Using Google Trend Data as an Initial Signal Indonesia Unemployment Rate

3 authors:

Rani Nooraeni
Sekolah Tinggi Ilmu Statistik

Nurmitra Sari Purba
Sekolah Tinggi Ilmu Statistik

Nugroho Puspito Yudho
Statistics Indonesia

62nd ISI World Statistics Congress
ISI 2019
Kuala Lumpur
18 - 23 August 2019

PROCEEDING
CONTRIBUTED PAPER SESSION
VOLUME 3

62nd ISI WORLD STATISTICS CONGRESS 2019

18 - 23 August 2019, Kuala Lumpur

Come | Connect | Create

www.isi2019.org

PROCEEDING

ISI WORLD STATISTICS CONGRESS 2019

CONTRIBUTED PAPER SESSION (VOLUME 3)

Published by:

Department of Statistics Malaysia
Block C6, Complex C
Federal Government Administrative Centre
62514 Putrajaya
MALAYSIA

Central Bank of Malaysia
Jalan Dato' Onn
P.O. Box 10922
50929 Kuala Lumpur
MALAYSIA

Malaysia Institute of Statistics
Department of Mathematical Sciences
Faculty of Sciences and Technology
43600 UKM Bangi, Selangor
MALAYSIA

Portal : https://www.isi2019.org
Email : lpc@isi2019.org

Published in February 2020

Suggested citation:
Department of Statistics Malaysia (DOSM). 2019. Proceeding of the 62nd ISI World Statistics Congress 2019: Contributed Paper Session: Volume 3, 2019. 444 pages

Preface

The 62nd International Statistical Institute World Statistics Congress (ISI WSC 2019) has a long tradition since 1887, held for the first time in Kuala Lumpur, Malaysia on 18 to 23 August 2019. ISI WSC 2019 is a global gathering of statistical practitioners, professionals and experts from industries, academia and official authorities to share insights in the development of statistical sciences.

The congress attracted an overwhelming number of participants across the regions. The scientific sessions were delivered over five days with parallel sessions and e-poster sessions running all day long. The scientific program reaches across the breadth of our discipline that comprised of Invited Paper Sessions (IPS), Special Topic Sessions (STS) and Contributed Paper Sessions (CPS). Papers presented exhibit the vitality of statistics and data science in all its manifestations.

I am very honoured to present the proceedings of ISI WSC 2019 to the authors and delegates of the congress. The proceedings contain papers presented in IPS, STS and CPS which were published in fourteen (14) volumes. Scientific papers were received from August 2018 and were carefully reviewed over few months by an external reviewer headed by Scientific Programme Committee (SPC) and Local Programme Committee (LPC). I am pleased that the papers received cover variety of topics and disciplines from across the world, representing both developed and developing nations.

My utmost gratitude and appreciation with the expertise and dedication of all the reviewers, SPC and LPC members for their contributions that helped to make the scientific programme as outstanding as it has been.

Finally, I wish to acknowledge and extend my sincere thanks to the member of National Organising Committee of ISI WSC 2019 from Department of Statistics Malaysia, Bank Negara Malaysia, Malaysia Institute of Statistics and International Statistical Institute for endless support, commitment and passion in making the ISI WSC 2019 a great success and congress to be remembered.

I hope the proceedings will furnish the statistical science community around the world with an outstanding reference guide.

Thank you.

Dr. Mohd Uzir Mahidin
Chairman
National Organising Committee
62nd ISI WSC 2019

Scientific Programme Committee of the 62nd ISI WSC 2019

Chair
Yves-Laurent Grize, Switzerland

Vice-Chairs
Kwok Tsui, Hong Kong, China
Jessica Utts, USA

Local Programme Committee Chair and Co-Chair
Rozita Talha, Malaysia
Ibrahim Mohamed, Malaysia

Representatives of Associations
Rolando Ocampo, Mexico (IAOS)
Cynthia Clark , USA (IASS)
Bruno de Sousa, Portugal (IASE)
Sugnet Lubbe, South Africa (IASC)
Mark Podolskij, Denmark (BS)
Beatriz Etchegaray Garcia, USA (ISBIS)
Giovanna Jona Lasinio, Italy (TIES)

At-Large Members
Alexandra Schmidt, Canada/Brazil
Tamanna Howlader, Bangladesh

Institutional/Ex Officio
Helen MacGillivray, ISI President, Australia
Fabrizio Ruggeri, Past SPC Chair, Italy

