5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16039

A Machine Learning approach to constructing weekly GDP tracker using Google Trends

Jean Christine A. Armas[1], Cherrie R. Mapa[1], Ma. Ellysah Joy T. Guliman[1], Michael Lawrence G. Castañares[1], Genna Paola S. Centeno[1]
[1]Department of Economic Research, Bangko Sentral ng Pilipinas, Philippines

Abstract
The outbreak of the COVID-19 pandemic further highlighted the limitation of existing traditional indicators as policy formulation, particularly during crisis periods, demands timely and granular data. We construct the first Weekly Growth Domestic Product (GDP) Tracker in the Philippines using topic- and category- based Google Trends search volumes with the aid of machine learning models. We find that our Weekly GDP Tracker is a useful high-frequency tool in nowcasting economic activity. We also show that the machine learning-based GDP tracker outperforms the traditional autoregression models under study in terms of lower root mean square error (RMSE) for both train and test datasets. On the whole, our Weekly GDP Tracker can serve as a useful complementary surveillance tool for monitoring economic activity.

Keywords: *Nowcasting; GDP; Google Trends; machine learning models; neural networks.*

1. Introduction

Timely and accurate information are essential inputs for policy formulation. Data becomes even more important during crisis periods, such as the COVID-19 pandemic, as high frequency and granular data are integral in crafting prompt and appropriate policy responses that can help attenuate the impact of a crisis. However, official economic statistics are typically published with a significant time lag. In the case of the COVID-19 pandemic, an extra layer of challenge emerged as collection of official statistics was hampered by the imposition of mobility restrictions during the height of the health crisis. These motivate the interest of policymakers, including monetary authorities, to tap alternative data sources to supplement existing official statistics or traditional indicators.

In the area of macroeconomic surveillance, the Gross Domestic Product (GDP) is the official and most comprehensive indicator for measuring economic activity. In the Philippines, the GDP is available on a quarterly basis and is published by the Philippine Statistics Authority (PSA) 40 days after the reference quarter except for the Q4 GDP which is released after 30 days.[1] Due to this publication lag, the Bangko Sentral ng Pilipinas (BSP) uses a number of models to nowcast the GDP as information on the output growth and the cyclical position of the economy in the business cycle are important considerations in policy formulation. Thus far, the BSP's nowcasting models for the GDP have employed traditional statistics with nowcast updates implemented on a monthly or quarterly basis.

This study attempts to build a high-frequency indicator of GDP growth that capitalizes on the use of alternative data. Our objectives are: (a) to construct a weekly GDP Tracker that can nowcast quarterly GDP growth using machine learning models trained on pre-identified Google search volumes; and (b) to evaluate the usefulness of Google Trends data in nowcasting GDP pending the availability of official statistics. To ensure that the use of Google Trends data is suitable for statistical and economic analysis, the topic- and category-based searches are selected based on the authors' expert and sensible judgment on mapping relevant internet searches with the components of National Accounts of the Philippines (NAP). To the authors' knowledge, this is the first empirical research in the Philippines to use both the topic- and category- based Google Trend searches in nowcasting GDP.

2. Data: Google Trends Data Selection and Statistical Pre-processing

This study uses data from two sources: PSA and Google Trends. The quarterly real GDP data (with 2018 base year) are sourced from the PSA. No other statistical processing was

1 PSA, Technical Notes on the National Accounts of the Philippines, available online at https://psa.gov.ph/statistics/technical-notes/node/168102.

performed for the GDP data apart from taking the difference in the natural logarithm of the real GDP to derive the year-on-year GDP growth rates. Thus, this section delves into the selection of Google Trends indicators incorporated in our study and the statistical pre-processing steps applied to these indicators.

2.1. Google Trends Selection

Google Trends data is an analytical tool provided by Google that allow users to determine relative interest on a particular search term. It shows how frequently a search term is entered into Google's search engine relative to all Google searches for a particular geographical region and time. This study utilizes Google search volume indices for the main categories, selected sub-categories and authors' pre-identified topics that are relevant to the estimation of GDP growth (Table 1).

