Forecasting Number of COVID-19 Cases in Indonesia with ARIMA and ARIMAX Models

1st Bimo Satrio Aji
School of Computing
Telkom University
Bandung, Indonesia
bimooaji@student.telkomuniversity.ac.id

2nd Indwiarti
School of Computing
Telkom University
Bandung, Indonesia
indwiarti@telkomuniversity.ac.id

3rd Aniq Atiqi Rohmawati
School of Computing
Telkom University
Bandung, Indonesia
aniqatiqi@telkomuniversity.ac.id

2021 9th International Conference on Information and Communication Technology (ICoICT) | 978-1-6654-0447-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICoICT52021.2021.9527453

***Abstract*—During the pandemic COVID-19, Indonesia has a significant number of positive cases among countries in Asia. In early December 2020, the death rate in Indonesia had been reached more than 3%. Meanwhile, the daily number of positive is also continued to increase, it happens due to lack of anticipation rules made by local authorities and central government. Thus, the preventive step such forecasting becomes a major issue in the area of science and technology, to make all stakeholders well-prepared against this pandemic. This paper provides the performance of The Autoregressive Integrated Moving Average (ARIMA) to forecast several COVID-19 and also examines Auto Regressive Integrated Moving Average with exogenous variables (ARIMAX) model by considering Google Trends as an external variable. We consider a daily dataset from the official website of Jakarta's COVID-19 and the Google Trends data based on certain queries as external variables on March 1 - November 25, 2020. According to ARIMA and ARIMAX models, we have ARIMAX model with Google Trends improving ARIMA's performance by reducing the MAPE by 0.8 %.**

Keywords—Forecasting, COVID-19, ARIMA, ARIMAX, Google Trends.

I. INTRODUCTION

COVID-19 is a disease caused by a new coronavirus called SARS-CoV-2. Following a study of a cluster of cases of 'viral pneumonia' in Wuhan, China, WHO (World Health Organization) first learned of this new virus on December 31, 2019 [1]. Globally, as of 3:30 p.m. GMT +7, 12 December 2020, there are 69,143,017 confirmed cases of COVID-19, including 1,576,516 deaths, reported to WHO z[2]. COVID-19 has a variety of effects on different people. Fever, a dry cough, and tiredness are some of the most common symptoms [3].

Since the virus's first confirmation in Indonesia on March 2, 2020, the incidence rate of COVID-19 has been steadily increasing. On December 12, 2020, there were 611,631 confirmed cases of COVID-19 in Indonesia, with 18,653 deaths, according to data obtained from the Indonesian National Disaster Management Authority [4]. Controlling COVID-19 cases transmission is part of the Indonesian National Disaster Management Authority strategic plan [5]. Therefore, it is necessary to predict the impact to prepare an effective response strategy [6].

Previous study has shown that Internet search monitoring is consistent, efficient, and represents population patterns in real time, giving it strong potential to complement existing epidemiological methods [7]. One in every two EU people (53%) between the ages of 16 and 74 reported searching online for health information related to accidents, disease, nutrition, improving health, or similar topics in the three months leading up to the 2019 survey on ICT use in households and by individuals [8]. Thus, Internet search engines are now essential for Internet users to find any information. The behavior of how and when people search may provide information or early indicators about potential concerns and expectations because a large population of people searches online for medical information.

In order to analyze how search engine data applied for alternative surveillance data, we referred to other previous studies using Google Trends. Google Trends is a Google, LLC public website that provides Google Search data that shows how frequently a specific search term is entered in comparison to all other search terms in various regions and languages [9]. In the previous study, for some certain queries, Google Trends was associated with national surveillance data in South Korea using the influenza survey [10]. Although there has never been a study carried out on the correlation between COVID-19 cases and COVID-19 search queries in Indonesian, the study showed that Google Trends correlates with the actual data.

In this research, ARIMA and ARIMAX models are chosen to develop forecasting models of COVID-19 cases. Hopefully, our proposed methods can contribute to help the Indonesian National Disaster Management Authority in dealing with the increase of COVID-19 cases in the future. In a previous study, the ARIMAX model used Google Trends data as a predictor variable to predict the amount of dengue fever in Surabaya and resulted in a 3 percent improvement in the Mean Absolute Percentage Error (MAPE) value compared to using the ARIMA model [11]. In other cases, the ARIMAX method have shown an increase than ARIMA method in accuracy level of training, testing, and next time forecasting processes [12]. This research provides a comparative study of the ARIMA and ARIMAX models with the Google Trends query for forecasting COVID-19 cases in Indonesia.