Liaison at ISI President's Office
Ada van Krimpen, ISI Director
Shabani Mehta, ISI Associate Director

Local Programme Committee of the 62nd ISI WSC 2019

Chairperson
Rozita Talha

Co-Chairperson
Prof. Dr. Ibrahim Mohamed

Vice Chairperson
Siti Haslinda Mohd Din
Daniel Chin Shen Li

Members
Mohd Ridauddin Masud
Dr. Azrie Tamjis
Dr. Chuah Kue Peng
Dr. Tng Boon Hwa
Prof. Dr. Ghapor Hussin
Prof. Madya Dr. Yong Zulina Zubairi
Prof. Madya Dr. Wan Zawiah Wan Zin @ Wan Ibrahim
Dr. Nurulkamal Masseran
Siti Salwani Ismail
Rosnah Muhamad Ali
Dr. Ng Kok Haur
Dr. Rossita Mohamad Yunus
Muhammad Rizal Mahmod

TABLE OF CONTENTS

Contributed Paper Session (CPS): Volume 3

applications to spatial, video, and network outlier detection. Data Mining and Knowledge Discovery, 28(1)

12. Shekhar S, Lu Chang-tien, Pu-sheng (2003). A Unified Approach to Detecting Spatial Outliers. Geo-Informatics,An International Journal on Advances of Computer Science for Geographic Information System, 7(2), pp.139-166.
13. Xin, L., Shaoliang, Z., & Pulin, Z. (2015). Spatial Outlier Detection of CO2 Monitoring Data Based on Spatial Local Outlier Factor. Journal of Engineering Science and Technology Review, 8(5), 110-116.

Using Google trend data as an initial signal Indonesia unemployment rate

Rani Nooraeni[1], Nurmitra Sari Purba[2], Nugroho Puspito Yudho[3]
[1,2]STIS, Polytechnic of Statistic, Jakarta, Indonesia
[3]BPS, Statistic Indonesia, Jakarta

Abstract

Macroeconomic data are generally published with a time lag. Nowcasting uses the latest available data to provide estimates of macroeconomic variables in the short term. Surely this prediction applies until the actual estimation is generated. The ability to search data through the internet has provided a new data source for researchers who are interested in predicting macroeconomic variables in the short term. This paper trying to learn the behavior intensity of keyword from google search in order to predict unemployment growth in Indonesia. Some modeling is done to get the results of nowcasting which is quite ideal using ARIMA, BSTS without and with Google Data as an additional variable. We also conduct several treatments related to variations of keywords obtained. selection of important keywords with stepwise regression and keyword grouping using the principal component analysis method to obtain a composite keyword that can produce a better model for predicting unemployment growth rates.

Keywords

unemployment; Google trend; nowcasting; BSTS; composite

1. Introduction

The government needs various macroeconomic indicators as the basis of determination national policy strategies. They use Macroeconomic indicators that resulted by national institutions. Some indicators that are often used as a basis for policy include GDP, inflation, unemployment, exchange rates, poverty, and others. The overall indicators are obtained from the process data collection such as census/surveys. The indicator of the unemployment was conducted either by BPS Statistics of Indonesia, the Ministry of Manpower, or other agencies that conduct separate surveys regarding the labor market. Even though many parties have collected data related to employment, each of the data collection processes has limitations, especially the period of dissemination. For example, the unemployment rate produced by BPS can only be disseminated twice in a year. This certainly influences policymakers because the basis of the policy is only able based on available data which does not necessarily match with the condition at the time the policy will be made. Because often policies need to be made frequently in line with changing times.

On the other hand, the development of the era has made information technology more sophisticated, from time to time the percentage of households that have computers and cell phones has increased. In sequence, they growth from 19.88 percent to 88.04 percent while households that have mobile phones have grown from 3.65 percent to 18,71 percent, during the period of 2005-2015. In line with the rapid increase in ownership and mastery of these technologies. Development of internet users is also growing rapidly. At the same period time, in Indonesia, the percentage of internet users increased from 7.4 percent to 34.9 percent. With this phenomenon, the impact of that depelovment is that people become easier to get information and also easy to give information. This has an impact on the use of search engines on the internet, more and more people are looking for information about anything through search engines that can cross between countries.

