The Journal of Finance and Data Science 9 (2023) 100106

Contents lists available at ScienceDirect

The Journal of Finance and Data Science

journal homepage: www.keaipublishing.com/en/journals/jfds

Research Article

Data-driven estimation of economic indicators with search big data in discontinuous situation

Goshi Aoki [a,c,1,4], Kazuto Ataka [a,c,*,2,4], Takero Doi [b,3], Kota Tsubouchi [c,4]

[a] 5322 Endo, Fujisawa-shi, Kanagawa 252-0882, Japan
[b] 2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan
[c] Kioi Tower 1-3 Kioicho, Chiyoda-ku, Tokyo 102-8282, Japan

ARTICLE INFO

Keywords:
Nowcasting
Business cycle indicators
Search big data
Japanese economy
Discontinuity
Humanneeds

ABSTRACT

Economic indicators are essential for policymaking and strategic decisions in both the public and private sectors. However, due to delays in the release of government indicators based on macroeconomic factors, there is a high demand for timely estimates or "nowcasting". Many attempts have been made to overcome this challenge using macro indicators and key variables such as keywords from social networks and search queries, but with a reliance on human selection. We present a fully data-driven methodology using non-prescribed search engine query data (Search Big Data) to approximate economic variables in real time. We evaluate this model by estimating representative Japanese economic indicators and confirm its success in nowcasting prior to official announcements, even during the COVID-19 pandemic, unlike human-selected variable models that struggled. Our model shows consistent performance in nowcasting indices both before and under the pandemic before government announcements, adapting to unexpected circumstances and rapid economic fluctuations. An exhaustive analysis of key queries reveals the pivotal role of libidinal drives and the pursuit of entertainment in influencing economic indicators within the temporal and geographic context examined. This research exemplifies a novel approach to economic forecasting that utilizes contemporary data sources and transcends the limitations of existing methodologies.

1. Introduction

Economic indicators announced by the government play critical roles in decision-making about upcoming actions in both the public and private sectors. A major drawback of these indicators is their lack of timeliness, due to the fact that they are based on macroeconomic factors such as inventory turnover and iron production. In the case of the Japan Cabinet Office's Indexes of Business Conditions, indices are announced two months later, e.g., a September value will be announced in November of the same year.

* Corresponding author.
E-mail addresses: goshiaoki@keio.jp (G. Aoki), ataka@sfc.keio.ac.jp (K. Ataka), tdoi@econ.keio.ac.jp (T. Doi), ktsubouc@yahoo-corp.jp (K. Tsubouchi).

Peer review under responsibility of KeAi Communications Co., Ltd.

[1] Faculty of Policy Management, Keio University, Kanagawa, Japan.
[2] Faculty of Environment and Information Studies, Keio University, Kanagawa, Japan.
[3] Faculty of Economics, Keio University, Tokyo, Japan.
[4] Yahoo Japan Corporation, Tokyo, Japan.

https://doi.org/10.1016/j.jfds.2023.100106
Received 29 December 2022; Received in revised form 16 August 2023; Accepted 7 September 2023
Available online 12 September 2023
2405-9188/© 2023 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Economists have been researching nowcasting of macroeconomic variables for years via various ways of modeling national economic indicators (Bragoli and Fosten, 2018; Richardson et al., 2021). Comprehensive reviews of previous studies on nowcasting in the literature of economics are provided in BańBura et al. (2011, 2013). Recent studies, particularly focusing on the Japanese case, have also contributed to this field. For example, Bragoli (2017), Chikamatsu et al. (2018) and Hayashi and Tachi (2022) have provided valuable insights into nowcasting in the context of Japan. Most of them show that macroeconomic indicators can be modeled on other macroeconomic variables and semi-macroeconomic statistical data, which are aggregated data on the number of actions taken by many households and firms related to the macroeconomic components, such as consumption and investment, if a time lag is allowed due to the data availability.

