Adu *et al.*
Journal of Electrical Systems and Inf Technol (2023) 10:12
https://doi.org/10.1186/s43067-023-00078-1

Journal of Electrical Systems and Information Technology

REVIEW Open Access

VAR, ARIMAX and ARIMA models for nowcasting unemployment rate in Ghana using Google trends

Williams Kwasi Adu[1*], Peter Appiahene[1] and Stephen Afrifa[2]

*Correspondence:
cplexadu@gmail.com

[1] Present Address: University of Energy and Natural Resources, Sunyani, Ghana
[2] Tianjin University, Tianjin, China

Abstract

The analysis of the high volume of data spawned by web search engines on a daily basis allows scholars to scrutinize the relation between the user's search preferences and impending facts. This study can be used in a variety of economics contexts. The purpose of this study is to determine whether it is possible to anticipate the unemployment rate by examining behavior. The method uses a cross-correlation technique to combine data from Google Trends with the World Bank's unemployment rate. The Autoregressive Integrated Moving Average (ARIMA), Autoregressive Integrated Moving Average with eXogenous variables (ARIMAX) and Vector Autoregression (VAR) models for unemployment rate prediction are fit using the analyzed data. The models were assessed with the various evaluation metrics of mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), median absolute error (MedAE), and maximum error (ME). The average outcome of the various evaluation metrics proved the significant performance of the models. The ARIMA (MSE = 0.26, RMSE = 0.38, MAE = 0.30, MAPE = 7.07, MedAE = 0.25, ME = 0.77), ARIMAX (MSE = 0.22, RMSE = 0.25, MAE = 0.29, MAPE = 6.94, MedAE = 0.25, ME = 0.75), and VAR (MSE = 0.09, RMSE = 0.09, MAE = 0.20, MAPE = 4.65, MedAE = 0.20, ME = 0.42) achieved significant error margins. The outcome demonstrates that Google Trends estimators improved error reduction across the board when compared to model without them.

Introduction

The vast amount of information provided by the internet such as Google [1, 2], Twitter [3], social media [4], or combinations of web-based data sources [5, 6] have necessitated its numerously used in recent decades to find the potential of digital information for predictions in a wide range of sectors. Study reviews that Google handles over 92% of all online search requests in the world [7], and has demonstrated to be valid [8], valuable [9], accurate [10], and beneficial [11] for predictions. Google Trends has proven to be a dependable source of trend data for online searches and it is being extensively used by researchers around the world mostly for a real-time prediction of macroeconomic trends [12, 13].

Information that people provide through the internet describes the current state of the people and offers a good understanding mostly of the economic processes, particularly

unemployment [14, 15]. Upon all these useful online sources with all the availability of high-frequency data and recent technological advancement, statistical information published on unemployment by nations is released with delays and may still be revised [16, 17]. The way of gathering data for unemployment estimation seems exasperating making it impossible to know how the economy is performing right now but only how it was several months or years ago. This challenge is almost common in all countries, with Ghana not an exemption. This results in Policymakers making assessments in real-time using inadequate information, and knowing the present unemployment state which could help them better understand whether an economy is contracting or expanding and respond [18]. This paper tackles the case by using real-time Google trends data for prediction of unemployment claims in Ghana.

According to the Ghana Statistical Service's most recent census, Ghana's UnEmployment Rate (UER) increased to 13.4% in 2021, up from 6% in 2010, with 32.8% of Ghanaians aged 15 to 24 unemployed. Ghana faces a desperate downturn in economy, and the economy robust growth over the last two decades has not converted into job creation or improved employment circumstances [19]. This unfortunate situation and pressure on jobs have resulted in the loss of hundreds of jobs [20]. It would be of communal interest to produce real-time estimates of the unemployment rate to help policy making to produce real-time unemployment rate. The novelty of this paper is as follows:

- this is the current paper that considers the use of ARIMA, ARIMAX, and VAR in predicting unemployment rate in Ghana.
- the paper considers Google Trends indicators to predict unemployment rate in Ghana, which in turn can be used for the West African sub region.
- the paper is the current to consider unemployment rate predictions in the literature.
- the current paper provides the strategies and benchmarks for governments, agencies and organizations to make informed decisions on unemployment in Ghana, Africa, and the world as a whole.

The rest of the paper is organized as follows: The next section discusses related literature on forecasting using online search data. Section "Methodology" describes the methodology used for identifying a large number of keywords that may help in the prediction of unemployment claims, also provides a brief overview of the models used for comparison of results. The results of the models are discussed in section "Results and discussion." Section "Conclusion and future works" gives the conclusion and discusses the importance of using different categories of keywords for the prediction of the unemployment claims.

Related works

Online search engines are frequently used for real-time research. Due to the huge amount of daily search queries, Ettredge et al. [21] took the first initiative by first looking into how real-time forecasting may be done by using the Internet and the study's findings reveal a strong link between Internet-related web search activity and unemployment rate in the USA [22, 23] continued by looking at how web search data, particularly Google,

could be utilized to improve forecasting of a range of economic parameters, such as jobless claims, retail sales, real estate demand, and vacation destination preferences. Several studies of real-time forecasting utilizing internet data, particularly Google Trends (GT) data, have been published since these papers, but this work focuses on unemployment prediction.

To anticipate UERs during the COVID-19 pandemic in Indonesia, Rizky et al. [2] used GT data query share for the keyword "phk" (work termination) and earlier series from the official labor force survey performed by Badan Pusat Statistik (Statistics Indonesia). As a result of using the GT index query as an exogenous variable to capture current conditions of a phenomenon that is occurring, results of predicting open UER using ARIMAX during the COVID-19 period generate forecast values that are reliable and near to reality. Petropoulos et al. [24] used text mining algorithms to develop a financial lexicon based on a collection of 10,000 Central Bank speeches. Google inquiries, according to experts, can predict future market volatility in a short time (one month). Tuhkuri [25] used the ETLAnow model and no Google search data to estimate official UER in the European Union (EU) - 28 countries. Google Inc.'s Google Trends database, as well as Eurostat's Labor Force Statistics, are the model's primary data sources. Findings suggest that Google searches are linked to the EU UER, even after controlling for country-level, delayed, and seasonal effects.

Tuhkuri [26] used GT's database from Google Inc. and Labor Force Statistics from the Current Population Survey and US Bureau of Labor Statistics. Results reveal that Google searches' predictive ability is inadequate for short-term forecasting, that the utility of Google data for forecasting purpose is occasional, and forecasting accuracy increases are relatively modest. Mulero and García-Hiernaux [1] used data from GT and the Spanish State Employment Service to examine a large number of potential explanatory factors for UERs. The results reveal an increase in expected accuracy of 10% to 25%.