An example of a phenomenal search engine is Google Search. Google Search is commonly used to find relevant information using various keywords. The intensity of search uses various kinds of keywords that produce data that is so large (Big Data) recorded by Google. Further, the data can show the latest socio-economic conditions. Then Google began to disseminate the intensity of the search in 2009 through the Google Trend interface. Of course, this is good news for researchers. They use this data to improve the accuracy of the latest predictions. Mitra, Sanyal, and Choudhury (2017), they were nowcasting Real Estate growth in India using Google Trend Data. The availability of digital data whose time references are more up-to-date, weekly, monthly or annual, this can be supporting data for data produced by official institutions through the census/survey data collection process.

Google provides three data sources that can be useful for social science (i.e. google trend, google correlate, and google Consumer Surveys (Stephens-Davidowitz & Varian, 2015), but they claim that besides social science there are many things that can be found from Google. Koop (2013) used google data to predict macroeconomic indicators, with Google Data we can make predictions about things in the most recent period, this is called Nowcasting, so we can use Google data to support indicators related to economic or social indicators of a country. In 2009, Google began to disseminate data from Google Search, including data from 2004 to the present, covering all countries in the world.

Based on the foregoing, this study will explore Google Data in describing the conditions of economic indicators in the current period. The focus of the macroeconomic indicators that we observe is the growth rate of unemployment in Indonesia. However, in the use of Google Search Data, caution is needed in determining keywords related to the problem to be studied, the behavior of Google Search data must also be identified so that the right method can be used to produce a good prediction model. One of the projection methods that can be used for this kind of data, internet data, is the Bayesian Structural Time Series. This paper utilizes data from the internet, Google trend (GT) data, to know the recent conditions of macroeconomic indicators in Indonesia. Then combinate the GT data with another macroeconomic variable to improve short-term predictive/nowcasting capabilities model.

Google Trend Data

When we are going to use GT data, there are some things that need to be underlined. First, data of GT has the potential to predict short-term but not for the long term. Second, it is rarely used for broadening macroeconomic variables such as inflation, industrial production, etc. It is more useful for predicting certain variables related to consumption, housing or the labor market. For example, Choi and Varian (2011) succeeded in predicting motor vehicle parts and cars, initial claims for unemployment benefits and tourist arrivals in Hong Kong

In line with the development of research that utilizes GT data, we found some literature/references that conducted research using GT Data to predict macroeconomic indicators. The Google variable can be used as additional information and check whether GT involvement can improve nowcasting capabilities.

2. Methodology

2.1 Data used

The data used in this study consisted of two sources, BPS-Statistics Indonesia and Google. Macroeconomic variables, namely unemployment rate, Consumen Price Index (CPI) and inflation are obtained from BPS-Statistics Indonesia. Google trend data collected based on several keywords that are relevant to the unemployment rate of Indonesia. Unemployment rate official data of Indonesia result from survey national of labour (SAKERNAS). BPS conduct this survey twice a year, February and August. A time reference that we use in build model prediction was from February 2005 to August 2018.

2.2 Selection of Keywords

The selection of keywords has a crucial role in producing precision prediction results. It takes several stages to finally produce keywords that are closely related to the subject we tell. In the use of keywords in the early stages can refer to keywords that have been used by previous researchers who have the same concern. In addition, we can find the relevant keywords from google correlate. However, not all of the keywords produced will be relevant to the subject we are observing. Thus, we have to combine all references, all procedures, and statistic method to get the most representative keyword.

Finally, based on relevancy, the following keywords are suggested to our study (in Bahasa):

bisnis	contoh cv	karir	lowongan kerja
bisnis indonesia	contoh lamaran pekerjaan	kerja	lowongan kerja di
bisnis online	employment	kerja di	lowongan kerja teknik
bursa kerja	job	kerja part time	lowongan pekerjaan
bursa lowongan	job application	lamaran kerja	lowongan teknik
career	jobsdb	lamaran pekerjaan	pekerjaan
cari duit	jobsdb.com	loker	peluang bisnis
cari kerja	jobstreet	lowongan	surat lamaran
cari uang	job vacancy	lowongan di	

The prediction of condition Unemployment rate in Indonesia will execute with only several important keywords. The selection will apply stepwise regression method to determine what keywords are most relevant to the research subject and Principal Component Analyse to combine number of variables.

2.3 ARIMA Model

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. ARIMA is an acronym that stands for Auto Regressive Integrated Moving Average. It has very good accuracy for short-term forecasting. Standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used. Parameter p is the number of lag observations included in the model, also called the lag order. Parameter d is the number of times that the raw observations are differenced, also called the degree of differencing. Parameter q is the size of the moving average window, also called the order of moving average.