Evans (2005) and Proietti and Giovannelli (2021) attempt to nowcast macroeconomic indicators using state space methods. (Carriero et al. 2015, 2020) and (D'Agostino et al. 2016) approach nowcasting them using Bayesian methods. Chauvet and Potter (2013) and Siliverstovs (2020), however, assess nowcast accuracy of those sophisticated models, and conclude that they produce average forecast accuracy that at best is comparable to that from an autoregressive model of order 2 (AR(2)) of macroeconomic indicators, which is the benchmark model in macroeconomics.

In addition, macroeconomic data can hardly represent irregular economic discontinuity, for instance, caused by the COVID-19 pandemic, because discontinuities in the economy are not immediately reflected in the macroeconomic factor. Most existing researches are conducted under situation without apparent discontinuity. Comelli (2015) examines the spillover effects of economic fluctuations in developed countries to developing countries during the period of the Great Recession. However, it is still not possible to address simultaneous global economic discontinuities such as COVID-19.

In order to overcome the drawbacks of existing macro-variable-driven methods, we here report the development of a new bigdata-driven approach that can reasonably nowcast the values of macroeconomic indicators without requiring aggregation of semi-macroeconomic data. Our approach requires only using query logs, “Search Big Data,” from a major search engine in the country where the economic indicators are published.

Our approach based on Search Big Data has two major advantages over previous macro-index-based approaches for developing a model of economic indicators. First, the data are immediately available: by definition, search engine query data can be obtained and analyzed in real time. Second, the data contain the signals of highly diversified human interests and activities in daily life, ranging from news topics to factory utilization and artificial intelligence (AI) (Choi and Varian, 2012; Della et al. 2009; Radinsky et al., 2008).

By leveraging this signal diversity, Askitas and Zimmermann (2009), D'Amuri and Marcucci (2010), Ettredge et al. (2005), and McLaren and Shanbhogue (2011) study unemployment rates, while Ataka (2016) forecast the results of a Japanese national election. Similarly, Hisada et al. (2020) examine the occurrences of COVID-19 patients, and Sasaki et al. (2021) calculate a real-time mood score for society. All of those works take advantage of the unique characteristics of Search Big Data.

There are some papers using methods that employ variables other than economic indicators. Big data from Search captures the needs of diverse consumers in far greater detail than macroeconomic indicators that aggregate a variety of economic activities. Aguilar et al. (2021) develop a daily sentiment indicator based on textual newspaper data. Cavallo and Rigobon (2016) create consumer price indexes by collecting a massive amount of retail price information. Chen et al. (2015) investigate signals of the business cycle using real-time Google search volume data. Chen et al. (2015), however, predict macroeconomic trends based on some queries pre-specified by the authors. Also, Bouayad et al. (2022), Jaemin and Owen (2019), Kohns and Bhattacharjee, 2023, Vosen and Schmidt (2011a, 2012), and Woloszko (2020) choose queries based on human knowledge. The selection rules of these papers can be biased due to subjective criteria, or the targeted candidate set can be quite limited, missing those seemingly irrelevant queries that are quite sensitive to economic conditions. Our machine-based methods are the opposite of the previous human knowledge-based methods.

In this study, we developed a fully data-driven approach to predict government economic indicators using only non-preset search engine query data. Our method is a statistical technique that utilizes the most highly correlated queries to model economic indicators through Search Big Data. During our model optimization process, we assessed four machine-based methods. We opted for a multiple regression analysis, selecting the most relevant search queries based on the strength of their correlations.

We used search big data from Yahoo! JAPAN, which approximately 80% of the Japanese population uses the internet (Ministry of Internal Affairs and Communications, 2021), and more than 60% of the nation's internet users use Yahoo! JAPAN Search, from which we obtained the query data. Moreover, we demonstrate the feasibility of constructing statistical models based on select search queries that exhibit strong correlations with certain economic indicators: specifically, Japan's Coincident Indexes of Business Conditions and Consumer Confidence Index (CCI), among billions of queries. We further show that statistical models based on queries preceding the target month can be used to nowcast economic indicators even during the unprecedented recession caused by the COVID-19 pandemic. Furthermore, our analysis of the queries used in the models reveals that human libido and desire for laughter are key drivers of the economy in Japan during the time period studied. In this way, we have developed a novel approach to modeling economic indicators while investigating the social forces underlying these trends by utilizing Search Big Data.