Lasso and Snijders [27] adopted GT method to forecast Brazil's UER. The findings reveal that Google search volumes for job-related phrases have significant predictive power, with biweekly search data forecasting the direction of the UER with over 80% accuracy, exceeding baseline methodologies based on seasonal trends by over 15%. Brake and Ramos [28] estimate the UER in the Netherlands using a variable based on the amount of Google search keywords. The predictive capability of the Google Indicator is determined by comparing the accuracy of a benchmark model to an upgraded model with the Google Indicator. According to the statistics, the Google improved models produce up to 27.8% more accurate estimations when considering a one-month forecast horizon.

Simionescu and Zimmermann [14] looked into how internet usage information is used in various industries, with unemployment modeling being a particular area of interest. The results of the research show that there is a lot of potentials that should be investigated further. A vast majority of nations base their unemployment estimation and modeling on internet data. However, the forecast's accuracy is based on each country's internet penetration, the age distribution of online users, and the stability of the generated internet variables. Maas [29] studied if Google search data, and other more traditional predictor elements, may be utilized to anticipate the UER in the USA. The

findings indicate that GT forecasting methods proposed in this study are most beneficial in short term.

Jung and Hwang [30] constructed unemployment prediction models for specific age groups using Google search queries related to them (the 30s and 40s) and known unemployment statistics from Statistics Korea. The findings demonstrate that employing a web search query to improve unemployment prediction models for Korea is still useful. Smit [31] investigates whether and to what extent Google search data may be utilized to forecast the US UER. They concluded that GTs enhances the anticipated accuracy of all currently used forecasting approaches.

Methodology

The study explored the effectiveness of the Google trends by adopting several testing techniques. Figure 1 displays a detailed procedure for the experiment. The steps below are a detailed explanation of Fig. 1.

1. To start, data from GT were joint with interpolated World Bank (WB) UER data to create a single, special dataset for the visualization and study of UER in Ghana using Granger causality.
2. Time Series (TS) data are split into training and test sets after input.

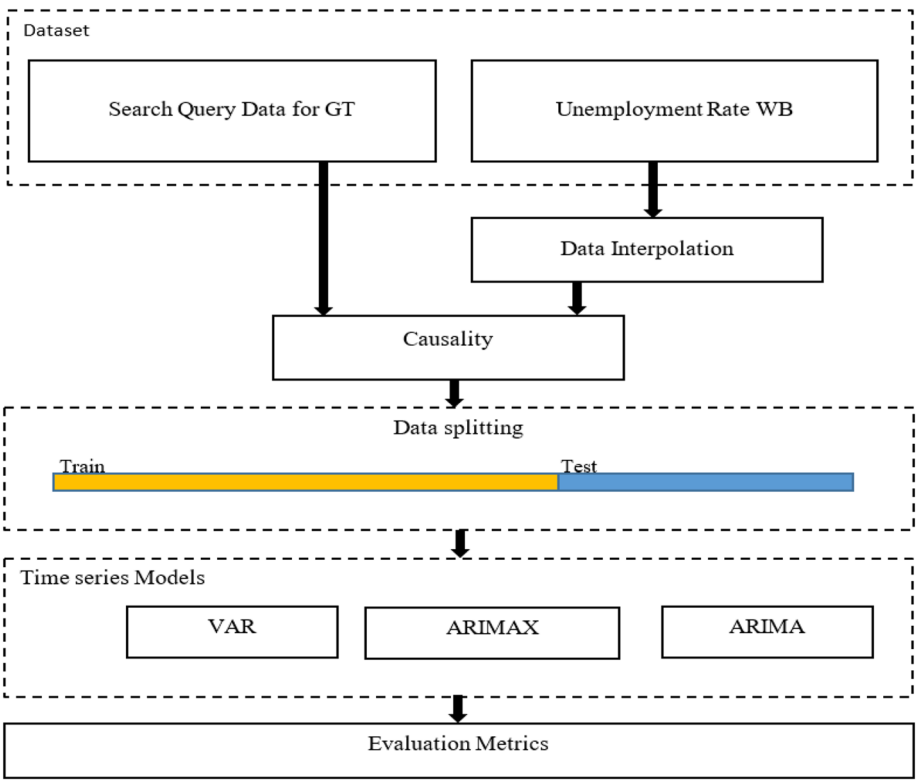

Fig. 1 Work flow of experimental design

3. Training sets and test sets were used to train and evaluate the models (ARIMA, ARIMAX, and VAR).

Data

The World Bank (WB) and Google Trends (GT) provided the data for this collection. Google launched the website Google Trends for search analysis in 2006. GT offers a search trend that starts with the year 2004 and shows the frequency with which a certain search phrase is entered into Google's search engine over time about the site's overall search traffic.

GT shows changes in internet interest for any TS in any nation or location over a selected period of time, such as one year, several years, four months, three weeks, thirty days, seven days, four hours, or one hour. Additionally, several sentences from various places can be compared simultaneously. The GT and World Bank data can be downloaded in ".csv" format. In short, GT calculates the number of searches represented mathematically in equation 1, 2, and 3 as follows [32]:

$$S(e)_{tot,m} = \sum_{k=1}^{\infty} s(e)_{k,m} \tag{1}$$

$$Qs(e)_{i,m} = \frac{S(e)_{i,m}}{S(e)_{tot,m}} \tag{2}$$

$$\mathrm{RSV}(e)_{i,m} = 100 * \frac{Qs(e)_{i,m}}{Qs(e)_{\max,m}} \tag{3}$$

where i = Terms or expressions of the study, k = possible terms to search on Google, m = months of the study. Additionally, $S(e)_{\mathrm{tot},m}$ = total search on Google for one-month m in a particular country, $S(e)_{i,m}$ = total searches on Google for a term i of our study for a month m and a country, $Qs(e)_{t,m}$ = Query share of a term in a certain month and country, and $\mathrm{RSV}(e)_{t,m}$ = Relative search volume of a term in a certain month and country.