A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model. Differencing, autoregressive, and moving average components make up a non-seasonal ARIMA. ARIMA (p,0,q) model can be written as a linear equation:

$$Y_t = c + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + e_t + \cdots + \theta_q e_{t-k}$$

where φ_1, φ_2 are parameters for Auto Regressive and θ_q is parameter of Moving Average. When we include either explanatory variable in to ARIMA model it is named ARIMAX model.

2.4 Bayesian Structural Time Series (BSTS) Model

One of the advantages of Bayesian modeling is to account for the uncertainty associated with parameter estimates and provide exact measures of uncertainty on the posterior distributions of these parameters, which is traditionally ignored in classical estimation models. In addition, Bayesian estimation and inference provide confidence intervals on parameters and probability values on hypotheses that are more in line with commonsense interpretations (Congdon, 2003).

BSTS combines the two approaches of Bayesian inference and structural time series. The second component, time series in a state space form, can be described with an observation and a transition equation given by

$$\boldsymbol{y_t} = Z_t^T \alpha_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \tag{1}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t \qquad \eta_t \sim N(0, Q_t) \tag{2}$$

With random Gaussian noise ε_t and η_t with variance σ_ε^2 2 and Q_t respectively. Equation (1) provides information regarding the relation between the observed variables $\boldsymbol{y_t}$ and the latent state variables α_t . Equation (2) describes the transition behavior of the latent state variables over time. The model matrix Z_t , T_t, R_t typically contain a mix of known values (often 0 and 1), and unknown parameters. Notation in the form of a state space model is convenient because one can effortlessly add further components to the state vector, such as seasonality and trend. A Model can be obtained by adding a regression component to the popular "basic structural model." This model can be written

$$y_t = \mu_t + \tau_t \boldsymbol{\beta}^T \boldsymbol{X}_t + \varepsilon_t \tag{3}$$

Where y_t , μt, τ_t , ω_t , ξ_t *and* ε_t representing target time series, local linier trend component, seasonal component, regression component and observation error terms respectively.

$$\mu_t = \mu_{t-1} + \delta_{t-1} + \mu_t \quad u_t \sim N(0, \sigma_u^2)$$

$$\tau_t = \sum_{s=1}^{s-1} \tau_{t-s} + w_t \qquad w_t \sim N(0, \sigma_\tau^2)$$

$$\delta_t = \delta_{t-1} + v_t \qquad v_t \sim N(0, \sigma_v^2)$$

S represent the number of season for y and τ_t denotes their join contribution to the observed target time series y_t . As is typical in Bayesian data analysis, forecasts from our model are based on the posterior predictive distribution. It is trivial to simulate from the posterior predictive distribution given draws of model parameter and state from their posterior distribution. Let ỹ denote the set of values to be forecast. The posterior predictive distribution of ỹ is

$$p(\tilde{y}|y) = \int p(\tilde{y}|\phi) p(\phi|y) d\phi$$

3. Result

The unemployment rate in Indonesia, based on SAKERNAS data, continued to decline from 10.26 in February 2005 to 5.3 in August 2018. This make conclusion that open employment opportunities has been increased. But in particularly point of period, the unemployment rate does not always decrease, at certain times the unemployment rate higher than the previous time. For example, the unemployment rate on February 2018 higher than rate on August 2018. It was increased from 5.1 to 5.3. This happened because of the influence of seasonal factors and other influences such as depreciation of the rupiah

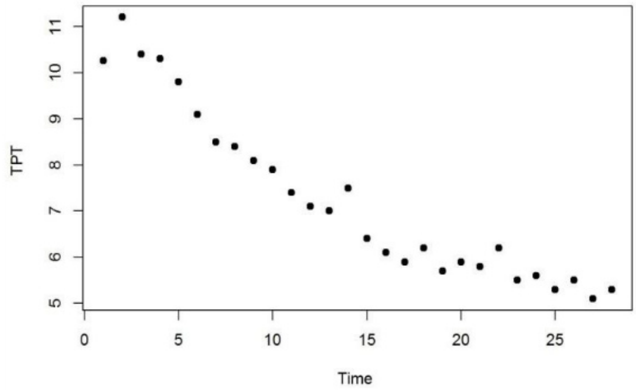

Figure 1. Unemployment Rate in Indonesia for the period 2005-2018

To predict the pattern of unemployment rates for the next recent period, GT data will be used based on 35 keywords that are relevant to Indonesia's unemployment rate. Based on 35 keywords, the most important keywords were selected using the stepwise regression method. Finally, four words are obtained, namely “bisnis indonesia”, “loker”, “bursa kerja”, and “job”. The macroeconomic variables that used in predicting unemployment rate are CPI and inflation rate. The interesting one that we can see from figure 2 is the correlation between Unemployment with “loker”. They have high negative correlation. The more people who search for this keyword then the unemployment rate will decrease.

