Nowcasting Tourist Arrivals to Prague: Google Econometrics

Zeynalov, Ayaz

IES, FSS, Charles University in Prague Abstract

2014

Nowcasting Tourist Arrivals to Prague: Google Econometrics

Ayaz Zeynalov*

IES, FSS, Charles University in Prague

Abstract

It is expected that what people are searching for today is predictive of what they have done recently or will do in the near future. This research will analyze the eligibility of Google search data to nowcast tourist arrivals to Prague. The present research will report whether Google data is potentially useful in nowcasting or short-term forecasting using by Support Vector Regressions (SVRs), which maps data to a higher dimensional space and employs a kernel function.

Keywords: Google trends, nowcasting, tourism forecasting

JEL Classification: C53; E17; L83

Introduction

People reveal useful information about their needs, wants, interests, and concerns from the internet. It might be the best explanation of Google's success, performed rapid growth

*Corresponding author: IES, FSS, Charles University in Prague, Opletalova 26, Prague 1, 110 00, Czech Republic. e-mail: azerayaz@gmail.com

with using internet information, made it publicly accessible and useful for consumer. It is expected that what people are searching for today is predictive of what they have done recently or will do in the near future.

The studies have focused on search for “predicts the present, show that search queries correlated to the contemporaneous activities [Askitas and Zimmermann, 2009, Hong, 2011, Choi and Varian, 2012, etc.]. In fact, Choi and Varian [2012] showed how to use search engine data for nowcasting the value of economic indicators, such as unemployment claims, automobile sales, consumer confidence and travel trends.

Several studies have discovered that the use of Google trend data is useful as an economic indicator. The research tested whether Google Trends Automotive Index improves the fit and efficiency of nowcasting models for automobile sales in Chile [Carriere-Swallow and Labbe, 2013]; demonstrates strong correlations between internet searches queries and unemployment rates in Germany [Askitas and Zimmermann, 2009]; forecasts the real price of oil on the basis of macroeconomic indicators and Google search data [Fantazzini and Fomichev, 2014]; uses Google Flu Trends data to describe the spread of flu in the United States during 2003-2009 [Dukic et al., 2012]; test whether Google queries can enhance predictions of youth unemployment in France [Fondeur and Karam, 2013]; offers significant benefits to forecasters of private consumption indicators based on search query time series provided by Google Trends [Vosen and Schmidt, 2011]; uses search query volume to forecast outcomes such as unemployment levels, auto and home sales, and disease prevalence in near real time [Goel et al., 2010]; analyzes factors that influence investor information demand around earnings announcements via Google searches [Drake et al., 2012]; emphasizes an approach to portfolio diversification based on popularity of a stock measured by search queries using Google Trends [Kristoufek, 2013].

Tourism forecasting had strong interested by many studies. The studies adjusted indicators of the inflow of tourists to obtained with a lead of almost one month of tourist arrival using Google Trends [Artola and Galn, 2012]; employed modelling and forecasting tourist arrivals to Hong Kong from China, South Korea, the UK and the USA [Song et al., 2011]; evaluated the different estimation methods of forecasting tourism data, which include 366 monthly series, 427 quarterly series and 518 annual series [Athanasopoulos et al., 2011]; analyzed external demand for Spanish tourist services within the framework of Structural Time Series Models which included different types of indices [Gonzalez and Moral, 1995]. This research will test whether Google Trends can help in monitoring tourist arrivals to Prague.

Google Trends provides free, large and practically real-time information, however it has some disadvantages. Firstly, Google shows only absolute data, where it provides an index which is relative to all searches. Secondly, internet users might type similar words even if they were looking for different topics, or different words even if they were searching for same topic. Thirdly, web search queries are related to personal characteristics such as education, income, age, etc. Google search is not perfect, however, based on the fact that it is one of best information data store web data, predicting present or near future might help improve efficiency in economic indictor via Google search query: it has the potential to act as a leading indicator.

This research will analyze the eligibility of Google search data to nowcast tourist arrivals to Prague. The present research will report whether Google data is potentially useful in nowcasting or short-term forecasting using with Support Vector Regressions (SVRs), which maps data to a higher dimensional space and employs a kernel function.