Our sample of search terms comprises 50 Google Trends which have been chosen based on the methodology as shown in Table 1. Our data window is restricted

Table 1 GT Keywords (50)

Acceptance letter	Distance education	Graduation	Loan application	School admission
American lottery	Distance learning	Health insurance	Mining jobs	Trade
Application for employment	Employment letter	How to make money	Nursing schools	Trading
Application letter	Employment	How to start a business	Nursing training	Training college
Business opportunities	Entrepreneur	Income tax	Online application	Unemployment
Career	Exchange rate	Job application	Online jobs	USA jobs
Companies in Ghana	Foreign exchange	Job interview	Online money	USA visa
Cover letter	Ghana economy	Job opportunities	Online schools	Vacancy
Curriculum vitae	Ghana jobs	Jobs in Ghana	Police recruitment	Visa application
Cv	Graduate	Jobs in USA	Scholarship	Visa

to begin with 2010–2020 - since this is the earliest data point for which Ghanaian migrated to using internet. The variable of interest is the unemployment rate date for Ghana downloaded from the World Bank website.

Interpolation

For extracting high-frequency data (such as monthly or weekly data) from low-frequency data (such as annual data), the Chow-Lin approach, a disaggregation method, is utilized [33]. The method makes sure that the high-frequency series' average, first, and last values correspond to those of the low-frequency series. The following two-step additive structure is the general temporal disaggregation framework for developing a high-frequency estimate, according to [33]. Equation 4 describes the Chow-Lin approach.

$$\upsilon_j = \overline{\upsilon}_j + \sum_{i=1}^{n} F_{ji}\left(y_i - \sum_{q=1}^{m} H_{iq}\overline{\upsilon}_q\right) \tag{4}$$

Make a preliminary high-frequency series $\overline{\upsilon}_j$ using auxiliary data from several indicator series. To incorporate this data, a generalized least squares regression strategy is frequently utilized. Analyze the differences in residuals between the observed low-frequency series and the high-frequency series that have been aggregated to the low-frequency scale (through the matrix $H \in f^{n\times m}$). Then, create a temporally consistent high-frequency version y_i by distributing these differences among the high-frequency periods using the distribution matrix $F \in R^{n\times m}$.

Causality (granger causality (GC))

GC test examines the connection between the current value of one variable and the historical values of another variable to find a causal direction between two or more time series [34]. According to [35] GC indexes of two series Y and X can be computed by finding the variance of the error samples. If X and Y are independent, then $X(var(\varepsilon)) = Y(var(\varepsilon))$, where $var(\varepsilon)$ denotes the variance of the error e. Otherwise, the two equations do not hold. For example, if X is the cause of Y, then $X\,(var(\varepsilon)) > Y\,(var(\varepsilon))$. It can be represented by the formula in Eq. 5 [36]

$$\begin{aligned} F_{(X\to Y)} &= \log \frac{X_{(var(\varepsilon))}}{Y_{(\mathrm{var}(\varepsilon))}} \\ F_{(Y\to X)} &= \log \frac{Y_{(var(\varepsilon))}}{X_{(\mathrm{var}(\varepsilon))}} \end{aligned} \tag{5}$$

If $F_{(X\to Y)} \geq 0$ and $F_{(Y\to X)} \geq 0$ then the indexes of causality can be analyzed. Specifically, if $F_{(X\to Y)} > F_{(Y\to X)}$, then X is the cause of Y, or the information flowing from X to Y is more than that from Y to X; if $F_{(X\to Y)} < F_{(Y\to X)}$, then Y is the cause of X.

Training, and test

The overall data set was split into training and test data sets with the shares close to 80% from 2010 to 2018 dataset, with the remaining 20% from 2019 to 2020 designated for testing. Table 2 shows specific splitting procedure that divides the dataset. In the

Table 2 Training, Test Sample

Data / train / Test	Year(s)	GT (50)	UER %	Percentage %
Total size 2010–2020	10 years	574	574	100% of data size
Training 2010–2018	8 years	470	470	80% of data size
Test set 2019–2020	2 years ($Y_{24/12}$)	104	104	20% of data size (100% of test set)
	1 year 9 months ($Y_{21/12}$)	91	91	88% of test set
	1 year 6 months ($Y_{18/12}$)	78	78	75% of test set
	1 year 3 months ($Y_{15/12}$)	65	65	63% of test set
	1 year ($Y_{12/12}$)	52	52	50% of test set
	9 months ($Y_{9/12}$)	39	39	38% of test set
	6 months ($Y_{6/12}$)	26	26	25% of test set
	3 months ($Y_{3/12}$)	13	13	13% of test set
	1 month ($Y_{1/12}$)	4	4	4% of test set

second step, the test set of two years frames is further divided into yearly (Y1), half-year, quartile, and monthly such that UER was tested in the different time frames.

Models

The data science project of TS forecasting is crucial for many processes that happen over time. TS forecasting is a practical method for figuring out how past data influence present results. Making short- and long-term projections and pattern-spotting using previous data allows for this. The TS used were ARIMA, VAR and ARIMAX.

VAR

VAR is a forecasting method that can be used when two or more TS interact. In other words, the TS in question has a two-way relationship. VAR models can be used to assess and predict multivariate TS data, which sets them apart from univariate autoregressive models. VAR models are often used in economics. For a VAR model with a large number of interconnected TS variables. Equation 6 represents the VAR model

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} + \begin{bmatrix} \phi_{11} \cdots \\ \phi_{21} \cdots \\ \vdots \\ \phi_{n1} \cdots \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{n,t-1} \end{bmatrix} + ... + \begin{bmatrix} \phi_{1,t-p} \cdots \\ \phi_{2,t-p} \cdots \\ \vdots \\ \phi_{n,t-p} \cdots \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{n,t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (6)$$

where the c is the intercept, ϕ coefficient of lags of y till order p, and ε error. Here, it is shown as a system of equations with one equation per TS variable. VAR is adaptable, requires less time and information [37], and makes it simple to integrate additional data [38]. VAR models, however, have the drawback of being unable to take into account when the measure of the dispersion between numbers in a data set changes across various time series values [39].

ARIMA

ARIMA combines the ideas of autoregression and moving average to provide forecasts that are linear combinations of previous variable values and forecast errors. ARIMA

is characterized by three factors: p, d, and q signify the number of lagged (or previous) data to consider for autoregression, the number of times the raw observations are differenced, as well as size of the moving average window, respectively.

The forecasting equation is structured in Eq. 7 as follows:

$$F_t = L_t + \Omega_1 D_{t-1}^{'} + \cdots + \Omega_p D_{t-p}^{'} + \beta_q E_{t-1} + \cdots + \beta_q E_{t-q} \tag{7}$$

where F_t = forecast point at time t, L_t = Level at time t (straight line approximation of all your data at one time point—calculated in ARIMA, it uses the mean of differenced data time smoothing constants), $D_{t-p}^{'}$ = Previous difference observed data points, E_{t-q}= Error in prediction on previous data points, and Ω and β are smoothing constants.