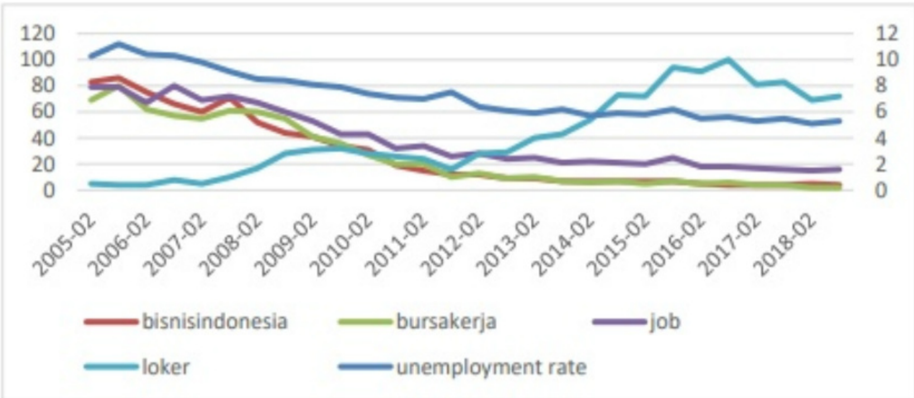

Figure 2. Inter-sample variation in keywords intensity

Based on plotting cumulative absolute error (fig 3), The performance of the prediction model with the ARIMA and Bayesian Time Series (BSTS) methods shows that the BSTS is able to produce better projection models than ARIMA. So the projection of the unemployment rate is continued with the BSTS method in several modeling scenarios. The first model, predicting unemployment rate using official data only. the second model, adding the 5 most important keywords into the first model. The other model, adding the other macroeconomic variables such as CPI or inflation into model 1 and model 2. The complete scenario model can be seen in figure 3.

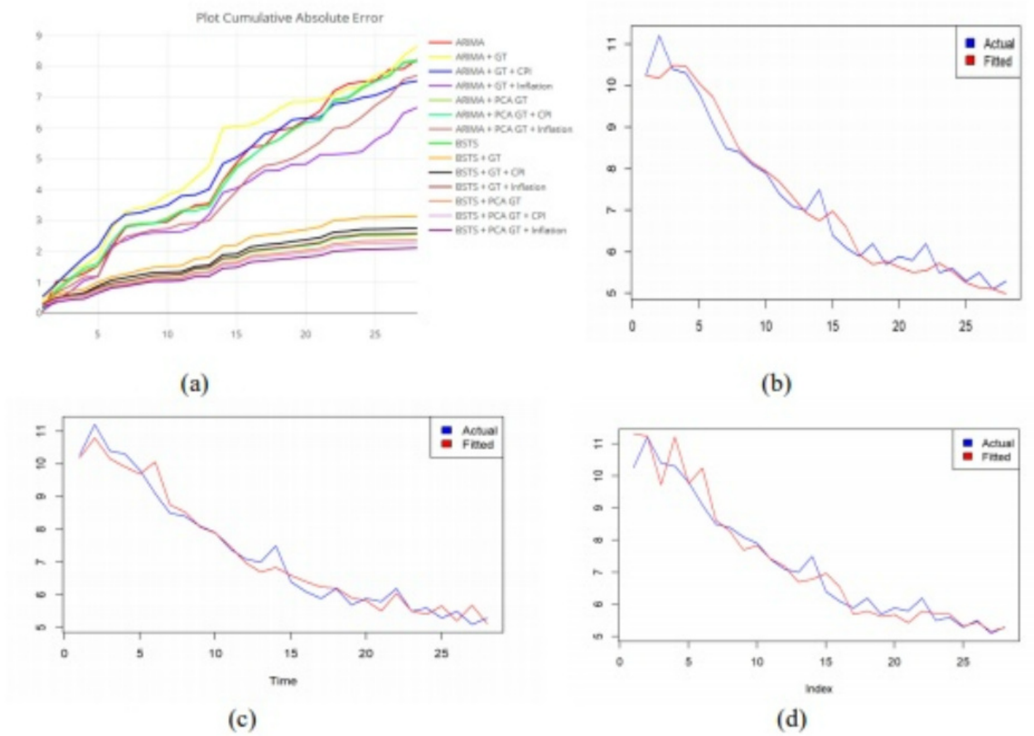

Figure 3. Plot (a) Cumulative Absolute Error for each scenario model, (b) ARIMA, (c) ARIMA+GT+Inflasi, (d)BSTS+PCA GT+Inflasi

The results of the unemployment projection using the ARIMA, ARIMAX and BSTS methods can be seen in figure 4. Projections with the BSTS method are more volatile and more relevant to official data than the ARIMA/ARIMAX method (figure 3). Based on the model BSTS and ARIMA, conditions of Unemployment Rate in February 2019 will be dcreased than August 2018.