II. LITERATURE REVIEW

A. Time Series Forecasting

Time Series is a sequence of observations made consecutively in time [13]. Forecasting is the process of predicting the future with as much accuracy as possible, using all available data, including historical data and knowledge of all future events that may affect the forecast [14]. Time series forecasting can be applied if there are conditions such as the availability of historical data information, where this information can be quantified in numerical form and it can be assumed that some aspects of past patterns will continue in the future.

B. ARIMA and ARIMAX

The Autoregressive Integrated Moving Average (ARIMA) model is a statistical model that uses past and present observations of the dataset. This model is useful in forecasting and evaluating the statistical relationship between the dataset to be predicted and the historical values of the dataset. An important aspect that must be considered in preparing the ARIMA model is stationarity. The data is said to be stationary if the statistical characteristics, such as mean and variance, are consistent over time. When viewed through the data plot, the data fluctuates only around the average. If some of the above characteristics are not met then the data is said to be non-stationary [15]. Stabilize the variance or mean by differencing the series or applying a logarithmic transformation to the dataset can be used to achieved stationary [16].

The formula of the $ARIMA\ (p, d, q)$ is given as follows:

$$\left(1 - \phi_p B - \cdots - \phi_p B^p\right)(1 - B)^d y_t = \left(1 + \theta_1 B + \cdots + \theta_q B^q\right)\varepsilon_t \qquad (1)$$

When y_t is the time observation time t, p is the Autoregressive (AR) parameter, q is the Moving Average (MA) parameter, d is the degree of first differencing involved and ε_t is an error at time t [14].

ARIMAX model or known as ARIMA with multiple regressors is a development of the basic ARIMA model with other external variables. The ARIMA model which is added with several variables that have a significant effect can increase the performance of the forecasting model [17]. In ARIMA and ARIMAX modeling, the accuracy of the forecasting results can be measured using Mean Absolute Percentage Error (MAPE) as follows:

$$\mathrm{M} = \frac{1}{\mathrm{n}} \sum_{\mathrm{t}=1}^{\mathrm{n}} \left| \frac{\mathrm{A_{t-F_t}}}{\mathrm{A_t}} \right| \times 100\% \qquad (2)$$

When n is the number of times the sum of the iteration occurs, A_t is the actual value and F_t is the forecast value, if the MAPE value is getting closer to 0, the accuracy will be higher. MAPE is popular among industry professionals because it is independent of scale and simple to read [18]. However, MAPE has a major disadvantage, when the real values are zero or close to zero, it generates infinite or unknown values, which is a common occurrence in some fields. MAPE produces extremely high percentage errors (outliers) when the actual values are small (usually less than one), whereas zero actual values produce infinite MAPE. Data with multiple zero values can be found in a variety of areas, including retailing, biology, and finance [19].

C. Google Trends

Google, LLC is a United States-based worldwide technology firm that focuses in Internet-related services and products such as web advertising technologies, search engines, cloud storage, software, and hardware. Along with Amazon, Facebook, Apple, and Microsoft, Google is considered one of the top five technology companies in the United States [20]. Google's web search service is well-known, and it is a major contributor to the company's growth.

Google Trends is a website owned by Google, LLC which contains trends in the use of keywords on the Google search engine website and trending news. Google Trends offers access to a sample of actual search requests made to Google that is largely unfiltered. It is anonymized (no one is identified personally), categorized (defining a search query topic), and aggregated (grouped). This allows us to demonstrate global interest in a topic, as well as interest at the city level [9]. The portal zdetermines the proportion of searches for a particular query made on Google Search. It also provides a Relative Search Volume (RSV), which is the query segment of a certain word for a specific location and time span, normalized by the highest query portion throughout time [21].

III. Data and Methodology

A. Data

This research used time series data from the total number of daily COVID-19 cases in Indonesia, instead of specific regions. The COVID-19 dataset was available at https://corona.jakarta.go.id. The ARIMAX model uses the Google Trend search index as an external variable. Google Trends search queries related to COVID-19 determined based on keywords related to COVID-19 searches. The queries used for ARIMAX modeling are “corona”, “covid”, “positif covid”, “covid-19”, “psbb”, and “demam”. In English language “positif” means positive. Other terms in Indonesian language “Pembatasan Sosial Berskala Besar” means large-scale social restrictions, shortened into “psbb”. In English “demam” means fever.