Many scholars who used time series recently explored ARIMA. However, the ARIMA model only applies to one variable, does not adequately describe some data turning points, and cannot adequately convey relationships between variables [40, 41]. As a result, it is insufficient to describe genuine issues.

ARIMAX

The ARIMAX model is an extension of the ARIMA model. The model includes other independent variables that are the X added to the end and stands for "exogenous variables." This involves adding a separate different outside variable to help measure our endogenous variable.

Equation 8 is structured as follows:

$$\Delta Pt = c + \beta X + \phi 1 \Delta Pt - 1 + \theta 1 \int t - 1 + \int t \tag{8}$$

where Pt and $Pt-1$ represent the values in the current period and 1 period ago, respectively. Similarly, ϵt and $\epsilon t-1$ are the error terms for the same two periods. C is just a baseline constant. $\phi 1$ and $\theta 1$, express what parts of the value $Pt-1$ and error $\epsilon t-1$ last period are relevant in estimating the current one. β is a coefficient which will be estimated based on the model selection and the data. X is the exogenous variable of interest. ARIMAX is helpful since it combines the time series and regression components into one model. However, it can be challenging to interpret the independent variable that may have an impact on the result.

Evaluation metrics

We compute the mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), median absolute error (MedAE), and maximum error (ME) to assess the forecasting accuracy of each model. Equations 9, 11, 12, 13, 14, and 14 represent the aforementioned evaluation metrics.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{9}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right|}{n} \tag{10}$$

Table 3 Descriptive statistics for UER and Interpolated UER

Statistics	Unemployment rate	interpolation unemployment rate
Count	11.000000	574.000000
Mean	5.358182	5.357195
Std	0.981140	0.953488
Min	4.120000	4.059191
25%	4.375000	4.305773
50%	5.450000	5.446064
75%	6.095000	6.218123
max	6.810000	6.971820

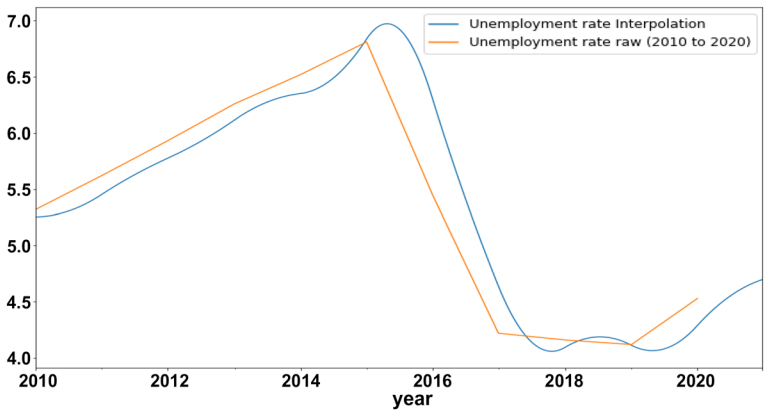

Fig. 2 Unemployment rate and an interpolation unemployment rate

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{11}$$

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \tag{12}$$

$$\mathrm{MedAE} = \mathrm{median}\left(\left|y - \widehat{y}_i\right|, \ldots, \left|y_n - \widehat{y}_n\right|\right) \tag{13}$$

$$\mathrm{ME} = \max\left(\left|y - \widehat{y}_i\right|, \ldots, \left|y_n - \widehat{y}_n\right|\right) \tag{14}$$

where y denotes current UER and $\widehat{y}$ is expected UER. Our study used six (6) different valuation metrics to evaluate the models. By employing more evaluation metrics, we were able to choose the optimum strategy while also confirming that each model was able to complete the underlying predicting task.

Table 4 GCT Analysis

Granger causality Test (squared residual (SSR) based *F* test)				
Acceptance letter Lag 1 $p(v)$: 0.2038 Lag 2 $p(v)$: 0.0041	Distance education Lag 1 $p(v)$: 0.2601 Lag 2 $p(v)$: 0.1064	Graduation Lag 1 $p(v)$: 0.0968 Lag 2 $p(v)$: 0.4066	Loan application Lag 1 $p(v)$: 0.2634 Lag 2 $p(v)$: 0.0340	School admission Lag 1 $p(v)$: **0.0393** Lag 2 $p(v)$: 0.0214
American lottery Lag 1 $p(v)$: 0.7104 Lag 2 $p(v)$: 0.8864	Distance learning Lag 1 $p(v)$: 0.5602 Lag 2 $p(v)$: 0.8839	Health insurance Lag 1 $p(v)$: **0.0019** Lag 2 $p(v)$: 0.0574	Mining jobs Lag 1 $p(v)$: 0.7852 Lag 2 $p(v)$: 0.1818	Trade Lag 1 $p(v)$: 0.3928 Lag 2 $p(v)$: 0.7702
Application for employment Lag 1 $p(v)$: 0.1214 Lag 2 $p(v)$: 0.0350	Employment letter Lag 1 $p(v)$: 0.4170 Lag 2 $p(v)$: 0.5553	How to make money Lag 1 $p(v)$: **0.0000** Lag 2 $p(v)$: 0.0000	Nursing schools Lag 1 $p(v)$: **0.0125** Lag 2 $p(v)$: 0.1505	Trading Lag 1 $p(v)$: **0.000** Lag 2 $p(v)$: 0.000
application letter Lag 1 $p(v)$: 0.9501 Lag 2 $p(v)$: 0.1191	Employment Lag 1 $p(v)$: **0.0279** Lag 2 $p(v)$: 0.0353	How to start a business Lag 1 $p(v)$: **0.0000** Lag 2 $p(v)$: 0.0000	Nursing training Lag 1 $p(v)$: 0.2080 Lag 2 $p(v)$: 0.4708	training college Lag 1 $p(v)$: 0.4841 Lag 2 $p(v)$: 0.4856
Business opportunities Lag 1 $p(v)$: 0.5516 Lag 2 $p(v)$: 0.4648	Entrepreneur Lag 1 $p(v)$: 0.4588 Lag 2 $p(v)$: 0.1039	Income tax Lag 1 $p(v)$: **0.0012** Lag 2 $p(v)$: 0.0156	Online application Lag 1 $p(v)$: 0.4383 Lag 2 $p(v)$: 0.3724	Unemployment Lag 1 $p(v)$: 0.0970 Lag 2 $p(v)$: 0.1388
Career Lag 1 $p(v)$: 0.0843 Lag 2 $p(v)$: 0.0015	Exchange rate Lag 1 $p(v)$: 0.7634 Lag 2 $p(v)$: 0.7559	Job application Lag 1 $p(v)$: 0.1591 Lag 2 $p(v)$: 0.0143	Online jobs Lag 1 $p(v)$: **0.0001** Lag 2 $p(v)$: 0.0036	USA jobs Lag 1 $p(v)$: 0.2268 Lag 2 $p(v)$: 0.2227
companies in Ghana Lag 1 $p(v)$: **0.0031** Lag 2 $p(v)$: 0.0534	Foreign Exchange Lag 1 $p(v)$: 0.1086 Lag 2 $p(v)$: 0.0025	Job interview Lag 1 $p(v)$: 0.0815 Lag 2 $p(v)$: 0.0443	Online money Lag 1 $p(v)$: **0.0322** Lag 2 $p(v)$: 0.1039	USA visa Lag 1 $p(v)$: **0.0000** Lag 2 $p(v)$: 0.0000
Cover letter Lag 1 $p(v)$: 0.1414 Lag 2 $p(v)$: 0.1310	Ghana economy Lag 1 $p(v)$: 0.7133 Lag 2 $p(v)$: 0.8431	Job opportunities Lag 1 $p(v)$: 0.2461 Lag 2 $p(v)$: 0.0003	Online schools Lag 1 $p(v)$: 0.3268 Lag 2 $p(v)$: 0.5662	Vacancy Lag 1 $p(v)$: 0.7254 Lag 2 $p(v)$: 0.0453
Curriculum vitae Lag 1 $p(v)$: **0.0135** Lag 2 $p(v)$: 0.0130	Ghana jobs Lag 1 $p(v)$: 0.1970 Lag 2 $p(v)$: 0.0301	Jobs in Ghana Lag 1 $p(v)$: 0.1561 Lag 2 $p(v)$: 0.0300	police recruitment Lag 1 $p(v)$: **0.0026** Lag 2 $p(v)$: 0.0004	Visa application Lag 1 $p(v)$: 0.3595 Lag 2 $p(v)$: 0.1187
Cv Lag 1 $p(v)$: 0.8731 Lag 2 $p(v)$: 0.0609	Graduate Lag 1 $p(v)$: 0.1661 Lag 2 $p(v)$: 0.3506	Jobs in USA Lag 1 $p(v)$: 0.4430 Lag 2 $p(v)$: 0.0679	Scholarship Lag 1 $p(v)$: 0.4433 Lag 2 $p(v)$: 0.0009	Visa Lag 1 $p(v)$: 0.1565 Lag 2 $p(v)$: 0.1066