2. Data and methodology

Figure 1 illustrates the proposed methodology for estimating economic indicators from Search Big Data. First, we extract all query data (over billions) from the server (for details, see subsection “Data”). Next, we extract the continuously searched queries (about 300,000) from all data (see subsection “Data”). To create a well-performed model, we calculate the single correlation between each extracted query and the target economic indicators. The queries with the highest absolute value between single correlations are extracted (see subsection “Query Selection for Training Step 1”) and remove multicollinearity (see subsection “Query Selection for Training Step 2”). Finally, after performing a t-test, we create the model (see subsection “Query Selection for Training Step 3”).

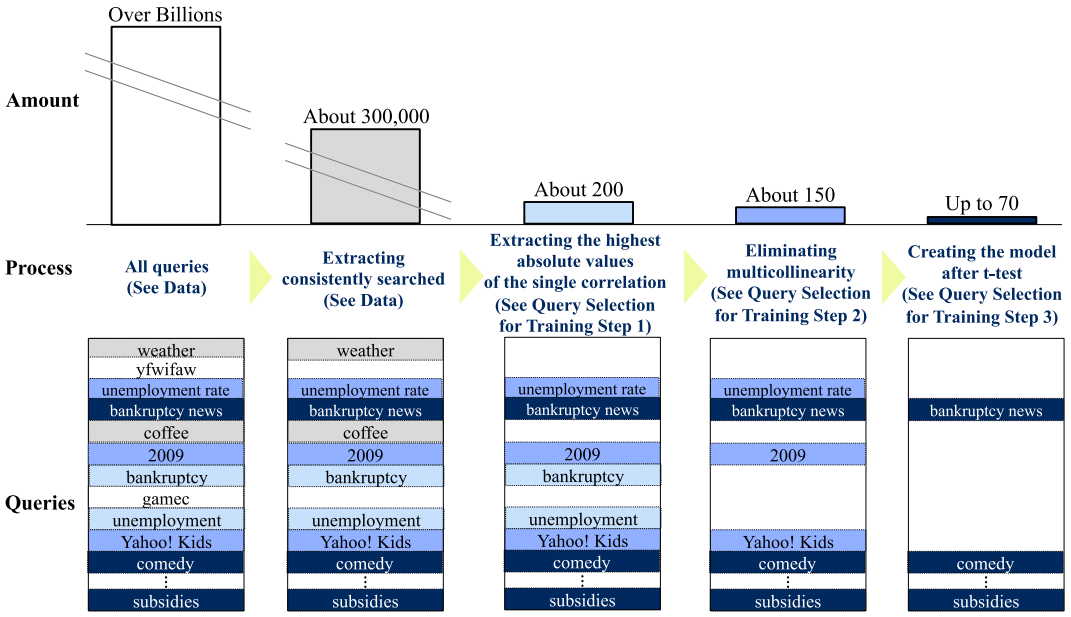

Fig. 1. Overview of our methodology for estimating economic indicators from Search Big Data.

2.1. Data

We use query and timestamp data that are collected between January 2006 and March 2021, which is obtained from Yahoo! JAPAN Search. Yahoo! JAPAN (an internet portal), with more than 50 million monthly active users, runs- one of the most popular search engine platforms in Japan. In order to use constantly searched queries, we extract with a threshold of 10 times per day from April 2020 to March 2021.