Google Trends dataset was downloaded on November 27, 2020 from the website of Google Trends available at https://trends.google.com. The geographical location is set to Indonesia. The dataset index range obtained is daily data, with a data range from March 1, 2020 - November 25, 2020. The dataset obtained from Google Trends shows that some of the keywords generated the string value "<1" due to the low trend on that date, so to handle it we round the value to 1. The COVID-19 and Google Trends dataset are divided into two: one set for the data training (parameter estimation), and another for data test. The data training and data test set ratio is 85:15, where the data training is implemented to build up a forecasting model while the data test (or validation) set is to validate and evaluate the model's accuracy.

B. Methodology

A time series of COVID-19 datasets will be used to construct the ARIMA model, followed by ARIMAX as external variables with Google Trends datasets. First, using the Augmented Dickey Fuller test, the stationary of the COVID-19 dataset will be tested, where the data is stationary if the p-value is less than 0.05 [22]. ARIMA is built on the concept that the dataset is stationary, that the variances and mean of the series are independent of time. In this research, we obtain stationary by differencing the data training. By removing variations in the time series level and thereby eliminating (or reducing) trends and seasonality, differencing will help to stabilize the time series average [14]. Differencing is performed by subtracting the previous observation from the current observation.

The order of p and q is determined from the result of differencing data using the Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF) after the COVID-19 dataset is stationary. Identification of the ARIMA model can be achieved by evaluating the following rules based on ACF and PACF plots [23].

TABLE I. ACF AND PACF IDENTIFICATION RULES

Model	ACF	PACF
$AR(p)$	Dies down (Drops exponentially)	Cuts off after lag p
$MA(q)$	Cuts off after lag q	Dies down (Drops exponentially)
$ARMA(p,q)$	Dies down (Drops exponentially)	Dies down (Drops exponentially)
$AR(p)$or $MA(q)$	Cuts off after lag p	Cuts off after lag q

The p and q values generated from the ACF and PACF analysis will be the limit combination in making the ARIMA model. Then, by using the COVID-19 dataset, we combine ARIMA's parameters (p, d, q) to find the best-fit parameters. Akaike's Information Criterion (AIC) method is used in selecting the best model by looking at the smallest AIC value. The model will be tested and sorted based on the five smallest AIC values.

The one that gives the lowest MAPE is the best-fit combination model. The MAPE is reported in this analysis as a percentage, which is the equation multiplied by 100. In calculating the MAPE value, the prediction results are obtained from the ARIMA model which is built from the training data and the actual data is the data test. The zbase zline zof zARIMAX zmodeling zwill be the ARIMA model with the best results.The best ARIMA model will perform each Google Trends queries as the external variables for the COVID-19 dataset. Similar to the ARIMA process, the best-fit combination model is chosen by the lowest MAPE. All ARIMA modeling and related statistical tests were carried out in Python using the Jupyter Notebook.

IV. RESULT AND DISCUSSION

The outcome of the ADF test for the time series COVID-19 is shown in Table II. The table shows that the non-differenced is not stationary ($d = 0$), and the series with first order of differencing is stationary ($d = 1$).

TABLE II. STATIONARY TEST RESULT

Differencing Order	ADF Statistic	p-value
$d = 0$	0.387229	0.981011
$d = 1$	-6.140016	8.013276e-08

Fig. 1 and Fig. 2 are the COVID-19 dataset log-transformed with differencing order = 0 and 1. The COVID-19 dataset is stationary after the first order of differencing, as shown in Table II, and the d parameter to be used is $d = 1$.

