances among different penalties for the VAR-L model is much smaller than the differences of forecast performances in Table 4 for different models, indicating that all VAR-L models have better forecast performances for oil consumption level than the other models in Table 4.

5. Conclusion

Forecasting US oil consumption is improved by considering google trends which are obtained from online google big data. We investigate the roles of selecting the google trend terms and the lags of predictors. Accordingly, the forecast is conducted with some google trend terms selected from the LASSO regression and with some lags of the predictors selected from the VAR-L model of Nicholson *et al.* (2017). An out-of-sample forecast shows that the VAR-L model has the superior forecast performance to the AR, VAR, ADL-L, VEC-L models and the superiority is acquired by selecting important google trend terms and selecting proper lag. We also find that the VAR-L model with the lag group penalty function tends to give better forecast performances than the VAR models with other penalty functions.

References

Baumeister C and Kilian L (2015). Forecasting the real price of oil in a changing world: a forecast combination approach, *Journal of Business and Economic Statistics*, **33**, 338–351.

Bulut L (2018). Google Trends and the forecasting performance of exchange rate models, *Journal of Forecasting*, **37**, 303–315.

Carriere-Swallow Y and Labbe F (2013). Nowcasting with google trends in an emerging market, *Journal of Forecasting*, **32**, 289–298.

Cepni O, Guney IE, and Swanson NR (2019). Nowcasting and forecasting GDP in emerging markets using global financial and macroeconomic diffusion indexes, *International Journal of Forecasting*, **35**, 555–572.

Cho SJ and Shin DW (2016). An integrated heteroscedastic autoregressive model for forecasting realized volatilities, *Journal of the Korean Statistical Society*, **45**, 371–380.

Choi JE and Shin DW (2018). Forecasts for leverage heterogeneous autoregressive models with jumps and other covariates, *Journal of Forecasting*, **37**, 691—704.

Fantazzini D and Fomichev N (2014). Forecasting the real price of oil using online search data, *International Journal of Computational Economics and Econometrics*, **4**, 4–31.

Guo JF and Ji Q (2013). How does market concern derived from the Internet affect oil prices?, *Applied Energy*, **112**, 1536–1543.

Hansen PR, Lunde A, and Nason JM (2011). The model confidence set, *Econometrica*, **79**, 453–497.

Kim HS and Shin DW (2019). Forecast of realized covariance matrix based on asymptotic distribution of the LU decomposition with an application for balancing minimum variance portfolio, *Applied Economics Letters*, **26**, 661–668.

Li J and Chen W (2014). Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models, *International Journal of Forecasting*, **30**, 996–1015.

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.

Messner JW and Pinson P (2019). Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting, *International Journal of Forecasting*, **35**, 1485–1498.

Nicholson WB, Matteson DS, and Bien J (2017). VARX-L: Structured regularization for large vector autoregressions with exogenous variables, *International Journal of Forecasting*, **33**, 627–651.

Niesert RF, Oorschot JA, Veldhuisen CP, Brons K, and Lange RJ (2020). Can Google search data help predict macroeconomic series?, *International Journal of Forecasting*, **36**, 1163–1172.

Sagaert YR, Aghezzaf EH, Kourentzes N, and Desmet B (2018). Tactical sales forecasting using a very large set of macroeconomic indicators, *European Journal of Operational Research*, **264**, 558–569.

Smeekes S and Wijler E (2018). Macroeconomic forecasting using penalized regression methods, *International Journal of Forecasting*, **34**, 408–430.

Stock JH and Watson MW (2005). *An Empirical Comparison of Methods for Forecasting Using Many Predictors*, Manuscript, Princeton University.

Tarassow A (2019). Forecasting U.S. money growth using economic uncertainty measures and regularisation techniques, *International Journal of Forecasting*, **35**, 443–457.

Wen F, Gong X, and Cai S (2016). Forecasting the volatility of crude oil futures using HAR-type models with structural breaks, *Energy Economics*, **59**, 400–413.

Yu L, Zhao Y, Tang L, and Yang Z (2019). Online big data-driven oil consumption forecasting with Google trends, *International Journal of Forecasting*, **35**, 213–223.

Zhang JL, Zhang YJ, and Zhang L (2015). A novel hybrid method for crude oil price forecasting, *Energy Economics*, **49**, 649–659.

Zhao Y, Li J, and Yu L (2017). A deep learning ensemble approach for crude oil price forecasting, *Energy Economics*, **66**, 9–16.

Received June 11, 2021; Revised October 15, 2021; Accepted October 28, 2021