We employ the Coincident Index (CI) from the Japan Cabinet Office's Indexes of Business Conditions (Cabinet Office, 2021) and the CCI, as they are representative indicators of real-time economic trends and publicly available in Japan (e-Stat). CI is an indicator compiled for the purpose of understanding the current state of the economy and contributing to future forecasts, and is calculated by integrating the movements of important and economy-sensitive indicators of various economic activities, such as production and employment. CCI is calculated using a questionnaire survey to obtain basic data for judging economic trends by ascertaining consumers' attitudes toward the outlook for their daily lives and their ownership of major durable consumer goods. In general nowcasting research, GDP is used as the target variables (Bragoli, 2017; Chikamatsu et al., 2018; Hayashi and Tachi, 2022), but in this study, we use CI and CCI instead. This is because GDP is announced only once every three months, and it is not adequate as an indicator for policy decisions in rapidly changing situations like COVID-19. Furthermore, previous studies using Search Query Data usually employ real monthly consumption indices (Jaemin and Owen, 2019; Vosen and Schmidt, 2011b; Vosen and Schmidt, 2012; Woloszko, 2020). Therefore, we choose the indicators CI and CCI, which are announced monthly.

2.2. Model

To model the economic indicators with Search Big Data, we perform a multiple regression analysis. In contrast to methods such as Principal Component Analysis (PCA), which emphasizes the extraction of orthogonal components that explain the largest variance in the data, and another statistical method like Partial Least Squares (PLS) that focuses on maximizing the covariance between the independent and dependent variables, our approach is a basic statistical method that leverages the most highly correlated queries to model economic indicators with Search Big Data. We perform multiple regression analyses and select the most relevant search queries based on their correlation strength. Table 1 and Table 2 list the top queries that have the strongest correlation with the economic indicators, along with the value of the Pearson correlation coefficients between each search query and the CI (Table 1 lists positive strong correlations and Table 2 lists negative ones).

Table 1
Top five queries with the highest positive Pearson correlation coefficients between the Web search query data and CI from June 2005 to February 2021. (* denotes names of the city in Japan.)

Search query	Query meaning	Correlation coefficient
Sapporo fuzoku	Sapporo* adult entertainment	0.66
Osaka fesutibaru hooru	concert hall	0.65
Sunlady's	employment agency name	0.65
Yokohama fuzoku	Yokohama* adult entertainment	0.63
Shibuya fuzoku	Shibuya* adult entertainment	0.63

Table 2
Top five queries with the highest negative Pearson correlation coefficients between the Web search query data and CI from June 2005 to February 2021.

Search query	Query meaning	Correlation coefficient
Ishida Akira	name of famous comedian	−0.79
8591	a company code in the securities market	−0.78
tosanjoho	bankruptcy information	−0.77
koyojoseikin	government subsidies for employment	−0.77
Audrey	a comedian	−0.76

Figure 2 illustrates the data decomposition method. The nowcasting target months span a total of 24 months, ranging from April 2019 to March 2021. A distinct model is constructed for each target month, resulting in a total of 24 models for nowcasting purposes. The validation data comprises the 12-month period immediately preceding the target month, while the training data encompasses the period from January 2006 up to the month prior to the validation period. For instance, when considering April 2019 as the target month, the validation period extends from March 2018 to March 2019, and the training period spans from January 2006 to February 2018.

2.3. Query selection for training

Four steps are performed to select the queries to be used to create the model.

2.3.1. Step 1: calculation of single correlation

First, data are extracted from January 2006 to the month prior to the target month for estimation, using approximately 300,000 queries that are consistently searched. Then, Pearson correlation coefficients between the economic indicators and search query data from this period are calculated to figure out the candidate queries for inclusion in the models.