Methodology and Data

Methodology

The study will employ how to get better nowcast on tourist arrivals by using SVRs. The empirical estimation will compare SVR with Google and autoregressive SVR, to detect whether Google search queries can add some insight into tourism prediction for Prague. Forecasting literature starts to chose a baseline model for getting meaningful predictive power. Afterward, the baseline model will test with and without Google data to analyze whether Google can improve forecasting on tourist arrivals.

Methodology will start to evaulate and compare three different univariate structure models: the Autoregressive Integrated Moving Average (ARIMA), the univariate Structural Time Series Model (STSM), and the Autoregressive (AR) SVR model. For choosing the best forecasting accuracy, the research will deal with traditional Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE).

Following Hong [2011]: "In the high dimensional feature space, there theoretically exists a linear function, f, to formulate the nonlinear relationship between input data (training data set) and output data", the linear function can be described as:

$$f(x) = w^T \phi(x) + b \tag{1}$$

where $f(x)$ denotes the forecasting values, $\phi(x)$ represents a mapping function in the feature space, w is a weight vector and b denotes a constant. The SVR method aims at minimizing the regularized risk function:

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} \Theta_\epsilon(y_i, w^T \phi(x_i) + b) \tag{2}$$

where $\Theta_\epsilon(y_i, w^T\phi(x_i) + b)$ is the ϵ-insensitive loss function:

$$\Theta_\epsilon(y_i, w^T\phi(x_i) + b)) = \begin{cases} |w^T\phi(x_i) + b - y_i| - \epsilon & \text{if } w^T\phi(x_i) + b - y_i| >= \epsilon \\ 0 & \text{otherwise} \end{cases}$$

The SVR focuses on finding the optimum hyper plane and minimizing the training error between the training data and the ϵ-insensitive loss function. Then, the SVR minimizes the overall errors:

$$\underset{w,b,\xi^*,\xi}{\text{minimize}} \qquad R_\epsilon(w, \xi^*, \xi) = \frac{1}{2}w^Tw + C\sum_{i=1}^{N}(\xi_i^* + \xi_i) \tag{3}$$

$$\text{subject to}$$
$$y_i - w^T\phi(x_i) - b \geqslant \epsilon + \xi_i^*, i = 1, 2, ..., N$$
$$-y_i + w^T\phi(x_i) + b \geqslant \epsilon + \xi_i, i = 1, 2, ..., N$$
$$\xi_i^*, \xi_i \geqslant 0, i = 1, 2, ..., N$$

where $\frac{1}{2}w^Tw$ is a regularization term that employees the trade-off between the complexity and accuracy of the regression to maintain regression flatness. $\sum\limits_{i=1}^{N}(\xi_i^* + \xi_i)$ penalizes training errors of $f(x)$ and y by using the ϵ-insensitive loss function. C is the regularization constant to emphasize trade-off between the empirical risk and regularization term.

After the quadratic optimization problem with inequality constraints are solved, the parameter vector w: in Eq.(1):

$$w = \sum_{i=1}^{N}(\beta_i^* - \beta_i)\phi(x_i) \tag{4}$$

where β_i^* and β_i, are obtained by solving a quadratic program and are the Lagrangian multipliers.

The general form of SVR function is:

$$f(x) = \sum_{i=1}^{N}(\beta_i^* - \beta_i)K(x_i, x_j) + b \tag{5}$$

where $K(x_i, x_j)$ expresses kernel function. The value of the kernel equals the inner product of two vectors, x_i and x_j, in the feature space $\phi(x_i)$ and $\phi(x_j)$, respectively. There are several types of kernel functions. The most used kernel function is the Gaussian radial basis function (RBF) with a width of σ: $K(x_i, x_j = exp(\frac{-|x_i - x_j|^2}{2\sigma^2})$. To get a more accurate forecasting model, it is highly related to the choice of hyper parameters C, ϵ and the kernel parameters (σ).