Table 1. List of Select Google Trends Variables

a. Google Trends categories and sub-categories		
Arts & Entertainment	Health	Pets & Animals
Autos & Vehicles	Hobbies & Leisure	Real Estate
Beauty & Fitness	Home & Garden	Reference
Books & Literature	Internet & Telecom	Science
Business & Industrial	Jobs & Education	Shopping
Computers & Electronics	Law & Government	Sports
Finance	News	
Food & Drink	Online Communities	
Games	People & Society	
Advertising & Marketing	Chemicals Industry	Pharmaceuticals & Biotech
Aerospace & Defense	Construction & Maintenance	Printing & Publishing
Agriculture & Forestry	Energy & Utilities	Retail Trade
Automotive Industry	Enterprise Technology	Small Business
Business Education	Entertainment Industry	Textiles & Nonwovens
Business Finance	Hospitality Industry	Transportation & Logistics
Business News	Industrial Materials & Equipment	
Business Operations	Manufacturing	
Business Services	Metals & Mining	
b. Google Trends topics		
Job	Subsidy	Unemployment Benefits
Pantawid Pamilyang	Tax	Investment
Pilipino Program	Unemployment	Resignation

Source: Google Trends; Authors' Selection

2.2. Data Pre-processing

Google Trends data series are collected on monthly (January 2004 - August 2022) and weekly (January 2014- August 2022) frequencies. For the monthly data series, corrections for known breaks in the time series data are implemented.

For the period 2004 to 2022, Google reported the following three (3) breaks in their time series data: (1) January 2011 due to geographical localization; as well as (2) January 2016 and (3) January 2022, both due to improvements in the data collection system. Only the breaks in 2011 and 2016 are corrected to address the potential issue of spikes in growth rates that may be attributed to changes in the data collection methods of Google.[2]

The breaks are addressed by introducing an adjustment that results into zero growth rate at the breakpoint. A backward correction approach was used to correct for the breaks starting from January 2016 back to January 2011. This approach, however, is a deviation from the study of Woloszko (2020) that made use of forward correction to address the breaks.[3]

3. Machine Learning Models

This study takes advantage of the empirically-documented ability of machine learning algorithms to generate relatively accurate and robust predictions. Machine learning models offer two key advantages: (a) capture the non-linearities in the data that could better explain the movements in real GDP especially during periods of extreme economic stress and heightened uncertainty; and (b) handle a wide array of variables without running into issues of overfitting through the multi-layer structure of machine learning models.

In this study, four (4) machine learning techniques are evaluated namely, Support Vector Regression (SVR), Decision Trees, Random Forest, and Artificial Neural Network (ANN). For all these models, the data are split into a train dataset (first 65 quarters) and a test dataset (last 5 quarters). Each model is trained to learn the patterns from the train dataset. Model performance in both the training and test sets are evaluated by computing for the root mean square error (RMSE). The results are also compared with the traditional autoregressive models. Table 2 below summarizes model performance based on RMSE for the train and test datasets. Notably, the ANN has outperformed other machine learning models and even the

---

[2] The impact of the January 2022 break on the data series is not yet evident at the time of the model construction.

[3] The backward break correction is done on each variable starting at the earliest date (i.e., 2004). For each variable, the difference between January 2011 (2016) and January 2010 (2015) is added to observations earlier than January 2011 (2016) inclusive. The break in 2022 will be addressed once there is evidence of a significant shift in the time series data.

traditional autoregression models. Therefore, it was chosen to be the main model for this study.

Table 2. Forecast evaluation of machine learning models vs. traditional time series models

	RMSE using Train data	RMSE using Test data
SVR	2.55	2.29
Decision Trees	0.50	12.41
Random Forest	1.04	5.62
ANN	0.53	1.49
ARIMA (1,1,1)	2.83	7.26
AR(1)	2.67	6.33

Source: Authors' estimates

4. Construction of the Weekly GDP Tracker

The Weekly GDP Tracker capitalizes on the use of a more frequent Google Trends data series to be able to extract leading information from this type of unconventional data source and thus, infer sensible predictions on economic activity or variations in business cycle especially during unprecedented periods.

The Tracker is constructed using a two-step model approach, broadly following Woloszko's (2020) methodology. First, the machine learner is trained to predict the quarterly GDP growth using topic- and category- based Google Trends searches, as outlined in the previous section. Second, by employing the frequency-neutrality assumption in Woloszko (2020), the estimated elasticities from the quarterly model were applied to the weekly Google Trends data series.