2.3.2. Step 2: multicollinearity

To prevent model instability, we eliminate multicollinearity. First, the queries obtained in Step 1 are ordered by the absolute values of their correlation coefficients. Next, we extract a number of queries equal to the number of months in the training period. We have chosen this threshold for the number of queries to be included in the regression model because of the linear nature of the regression; theoretically, including more queries than the training period would lead to overfitting. The queries with the highest absolute values of the correlation coefficients are added to a list for inclusion in a model. The variance inflation factor (VIF) is calculated for the queries with the highest and second highest absolute values in the list. If a query's VIF is less than 10, then it is added to the model; otherwise, the query is removed. Next, the VIF is calculated for the query with the third largest absolute value, and it is added to the list for inclusion in the model. When the query list contains two queries, the VIF is calculated for both queries. This process is performed for all of the extracted queries.

2.3.3. Step 3: t-test

From the list of queries obtained in Step 2, the top w queries with the strongest correlations are chosen. With these queries, a multiple regression model are calculated to estimate their respective p-values. Subsequently, we iteratively remove the query with the highest p-value above the 0.05 threshold until all remaining queries exhibit p-values below this threshold. This procedure guarantees that the null hypothesis is rejected at the 5% significance level. The value of w varies from 1 to the number of queries remaining after eliminating multicollinearity in Step 2.

2.3.4. Step 4: validation

In this step, we focus on selecting the top w queries with the smallest mean squared errors (MSEs) for the final estimation. This is accomplished through a validation process that leverages the data from the 12 months immediately preceding the target month. For

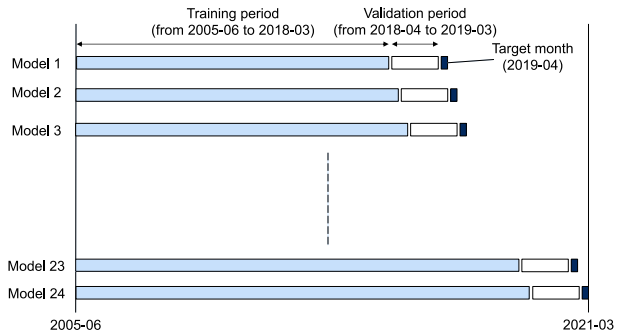

Fig. 2. Overview of our methodology for the data decomposition method.

each target month within the estimation period of April 2019 to March 2021, we calculate the MSE for each of the top w queries with the strongest correlations, using data from 13 months prior to the month before the target month. From these results, we identify the top h queries with the lowest MSEs. The model constructed from these h queries is then adopted as the final estimation model. This validation process ensures that the most accurate and reliable model is selected for each target month, enhancing the overall robustness and reliability of the nowcasting results.

3. Empirical analysis and results

3.1. Overview of empirical analysis

To evaluate this multiple regression model based on search queries (multi-query model: *Our Model*), nine models are used as controls (comparative methods) to evaluate the nowcasting performance of economic indicators under "Before COVID-19" and "Under COVID-19" situations.

The "Before COVID-19" period is defined as the period before April 7, 2020, when the Japanese government declared a state of emergency (SoE) for the first time and urged its citizens to refrain from going out, and the "Under COVID-19" period is defined as the period from April, 2020, to March 2021. In each case, the estimates for the month are extrapolated.

One of the models used for comparison is a single regression model, also using search queries (Single-Query), and another is a multiple regression model using macroeconomic indicators(Macro). In addition to these models, we also compare our model with seven other methods that utilize Google Trends data in existing research. These methods rely on pre-specified queries for regression based on economic knowledge gained from past studies. The seven methods are from the following studies (Bouayad et al., 2022; Chen et al., 2015; Jaemin and Owen, 2019; Kohns and Bhattacharjee, 2023; Vosen and Schmidt, 2011a, 2012; Woloszko, 2020). Since the queries used in these existing studies are not in Japanese, we use Google Translate to translate the words into Japanese and employ them as the target queries for our analysis.