The bold values are the estimators whose *p*-value were lesser than 0.05. This was used as a benchmark for the study

Results and discussion

Interpolation

The basic goal of temporal disaggregation methods is to create a new TS while preserving the short-term behavior of higher frequency indicator series. This TS must be coherent with low-frequency data. For the UER and interpolated UER in question, standard descriptive data are provided in Table 3. The table demonstrates that the UER and the interpolated UER are nearly equal. Visual representations of the UER and interpolated UER are shown in Fig. 2.

The graph shows that the interpolated unemployment rate, which comes from a dataset of 574 recordings, and the actual unemployment rate, which comes from a dataset of 11 records, both vary in the same way over time, proving that our dataset is equal in mean and standard deviation.

Cross-correlation function (CCF) analysis

Table 4 outlines the keywords whose trends were highly linked with UER using Granger causality Test (GCT). We compiled a list of terms from Table 4 with a high p for lag << 1 0.05 that are associated with the UER.

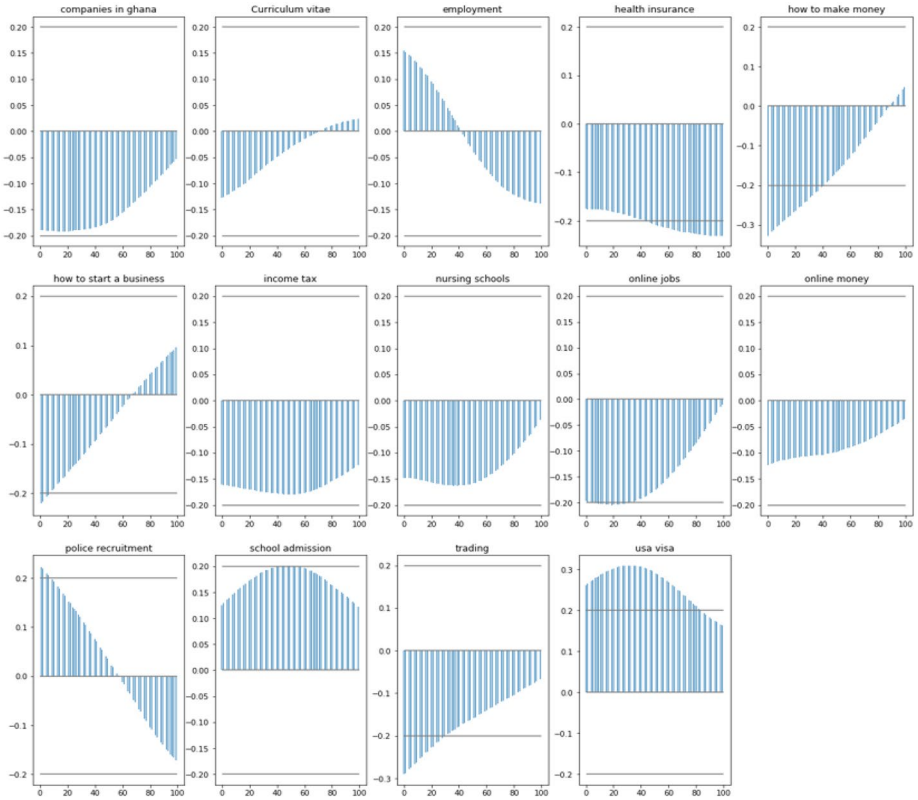

Fig. 3 GCT analysis results of selected GT (*x*) and WB UER (*y*) keywords

The table demonstrates how 14 of the 50 GT (x1 to 50) estimators for Ghana are related to the WB UERs series (*y*). The cells in the table with $p(v)$ values that are less than 0.05 for the first lag were chosen. Figure 3 displays a graph of GT estimators with p0.05 analysis results. The graph shows that there is a range of correlations between +1 and -1, where +1 represents the total positive correlation, 0 represents the absence of any correlation, and -1 represents the total negative correlation. The lags and past values of the 14 indicators are statistically significant in the equation and predicting the future values of unemployment rate.