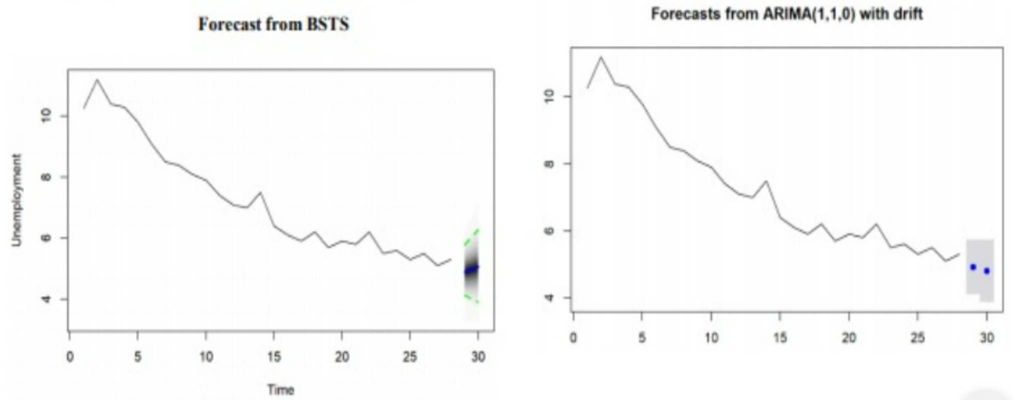

Figure 4. The Forecast plot with BSTS and ARIMA method

4. Discussion and Conclusion

The unemployment rate in Indonesia is predicted to fluctuate, this is due to seasonal influences. With Indonesia's background as an agrarian country, the dominance of sectors that absorb the most labor force in agricultural sector so that the harvest or non-harvest periods greatly affect the amount of unemployment. In addition, the stability of the rupiah value also affected fluctuations of the unemployment rate. Thus government efforts are needed to further open and enlarge employment opportunities for the Indonesian workforce as they reduce the effects of seasonal factors.

Based on the absolute error value, nowcasting the unemployment rate is better produced by BSTS than ARIMA method. the projection of BSTS can show seasonal effects such as the actual phenomenon of unemployment condition in Indonesia. While ARIMA was projecting the unemployment rate will decrease in time by the time. The use of Google Trend data as an additional information in projecting unemployment can improve nowcasting performance. Likewise, when combined with other macroeconomic variables, the involvement of Google trend data can improve the performance of the unemployment nowcasting model. Thus the role of Google trend data can increase the performance of the projection macroeconomic indicators.

References

1. Irfan, A. S. (2017). Tracking Labour Market Condition using Google Search Data-Indian Case Study.
2. Koop, G. (2013). Macroeconomic Nowcasting Using Google Probabilities.
3. Mitra, P., Sanyal, A., & Choudhury, S. (2017). Nowcasting Real Estate sector growth using Google Trend Data: An empirical retrospect for India. Reserve Bank of India, 1-32.
4. Stephens-Davidowitz, S., & Varian, H. (2015). A Hands-on Guide to Google Data. 1-25.
5. Congdon, Peter (2007). Introduction: The Bayesian Method, its Benefits and Implementation

9. Leffondre, K., Abrahamowicz, M. and Siemiatycki, J. (2003). Evaluation of Cox's model and logistic regression for matched case-control data with time-dependent covariates: a simulation study. Statistics in Medicine, 22, 3781–3794.

Index

Index

ISIWSC2019

Organised by :

DEPARTMENT OF STATISTICS MALAYSIA
MINISTRY OF ECONOMIC AFFAIRS

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

MALAYSIA INSTITUTE OF STATISTICS

Supported by:

MINISTRY OF TOURISM, ARTS & CULTURE MALAYSIA

Malaysia Convention & Exhibition Bureau

Visit Truly Asia Malaysia 2020

ISBN 978-967-2000-69-3

#ISIWSC2019