The application of elasticities from quarterly model on the weekly Google Trends requires calibrating the weekly Google Trends series to match the beak-corrected monthly Google Trends indices and using its 12-week moving average as input in the trained model.

5. Analysis of the Model Results

An important question on the use of the Weekly GDP Tracker is how well it can nowcast economic activity? Overall, the results show that the Weekly Tracker is a useful complementary surveillance tool to official statistics in terms of providing relevant leading information on the likely trend or path of output growth as well as capturing the important crisis episodes or business cycle fluctuations.

We find relatively good prediction performance of our chosen machine learning model, primarily due to its ability to find the optimal parameters that could fit the training dataset. At the same time, as presented in Table 2, the RMSE for both training and test datasets percent of the ANN are significantly lower when compared to the traditional time series models such as Autoregression (AR) and Autoregressive Integrated Moving Average (ARIMA). The predictive performance of our quarterly model is also assessed with regard to the movements of actual output growth during the 2020 COVID-19 pandemic. We note that the COVID-19 crisis period as well as the subsequent rebound following the downturn are well-captured by the predicted values for output growth. In particular, our Weekly GDP Tracker was able to capture about 96 percent of the slump observed in actual GDP growth in Q2 2020.

The weekly nowcast GDP growth from January 2015 to August 2022 is plotted against the actual quarterly year-on-year GDP growth in Figure 1. As shown, the Weekly Tracker closely tracks the general direction of the actual GDP, pointing to the usefulness of the Weekly Tracker in providing reasonable near real-time measurement of economic activity.

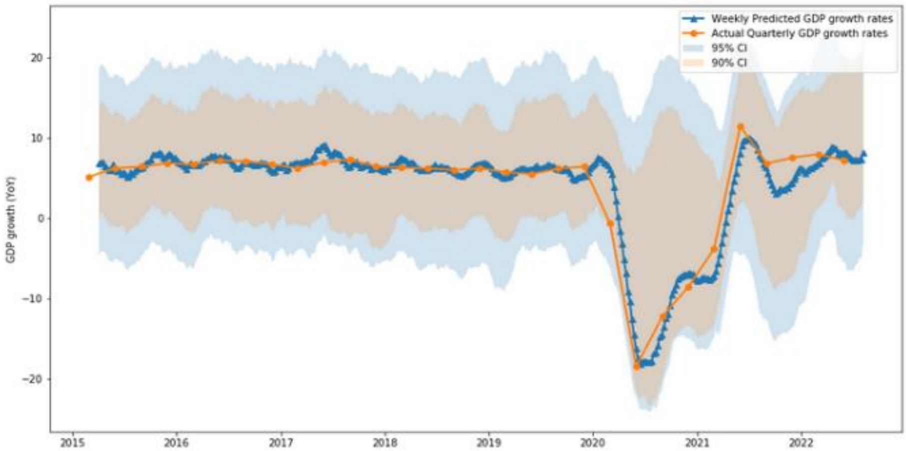

Figure 1. Weekly GDP tracker and actual quarterly GDP (in percent, %) (January 2015 – August 2022). Source: Authors' estimates

5.1. Most Important Predictors of GDP Growth Based on SHAP Values

One of the common problems with machine learning algorithms, especially with neural networks, is their black-box nature that makes analytical interpretation of the results difficult. One way to address this is to use an interpretability technique known as the Shapley values, which estimate the average marginal contribution of a variable to the prediction over all possible variable combinations or coalitions. This study uses the SHapley Additive exPlanations (SHAP) a fast algorithm implementation to extract Shapley values.

Based on the train dataset, Google searches for investment, unemployment, real estate, business news, and agriculture and forestry along with consumption-related searches for autos and vehicles and hobbies and leisure are found as the top contributors to the predicted GDP growth rate based on their SHAP values (Figure 4). Intuitively, higher google searches for investment and real estate may be correlated with higher economic growth while higher searches for unemployment may be associated with weaker GDP growth.