Model result

According to the experimental design aforementioned, detailed experiments with different TS models were conducted using univariate or multivariate models. The models and order utilized in building series are ARIMA (1, 2, 1), ARIMAX (4, 1, 3) and VAX (3, 0). Table 5 illustrates the data evaluation metrics for the Models.

Evaluation of the models

The selected significant prospective determinants of the unemployment rate are taken into account with the aid of various evaluation metrics. Consideration was given to the significant y chosen for the unemployment rate in all periods. Table 5 provides an overview of the performance metrics MSE, RMSE, MAE, MAPE, MedAE, and ME for all the periods. The results show that over the first five measurement periods, the model

Table 5 Evaluation results over the TS models for the selected periods

	Model name	MSE	RMSE	MAE	MAPE	MedAE	ME
$Y_{1/12}$	ARIMA	0.00	0.00	0.00	0.19	0.00	0.00
	ARIMAX	0.00	0.00	0.00	0.18	0.00	0.00
	VAR	0.00	0.00	0.00	0.04	0.00	0.00
$Y_{3/12}$	ARIMA	0.00	0.02	0.02	0.69	0.01	0.04
	ARIMAX	0.00	0.00	0.01	0.67	0.01	0.04
	VAR	0.00	0.00	0.02	0.40	0.01	0.04
$Y_{6/12}$	ARIMA	0.01	0.08	0.06	1.76	0.05	0.18
	ARIMAX	0.01	0.01	0.06	1.70	0.04	0.17
	VAR	0.01	0.01	0.06	1.45	0.05	0.16
$Y_{9/12}$	ARIMA	0.03	0.18	0.14	3.44	0.10	0.39
	ARIMAX	0.03	0.03	0.13	3.36	0.10	0.38
	VAR	0.02	0.02	0.12	2.91	0.10	0.31
$Y_{12/12}$	ARIMA	0.10	0.31	0.24	5.78	0.18	0.67
	ARIMAX	0.09	0.09	0.23	5.67	0.18	0.66
	VAR	0.06	0.06	0.19	4.58	0.16	0.49
$Y_{15/12}$	ARIMA	0.22	0.47	0.36	8.57	0.28	0.99
	ARIMAX	0.21	0.21	0.35	8.40	0.28	0.97
	VAR	0.11	0.11	0.27	6.29	0.24	0.63
$Y_{18/12}$	ARIMA	0.41	0.64	0.49	11.51	0.40	1.28
	ARIMAX	0.39	0.39	0.48	11.23	0.39	1.23
	VAR	0.17	0.17	0.34	7.75	0.32	0.70
$Y_{21/12}$	ARIMA	0.64	0.80	0.62	14.43	0.53	1.55
	ARIMAX	0.60	0.60	0.60	13.98	0.53	1.47
	VAR	0.22	0.22	0.39	8.85	0.40	0.71
$Y_{24/12}$	ARIMA	0.92	0.96	0.76	17.27	0.69	1.79
	ARIMAX	0.92	0.92	0.76	17.27	0.69	1.79
	VAR	0.25	0.25	0.43	9.57	0.49	0.71
Average	ARIMA	0.26	0.38	0.30	7.07	0.25	0.77
	ARIMAX	0.25	0.25	0.29	6.94	0.25	0.75
	VAR	0.09	0.09	0.20	4.65	0.20	0.42

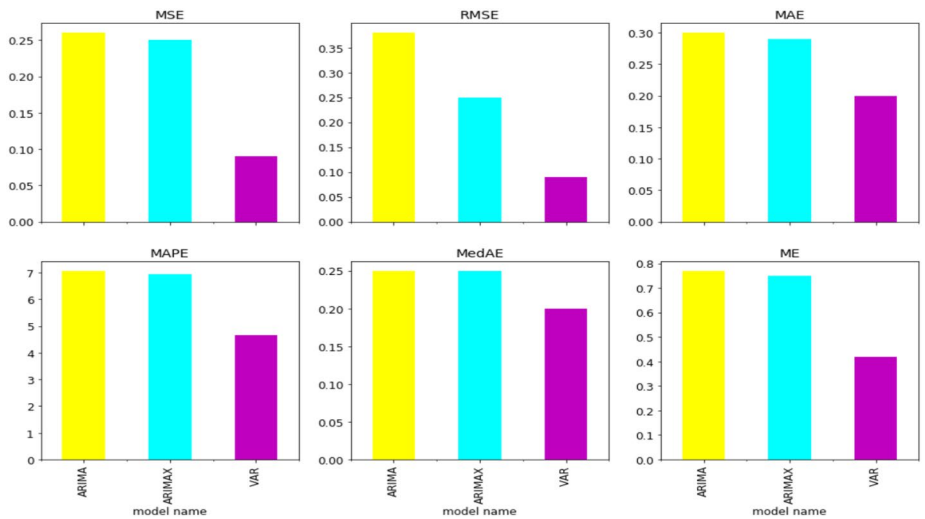

Fig. 4 Visualization comparison of the average evaluation result for the models

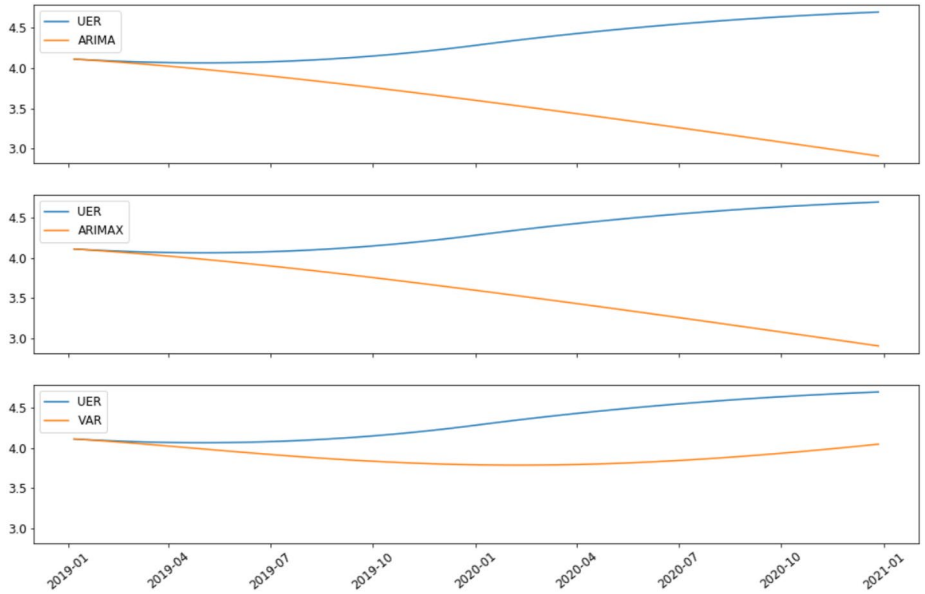

Fig. 5 Real UER and forecasted UER for Ghana for the 2 years over Models

was able to forecast with little error. Additionally, for all models, the error margin rises as the anticipated period grows. Furthermore, for nearly all periods and virtually all evaluation techniques, VAR was able to forecast with minimum error.