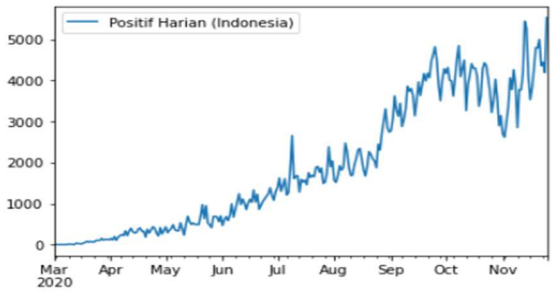

Fig. 1. The plot of log-transformed dependent variable d=0

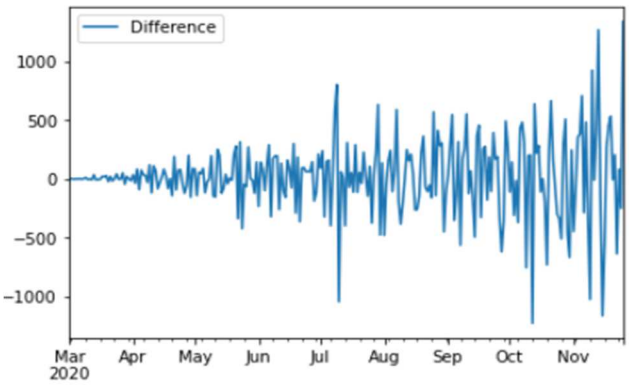

Fig. 2. The plot of log-transformed dependent variable d=1

As mentioned in the previous section, to obtain the p and q values we generated from the ACF and PACF. Figures 3 and 4 show the AFC and PACF plots for $d = 1$.

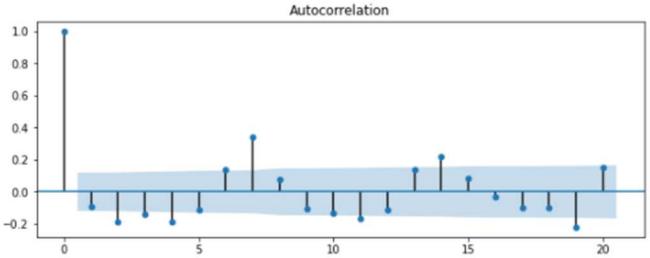

Fig. 3. The plot of ACF for d=1

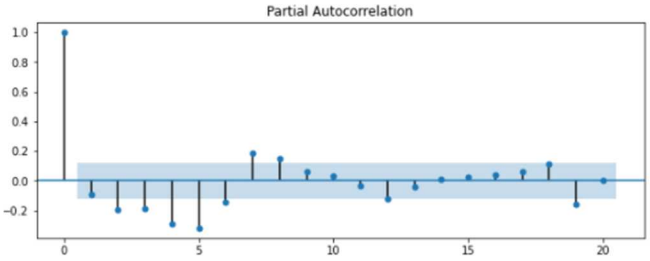

Fig. 4. The plot of PACF for d=1

Cut-off lags for ACF at lag 5 and cut-off lags for PACF at lag 9. The possible values of p and q in $ARIMA\ (p, d, q)$ are combinations of p and q from lag 0 to cut-off lag of p and q. These parameters produce 60 combinations of p, d, q and the 5 best parameter combinations that have the smallest AIC value are selected. Using the ARIMA method, the combinations are checked one by one to find the AIC value.

TABLE III. FIVE BEST MODEL COMBINATIONS WITH SMALLEST AIC

$ARIMA\ (p, d, q)$	AIC
ARIMA (5,1,2)	3113.083552
ARIMA (4,1,2)	3113.091031
ARIMA (6,1,2)	3113.160966
ARIMA (6,1,4)	3113.246304
ARIMA (6,1,3)	3113.316498

The model in table III is used to forecasting for the data training and calculating the MAPE value generated from each prediction. The model of ARIMA is chosen based on the smallest value of MAPE between models.

TABLE IV. THE MAPE RESULT OF EACH PARAMETER'S COMBINATION FOR THE ARIMA MODEL

$ARIMA\ (p,d,q)$	MAPE
$ARIMA$ (5,1,2)	13.41%
$ARIMA$ (4,1,2)	13.64%
$ARIMA$ (6,1,2)	13.19%
$ARIMA$ (6,1,4)	13.17%
$ARIMA$ (6,1,3)	13.20%

There have been 229 records of training data used, the resulted MAPE value varies from each model selected as shown in Table IV. The best fit parameter's combination for the ARIMA model is (6,1,4). The $ARIMA$ (6,1,4) model result in MAPE is 13.17%. To add Google Trends to the ARIMA model as an external variable, the model is tested using the ARIMA method with the addition of external variables. Table V shows the result of the ARIMA (6,1,4) model fitted to each query from Google Trends.