3.2. Comparative methods

- **Macro** is a multiple regression model using macroeconomic indicators following the work of Cepni et al. (2019). The model based on economic indicators is developed using the macro indicators shown in Table 3. In order to avoid getting values for the CI and CCI themselves, we do not use the economic indicators that are actually used to create these indicators, but instead use the economic indicators released monthly by the Bank of Japan and the government.
- **Single-Query** is a single regression model. The query with the highest correlation with the economic index in the data training period, from 2006, up to the month immediately preceding the nowcast month is used to produce a nowcast of the economic index for the following month.
- **Vosen (2011)** is a multiple regression model using Search Query Data following the work of Vosen and Schmidt (2011a). It employs 56 consumption-relevant categories from Google Trends.
- **Vosen (2012)** is a multiple regression model using Search Query Data following the work of Vosen and Schmidt (2012). It employs 41 consumption-relevant categories from Google Trends.
- **Chen (2015)** is a multiple regression model using Search Query Data following the work of Chen et al. (2015). It employs three queries such as "Recession," "Foreclosure Help" and "Layoff."
- **Jaemin (2019)** is a multiple regression model using Search Query Data following the work of Jaemin et al. (Jaemin and Owen, 2019). It employs 64 categories that are intuitively related to the Bureau of Economic Analysis' categorization, which are from Google Trends.
- **Woloszko (2020)** is a multiple regression model using Search Query Data following the work of Woloszko (2020). It employs 115 categories from Google Trends.

Table 3
Economic variables used for nowcasting.

Economic Variables	Sources
Effective Exchange Rates(Nominal)	Bank of Japan
Effective Exchange Rates(Real)	Bank of Japan
Nominal Consumption Activity Index	Bank of Japan
Real Consumption Activity Index	Bank of Japan
Producer Price Index	Corporate Goods Price Index
Value of Wholesale	Current Survey of Commerce
Exchange Rates	Finance and Economic Statistics
Labor Productivity Index	Labour Productivity Statistics
Balance of Payments	Ministry of Finance
Hours Worked Indices	Monthly Labor Survey
Unemployment Insurance	Statistics of Employment Insurance
Indices of Operating Ratio Manufacturing	Trade and Industrial Statistics

- **Kohns (2022)** is a multiple regression model using Search Query Data following the work of Kohns et al. (Kohns and Bhattacharjee, 2023). It employs 37 categories which capture economic activity ranging from recession, labor market, personal finance, consumption to supply side activities, which are from Google Trends.
- **Bouyard (2022)** is a multiple regression model using Search Query Data following the work of Bouyard (Bouayad et al., 2022). It employs 12 categories which depict the spending behavior of the Moroccan consume, which are from Google Trends.

3.2.1. Metrics

We use the mean squared error (MSE) to evaluate each model (comparative method). The MSE represents the difference between the original and estimated values extract by squaring the average difference across the data. The MSE is calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i represents the original value, $\hat{y}_i$ represents the estimated value and n represents the number of the dataset.

3.3. Results

3.3.1. Overall performance evaluation

Table 4 summarizes the MSE results in each period. As seen in the table, the model using only economic statistics (denoted as *Macro*) performed well during the Before COVID-19 period but showed a significant decrease in performance during the Under COVID-19 period. Single regression models (denoted as *Single-Query*) show low accuracy for both periods. Also, human knowledge-based models exhibit satisfactory results in the Before COVID-19 period, but demonstrate poorer performance in the Under COVID-19 period. In contrast, the proposed multi-query-based method (denoted as *Our Model*) showed high nowcasting performance, especially during the Under COVID-19 period.

Figure 3 shows the actual and nowcasted CI values from April 2019 to March 2021. Similarly, Fig. 4 shows these values for the CCI. The vertical axis shows the values of CI and CCI. The horizontal one shows the time period of estimation. *Actual* is the official value announced by the government, *Our Model* is the proposed method. The others depict the value estimated by each methodology (*Macro*, *Single Query* and human knowledge-based models).

The nowcasted values obtained with multi-query-based model (*Our Model*) show that the CI changes were captured with no time lag during the rapid economic downturn in the period prior to April 2020, when a COVID-19 state of emergency (SoE) was first declared in Japan. In May 2020, the SoE was lifted, and the subsequent changes reflecting the economic recovery were also captured by our method.