We created an average based on each evaluation metric results for all models, as shown in Table 5, to decide and choose the best models for the forecast. The VAR model had the best and least average error values, with MSE = 0.09, RMSE = 0.09, MAE = 0.20, MAPE = 4.65, MedAE = 0.20, and ME = 0.42, as demonstrated by the average findings in Fig. 4. This demonstrated how better the proposed model VAR (multivariate TS) with GT estimators is compared to ARIMA and ARIMX. The VAR was able to detect a minor growth even if the models did not follow the major trend of UER change. The graph demonstrated how much better and more effective the VAR model is than the other models.

Figure 5 shows the actual UER for Ghana as well as the predicted visualization for each of the models over the two-year timeframe. Except for VAR, which is somewhat in line and reflected the modest shift, all models were not in line with the UER, according to the figure. The VAR Model outperforms all other models (ARIMA and ARIMAX). Most models in economic condition approximation perform well in a stable environment, but they lack the prudence to foresee hidden economic change. In both steady and dynamic settings, the VAR Model linking input factors derived from rich high-frequency timely variables for predicting UER perform better.

Conclusion and future works

The issue is not a dearth of data, but rather a dearth of information that can be used for planning, strategy, and decision-making. Using big data, such as Google Trends, can assist the entire government system. Google Trends provides access to a huge unfiltered collection of actual Google search requests. People use Google for a wide range of

informational and topical searches, making it a valuable search engine. 50 words or phrases were of interest. Google Trends (GT) search query data were used to derive values for search relating to Jobs, society, social services, and economic indicators. The study identified a number of factors that influence the unemployment rate, including "how to make money," "how to start a business," "jobs in Ghana," "jobs in the USA," "online money," "nurse application," "visa application," and "police recruiting." This study proposes a technique to first implementing pre-processing to overcome the difficulty of handling the vast data and describes an in-depth look into the use of ARIMA, ARIMAX and VAR in nowcasting unemployment in Ghana as a use-case.

In terms of prediction accuracy, error margin, and model reliability, results show that the VAR method surpassed all other techniques. VAR (MSE = 0.09, RMSE = 0.09, MAE = 0.20, MAPE = 4.65, MedAE = 0.20, ME = 0.42) achieved significant error margins. This is compelling evidence that real-time UER forecasting at a daily level of generality is possible. Most models in economic condition approximation perform well in a stable environment, but they lack the prudence to foresee hidden economic change. In both steady and dynamic settings, the VAR Model linking input factors derived from rich high-frequency timely variables for predicting UER perform better. The objective of successful citizen care management can be attained with the use of Google Trends by offering effective data-driven services to citizens and predicting their needs based on the analysis of surveys taken among various groups of citizens. In future, more data will be collected to train with artificial intelligence techniques to generate decision support systems.

In the current study, we have highlighted a few predictor variables that contribute to the nation's unemployment rate and are crucial in figuring out unemployment. The government can also use this study's crucial information to make data-driven decisions. The government will be assisted in strengthening technical and vocational institutions. These will then bring in revenue and be put toward development. Additionally, it will be useful in establishing the state of the economy while formulating monetary policy. We recommend using machine learning model for future work.

Acknowledgements
We express our sincere gratitude to Mrs. Nancy Addia who encouraged and motivated us throughout the research. Finally, we would like to thank Google and World Bank, for making the data available.

Author contributions
"Conceptualization, WKA and PA; methodology, SA and WKA; software, WKA.; validation, WKA, PA and SA; formal analysis, WKA; investigation, WKA, PA and SA; resources, PA; data curation, WKA; writing—original draft preparation, WKA; writing—review and editing, PA; visualization, WKA; supervision, PA; project administration, PA; funding acquisition, PA. All authors have read and agreed to the published version of the manuscript."

Funding
Not applicable.

Availability data and materials
The data presented in this study are publicly available through the Fig Share repository via Afrifa, Stephen (2022): unemployment_data.csv. figshare. Dataset. https://doi.org/10.6084/m9.figshare.20311167.v1.

Declarations

Competing interests
Competing interest statement declared by the corresponding author on behalf of all authors. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 26 August 2022 Accepted: 13 February 2023
Published online: 27 February 2023