This study will use SVRs to estimate:

$$tourist_t = \sum_{i=1}^{\kappa} \alpha_i tourust_{t-i} \tag{6}$$

$$tourist_t = \sum_{i=1}^{\kappa} \alpha_i tourust_{t-i} + \sum_{i=1}^{\rho} \beta_i google_{t-i} \tag{7}$$

where t represents time, and $t - i$ represents lags. The number of lags (κ) and ρ will be chosen by Akaike Information Criteria (AIC) and Root Mean Square Error (RMSE). The number of lags (κ) in Eq.(5) will be kept the same as in Eq.(6), due to maintain pure AR-SVR for the Google-based SVR.

Before the nowcasting analysis, the degree of correlation between the number of tourist arrivals and Google trends will be determined. This visual evaluation will help to understand whether the time series has co-movement, for instance, whether tourist arrivals have seasonality or not. This preliminary analysis will emphasize whether Google can provide insight on tourist arrivals to Prague. Furthermore, visual evaluation will be developed with the Granger causality test, which will reveal whether contemporaneous search queries can help better predictions for tourist arrivals. Visual evaluation and the

Granger test might be misleading, therefore, SVR will help to develop and evaluate whether findings are consistent with robust methods for “nowcasting”.

Data

Weekly data will be obtain from the Prague Immigration Department from January, 2008 to September, 2014. The study will collect search volume histories related to the simple search term “prague” under the Google Trends. Mostly, macroeconomic variables are available at least on a monthly basis, which causes multi-frequency problem [Fondeur and Karam, 2013]. Using weekly dataset in this research will help to rid of multi-frequency problems.

References

Concha Artola and Enrique Galn. Tracking the future on the web: construction of leading indicators using internet searches. Banco de Espania Occasional Papers 1203, Banco de Espania, April 2012.

Nikolaos Askitas and Klaus F. Zimmermann. Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly (formerly: Konjunkturpolitik), Duncker & Humblot, Berlin*, 55(2):107–120, 2009.

George Athanasopoulos, Rob J. Hyndman, Haiyan Song, and Doris C. Wu. The tourism forecasting competition. *International Journal of Forecasting*, 27(3):822–844, July 2011.

Yan Carriere-Swallow and Felipe Labbe. Nowcasting with Google Trends in an Emerging Market. *Journal of Forecasting*, 32(4):289–298, 07 2013.

Hyunyoung Choi and Hal Varian. Predicting the Present with Google Trends. *The Economic Record*, 88(s1):2–9, 06 2012.

Michael S. Drake, Darren T. Roulstone, and Jacob R. Thornock. Investor Information Demand: Evidence from Google Searches Around Earnings Announcements. *Journal of Accounting Research*, 50(4):1001–1040, 09 2012.

Vanja Dukic, Hedibert F. Lopes, and Nicholas G. Polson. Tracking Epidemics With Google Flu Trends Data and a State-Space SEIR Model. *Journal of the American Statistical Association*, 107(500):1410–1426, December 2012.

Dean Fantazzini and Nikita Fomichev. Forecasting the real price of oil using online search data. *International Journal of Computational Economics and Econometrics*, 4(1/2): 4–31, 2014.

Y. Fondeur and F. Karam. Can Google data help predict French youth unemployment? *Economic Modelling*, 30(C):117–125, 2013.

S. Goel, J. Hofman, S. Lehaie, D. M. Pennock, and D. J. Watts. Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America*, 7:1748617490, 2010.

Pilar Gonzalez and Paz Moral. An analysis of the international tourism demand in Spain. *International Journal of Forecasting*, 11(2):233–251, June 1995.

Wei-Chiang Hong. Electric load forecasting by seasonal recurrent SVR (support vector regression) with chaotic artificial bee colony algorithm. *Energy*, 36(9):5568–5578, 2011.

Ladislav Kristoufek. Can Google Trends search queries contribute to risk diversification? Papers 1310.1444, arXiv.org, October 2013.

Haiyan Song, Gang Li, Stephen F. Witt, and George Athanasopoulos. Forecasting tourist arrivals using time-varying parameter structural time series models. *International Journal of Forecasting*, 27(3):855–869, 2011.

Simeon Vosen and Torsten Schmidt. Forecasting private consumption: survey based indicators vs. Google trends. *Journal of Forecasting*, 30(6):565–578, September 2011.