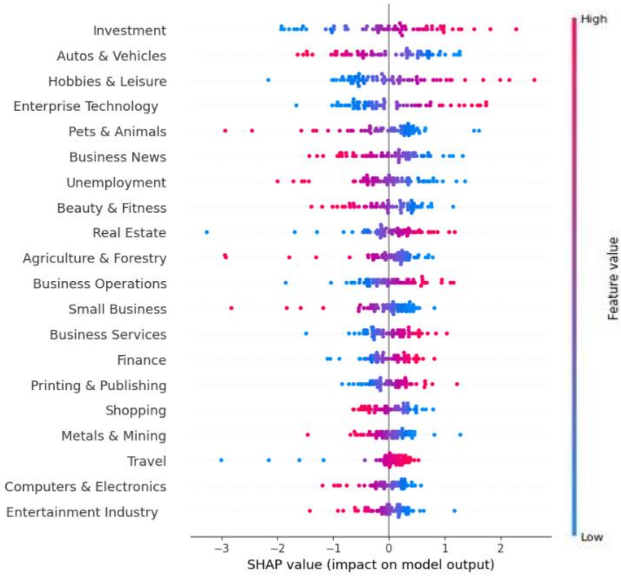

Figure 2. Most important predictors of GDP based on SHAP value. Source: Authors' estimates Note: The color bar in the illustration corresponds to the raw values of the variables for each instance on the graph. That is, variables with high and low values appear as red and blue dots, respectively. Meanwhile, the horizontal axis shows whether the effect of that value is associated with a higher or lower prediction. For example, high searches for investment (denoted by red dots) have positive impact on predicted GDP growth (shown in the horizontal axis).

6. Conclusion

The pandemic highlighted the importance of unconventional data sources, such as internet data, in macroeconomic surveillance. To the authors' knowledge, this is the first empirical research in Philippine literature to capitalize on the use of Google – the world's largest search engine– in near-real time tracking of output growth. To ensure that the use of Google Trends is suitable for statistical and economic analysis, topic- and category- based searches are selected based on the authors' expert and sensible judgment.

In this study, the Weekly GDP Tracker was constructed broadly following the approach of Woloszko (2020). The evaluation shows that the Weekly GDP Tracker is a useful high-frequency tool in tracking economic activity. The broad goal is not to replace the existing suite of economic models by the BSP but to contribute to surveillance by providing a high frequency monitoring tool that takes advantage of the readily available alternative data and the predictive capacity of advanced algorithms. Thus, pending the availability of quarterly national accounts, the Weekly GDP Tracker can serve as complementary surveillance tool of economic activity.

Like all models, prediction errors are anticipated, especially when the sample size to train or learn from is limited. Nonetheless, a key advantage of the weekly tracker is its greater flexibility to re-train and update nowcast predictions for GDP as new actual data becomes available. Hence, the model's learning capacity to predict output growth estimates can be further improved as new data comes in.

References

Armas, J. and P. Tuazon (2021). 'Revealing investors' sentiment amid Covid-19: the Big Data evidence based on internet searches' in Bank for International Settlements eds., New developments in central bank statistics around the world, vol. 55, Bank for International Settlements.

Assunção, J., Fernandes, P. (2022). 'Nowcasting GDP: An Application to Portugal'. Forecasting. 4, pp. 717-731, https://doi.org/10.3390/forecast4030039.

D'Amuri, F. and J. Marcucci (2012). 'The predictive power of google searches in forecasting unemployment.'. Bank of Italy Working Paper 891.

Gil, M., J. Perez, A. Sánchez, and A. Urtasun (2018): "Nowcasting private consumption: traditional indicators uncertainty measures, credit cards and some internet data". Working Paper 1842, Banco de España.

Haykin, S. (2009) Neural Networks and Learning Machines, 3rd edition, Pearson Prentice Hall.

Indaco, A. (2020). From twitter to GDP: Estimating economic activity from social media. Regional Science and Urban Economics.

Molnar, C. (2020), Interpretable Machine Learning: A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/.

Shapley, Lloyd S. (1953). "A value for n-person games." Contributions to the Theory of Games 2.28 pp. 307-317

Tuhkuri, J. (2016). 'Forecasting Unemployment with Google Searches'. ETLA Working Papers, No. 35 The Research Institute of the Finnish Economy (ETLA), Helsinki

Vosen, S. and Schmidt, T. (2011). 'Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends'. Journal of Forecasting. 30, pp. 565-578. DOI: 10.1002/for.1213

Woloszko, N. (2020), "Tracking activity in real time with Google Trends", OECD Economics Department Working Papers, No. 1634, OECD Publishing, Paris, https://dx.doi.org/10.1787/6b9c7518-en