References

1. Mulero R, García-Hiernaux A (2021) Forecasting Spanish unemployment with Google Trends and dimension reduction techniques. SERIEs 12(3):329–349. https://doi.org/10.1007/s13209-021-00231-x
2. Rizky O, Fajar M, Prasetyo OR, Nonalisa S (2020) Forecasting unemployment rate in the time of COVID-19 pandemic using Google Trends Data (Case of Indonesia). Munich Pers. RePEc Arch, no. 105042
3. Nirmala CR, Roopa GM, Kumar KRN (2015) Twitter data analysis for unemployment crisis. In: Proceedings of 2015 international conference applications theoretical computer communications and technology. iCATccT 2015, pp 420–423. https://doi.org/10.1109/ICATCCT.2015.7456920
4. Ryu PM (2018) Predicting the unemployment rate using social media analysis. J Inf Process Syst 14(4):904–915. https://doi.org/10.3745/JIPS.04.0079
5. Mavragani A, Ochoa G, Tsagarakis KP (2018) Assessing the methods, tools, and statistical approaches in Google trends research: Systematic review. J Med Internet Res 20(11):1–20. https://doi.org/10.2196/jmir.9366
6. Twumasi E, Frimpong EA, Kwegyir D, Folitse D (2021) Improvement of grey system model using particle swarm optimization. J Electr Syst Inf Technol. https://doi.org/10.1186/s43067-021-00036-9
7. Naccarato A, Falorsi S, Loriga S, Pierini A (2018) Combining official and Google Trends data to forecast the Italian youth unemployment rate. Technol Forecast Soc Change 130:114–122
8. McCallum ML, Bury GW (2014) Public interest in the environment is falling: a response to Ficetola (2013). Biodivers Conserv 23(4):1057–1062
9. Jun SP, Park DH (2016) Consumer information search behavior and purchasing decisions: empirical evidence from Korea. Technol Forecast Soc Change 107:97–111. https://doi.org/10.1016/j.techfore.2016.03.021
10. Han SC, Chung H, Kang BH (2012) It is time to prepare for the future: forecasting social trends. In: Kim Th, Ma J, Fang Wc, Zhang Y, Cuzzocrea A (eds) Computer applications for database, education, and ubiquitous computing. EL DTA 2012. Communicat. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35603-2_48.
11. Vosen S, Schmidt T (2011) Forecasting private consumption: Survey-based indicators vs. Google trends. J Forecast 30(6):565–578. https://doi.org/10.1002/for.1213
12. Kundu S, Singhania R (2020) Forecasting the United States unemployment rate by using recurrent neural networks with Google Trends data. 11(6). https://doi.org/10.18178/ijtef.2020.11.6.679
13. Heidary J, Rastegar H (2022) A novel computational technique using coefficient diagram method for load frequency control in an interconnected power system. J Electr Syst Inf Technol 9(1):1–24. https://doi.org/10.1186/s43067-022-00062-1
14. Simionescu M, Zimmermann KF (2017) "Big Data and Unemployment Analysis," GLO Discuss. Pap., p. No. 81
15. Hacıevliyagil N, Drachal K, Eksi IH (2022) Predicting house prices using DMA method: evidence from Turkey. Economies 10(3):1–27. https://doi.org/10.3390/economies10030064
16. Naccarato A, Pierini A, Falorsi S (2015) Using Google Trend data to predict the Italian unemployment rate. Dep. Work. Pap. Econ. - Univ. "Roma Tre
17. Junior MA, Appiahene P, Appiah O (2022) Forex market forecasting with two - layer stacked Long Short - Term Memory neural network (LSTM) and correlation analysis. J Electr Syst Inf Technol 1:1–24. https://doi.org/10.1186/s43067-022-00054-1
18. Simionescu M, Cifuentes-Faura J (2022) Forecasting National and Regional Youth Unemployment in Spain Using Google Trends. Soc Indic Res 164(3):1187–1216. https://doi.org/10.1007/s11205-022-02984-9
19. Simionescu M, Cifuentes-Faura J (2022) Can unemployment forecasts based on Google Trends help government design better policies? An investigation based on Spain and Portugal. J Policy Model 44(1):1–21. https://doi.org/10.1016/j.jpolmod.2021.09.011
20. Şentürk G (2022) Can Google search data improve the unemployment rate forecasting model? AN empirical analysis for Turkey. J Econ Policy Res 9(2):229–244. https://doi.org/10.26650/jepr963438
21. Ettredge M, Gerdes J, Karuga G (2005) Using web-based search data to predict macroeconomic statistics. Commun ACM 48(11):87–92. https://doi.org/10.1145/1096000.1096010
22. Choi H, Varian H (2009) Predicting the present with Google Trends. Tech. report, Google. [Cited 1 April 2012.]
23. Choi H, Varian H (2009) Predicting initial claims for unemployment insurance using Google Trends. Tech. report, Google. [Cited 1 April 2012.]
24. Petropoulos A, Siakoulis V, Stavroulakis E, Lazaris P, Vlachogiannakis N (2021) Employing Google Trends and deep learning in forecasting financial market turbulence. J Behav Financ. https://doi.org/10.1080/15427560.2021.1913160
25. Tuhkuri J (2016) ETLAnow: a model for forecasting with Big Data forecasting unemployment with Google Searches. ETLA Reports 54, no. 54, p 20
26. Tuhkuri J (2016) Forecasting unemployment with Google Searches. ETLA Work. Pap. No 35
27. Lasso F, Snijders S (2016) The power of Google search data2 an alternative approach to the measurement of unemployment in Brazil
28. te Brake G, Ramos R (2017) Unemployment ? Google it ! Analyzing the usability of Google queries in order to predict unemployment
29. Maas B (2019) Short-term forecasting of the US unemployment rate. J Forecast. https://doi.org/10.1002/for.2630
30. Jung JU, Hwang J (2019) Application of Google Search queries for predicting the unemployment rate for Koreans in their 30s and 40s. 17(9):135–145
31. A. O. O. Smit (2018) Unemployment rate forecasting using Google trends, Bachelor Thesis in Econometrics & Operations Research erasmus university rotterdam erasmus school of economics, pp 1–22

32. Jimenez A, Santed-Germán MA, Ramos V (2020) Google Searches and Suicide Rates in Spain, 2004–2013: Correlation Study. JMIR Public Heal Surveill 6(2):2004–2013. https://doi.org/10.2196/10919
33. Mosley L, Eckley I, Gibberd A (2021) Sparse temporal disaggregation, no. 2019, pp 1–33
34. Ghouali S et al (2017) The granger causality effect between cardiorespiratory hemodynamic signals to cite this version : HAL Id : hal-01573108 The Granger Causality Effect between. https://doi.org/10.5176/2251-1911
35. Chen B, Ma R, Yu S, Du S, Qin J (2019) Granger causality analysis based on quantized minimum error entropy criterion. IEEE Signal Process Lett 26(2):347–351. https://doi.org/10.1109/LSP.2019.2890973
36. Bressler SL, Seth AK (2011) Wiener–Granger causality: a well established methodology. Neuroimage 58(2):323–329. https://doi.org/10.1016/j.neuroimage.2010.02.059
37. Bai P, Safikhani A, Michailidis G (2022) Multiple change point detection in reduced rank high dimensional vector autoregressive models. J Am Stat Assoc. https://doi.org/10.1080/01621459.2022.2079514
38. Odekina GO, Adedotun AF, Imaga OF (2022) Modeling and forecasting the third wave of Covid-19 incidence rate in Nigeria using vector autoregressive model approach. J Niger Soc Phys Sci 4(1):117–122. https://doi.org/10.46481/jnsps.2022.431
39. Cho H, Maeng H, Eckley IA, Fearnhead P (2022) High-dimensional time series segmentation via factor-adjusted vector autoregressive modelling, pp 1–62
40. Victor-Edema UA, Essi PID (2016) Autoregressive integrated moving average with exogenous variable (ARIMAX) model for Nigerian Non Oil Export 8(2014):2010–2015
41. Yucesan M, Gul M, Celik E (2018) Performance comparison between ARIMAX , ANN and ARIMAX-ANN hybridization in sales forecasting for furniture industry. RES Gate. https://doi.org/10.5552/drind.2018.1770

Publisher's note