TABLE V. THE MAPE RESULT OF ARIMA(6,1,4) WITH EXTERNAL VARIABLE

$ARIMAX\ (p,d,q)$	Google Trends Queries	MAPE
$ARIMAX$ (6,1,4)	"corona"	13.14%
$ARIMAX$ (6,1,4)	"covid"	13.09%
$ARIMAX$ (6,1,4)	"positif covid"	13.10%
$ARIMAX$ (6,1,4)	"covid-19"	13.14%
$ARIMAX$ (6,1,4)	"psbb"	13.14%
$ARIMAX$ (6,1,4)	"demam"	13.37%

The result of this test indicates that there is a correlation between the number of COVID-19 cases with some Google Trends queries. $ARIMAX$ (6,1,4) with "covid" as an external variable has the best fit MAPE of the ARIMAX model, with a MAPE of 13.09%. According to the average value of the data used, the MAPE value shows that the model has good accuracy. The MAPE value of the ARIMAX model has decreased, although not significantly. From the ARIMAX model, we found that including Google Trends as an external variable decreased the MAPE by 0.8%. Table VI shows the performance comparisons between ARIMA and ARIMA models.

TABLE VI. PERFORMANCE COMPARISON OF ARIMA (6,1,4) AND ARIMAX (6,1,4) MODELS

Model	Google Trends Queries	MAPE
$ARIMA$ (6,1,4)	-	13.17%
$ARIMAX$ (6,1,4)	"covid"	13.09%

In Fig. 5 and Fig. 6 shows the fitted value of $ARIMA$ (6,1,4) and $ARIMAX$ (6,1,4) with "covid" as an external variable. The result of the COVID-19 forecast does not seem to meet COVID-19 observation data. However, the predictions are consistently predicted around the mean value of observed data which causes a small MAPE value.

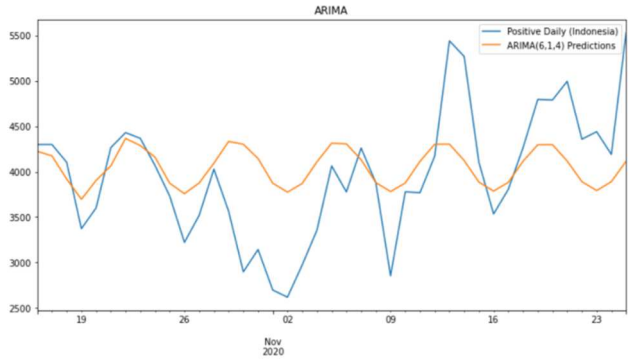

Fig. 5. The plot of COVID-19 forecasted by ARIMA(6,1,4)

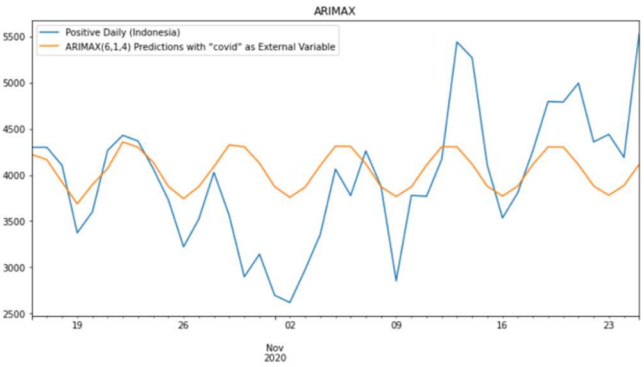

Fig. 6. The plot of COVID-19 forecasted by ARIMAX(6,1,4) with "covid" as an external variable

The results indicate that time series analysis techniques would be used to predict more complicated time series in the future, such as COVID-19 anomalies in a country or region, COVID-19 case prediction based on social media public opinion, and COVID-19 case end prediction in a country.

The addition of certain queries from Google trends as an external variable statistically increases the accuracy of the forecast model, but not on the graph. In this study, there are limitations to the use of Google Trends. Queries selection is done by selecting words related to COVID-19, without any prior survey. Although there is a decrease in the MAPE value in the ARIMAX model, the decrease is not significant. In futher research, the selection of keywords in Google Trends queries can be determined by conducting a survey to the public regarding their searches related to COVID-19.

V. CONCLUSION

According to the results of this research, the ARIMA approach can be implemented to forecast COVID-19 cases in Indonesia, with reliable accuracy in MAPE value of 13.17%. The best ARIMA model is ARIMA(6,1,4).The addition of Google Trends to forecast the number of cases of COVID-19 in Indonesia using the ARIMAX model leads to a 0.8 percent increase in model accuracy. However, in further studies, the selection process of Google Trends queries as an external variable must be refined.

REFERENCES

[1] WHO, "Coronavirus disease (COVID-19)." https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19 (accessed Dec. 12, 2020).

[2] WHO, "Coronavirus disease (COVID-19)." https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (accessed Dec. 12, 2020).

[3] WHO, "Coronavirus." https://www.who.int/health-topics/coronavirus#tab=tab_3 (accessed Dec. 13, 2020).

[4] National Agency for Disaster Management (BNPB) Indonesia, "Peta Sebaran | Satgas Penanganan COVID-19." https://covid19.go.id/peta-sebaran (accessed Dec. 12, 2020).

[5] National Agency for Disaster Management (BNPB) Indonesia, "Pemda Diminta Tingkatkan Penanganan Untuk Menurunkan Kasus Aktif - Berita Terkini | Satgas Penanganan COVID-19." https://covid19.go.id/p/berita/pemda-diminta-tingkatkan-penanganan-untuk-menurunkan-kasus-aktif (accessed Dec. 12, 2020).

[6] C. S. Lutz et al., "Applying infectious disease forecasting to public health: A path forward using influenza forecasting examples," BMC Public Health, vol. 19, no. 1, pp. 1–12, 2019, doi: 10.1186/s12889-019-7966-8.

[7] L. C. Madoff, D. N. Fisman, and T. Kass-Hout, "A new approach to monitoring dengue activity," PLoS Negl. Trop. Dis., vol. 5, no. 5, pp. 3–7, 2011, doi: 10.1371/journal.pntd.0001215.

[8] Eurostat, "53% of EU citizens sought health information online – Products Eurostat News-Eurostat." https://ec.europa.eu/eurostat/web/products-eurostat-news/-/DDN-20200327-1 (accessed Dec. 12, 2020).

[9] Google LLC, "FAQ about Google Trends data - Trends Help." https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052 (accessed Dec. 13, 2020).

[10] S. Cho et al., "Correlation between national influenza surveillance data and Google Trends in South Korea," PLoS One, vol. 8, no. 12, 2013, doi: 10.1371/journal.pone.0081422.

[11] W. Anggraeni and L. Aristiani, "Using Google Trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia," Proc. 2016 Int. Conf. Inf. Commun. Technol. Syst. ICTS 2016, pp. 114–118, 2017, doi: 10.1109/ICTS.2016.7910283.

[12] W. Anggraeni, R. A. Vinarti, and Y. D. Kurniawati, "Performance Comparisons between Arima and Arimax Method in Moslem Kids Clothes Demand Forecasting: Case Study," Procedia Comput. Sci., vol. 72, pp. 630–637, 2015, doi: 10.1016/j.procs.2015.12.172.

[13] G. M. L. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Time Series Analysis: Forecasting and Control, 5th Edition. 2016.

[14] R. J. Hyndman and G. Athanasopoulos, Forecasting : Principles and Practice. 2018.

[15] Statistics Office of Indonesia (Badan Pusat Statistiks), Seasonal Adjustment dan Peramalan PDB Triwulan. 2010.

[16] R. P. Soebiyanto, F. Adimi, and R. K. Kiang, "Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters," PLoS One, vol. 5, no. 3, pp. 1–10, 2010, doi: 10.1371/journal.pone.0009450.

[17] D. E. Hinkle, W. Wiersma, and S. G. Jurs, "Applied statistics for the behavioral sciences." Houghton Mifflin ; [Hi Marketing] (distributor), Boston, Mass.; [London], 2003, [Online]. Available: http://catalog.hathitrust.org/api/volumes/oclc/50716608.html.

[18] R. Byrne, "Beyond Traditional Time-Series: Using Demand Sensing to Improve Forecasts in Volatile Times," J. Bus. Forecast., vol. 31, no. 2, 2012.

[19] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," Int. J. Forecast., vol. 32, no. 3, 2016, doi: 10.1016/j.ijforecast.2015.12.003.

[20] Wikipedia,"Google-Wikipedia." https://en.wikipedia.org/wiki/Google (accessed Dec. 14, 2020).

[21] H. Choi and H. Varian, "Predicting the Present with Google Trends," Econ. Rec., vol. 88, no. SUPPL.1, pp. 2–9, Jun. 2012, doi: 10.1111/j.1475-4932.2012.00809.x.

[22] E. E. Holmes, M. D. Scheuerell, and E. J. Ward, Applied Time Series Analysis for Fisheries and Environmental Sciences. 2020.

[23] S. G. Makridakis, S. C. Wheelwright, and R. J. Hyndman, Forecasting: Methods and Applications. Wiley, 1997.