On the other hand, the *Macro* model using economic statistics showed high accuracy Before COVID-19 period, but it could not detect the sudden CCI decline due to the COVID-19 pandemic. After Japan's second SoE on January 8, 2021, the actual CI value did not fall, but the actual CCI did. The Macro model neither estimated value fell in both CI and CCI, while the *Our Model* estimated fell in both indices. Our model also estimated an extreme recovery, which was not originally the case with the CCI indicator.

In addition, Before the COVID-19 pandemic, the human knowledge-based models demonstrated relatively accurate predictions for both the CI and CCI. However, Under COVID-19, these models showed poor performances. For the CI, the model struggled to adapt to the rapid changes, resulting in less accurate predictions. In contrast, while the model's predictions for the CCI exhibited some responsiveness to the changes, most of the predictions still deviated from the actual values.

3.4. Queries extracted for models

Tables 5 and 6 show the importance of each query used in trained models. We ran a simulation predicting economic indicator value for 24 months. The column Query utilization (%) in the table shows how many of the 24 months of models generated contain the same

Table 4
MSE in estimating CI and CCI values Before COVID-19 and Under COVID-19 period with Deterioration rate. Note: The bold represents the best performance.

	CI			CCI		
Model	Before COVID-19	Under COVID-19	Under/ Before	Before COVID-19	Under COVID-19	Under/ Before
Our Model	31.8	**34.4**	1.1	21.9	**22.5**	1.0
Macro	**9.0**	98.7	11.0	**7.0**	34.7	5.0
Single-Query	497.5	314.6	0.6	15.7	52.5	3.3
Vosen (2011)	25.1	104.6	4.2	13.8	53.4	3.9
Vosen (2012)	20.2	166.6	8.3	17.1	71.0	4.2
Chen (2015)	11.4	174.2	15.3	14.9	101.9	6.8
Jaemin (2019)	24.8	96.6	3.9	19.1	48.6	2.5
Woloszko (2020)	14.7	92.8	6.3	18.5	63.0	3.4
Kohns (2022)	14.4	107.7	7.5	19.6	133.2	6.8
Bouayad et al. (2022)	16.6	121.1	7.3	21.9	45.1	2.1

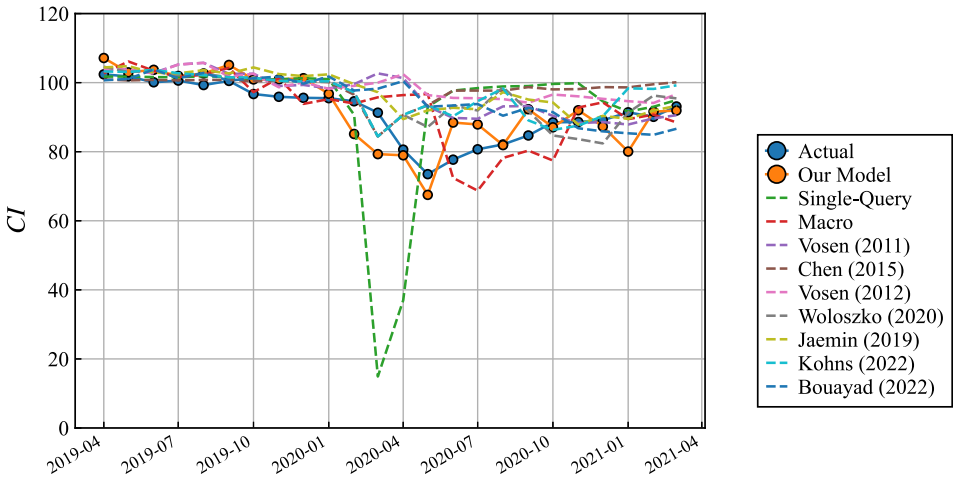

Fig. 3. Actual and nowcasted values of the CI from April 2019 to March 2021.

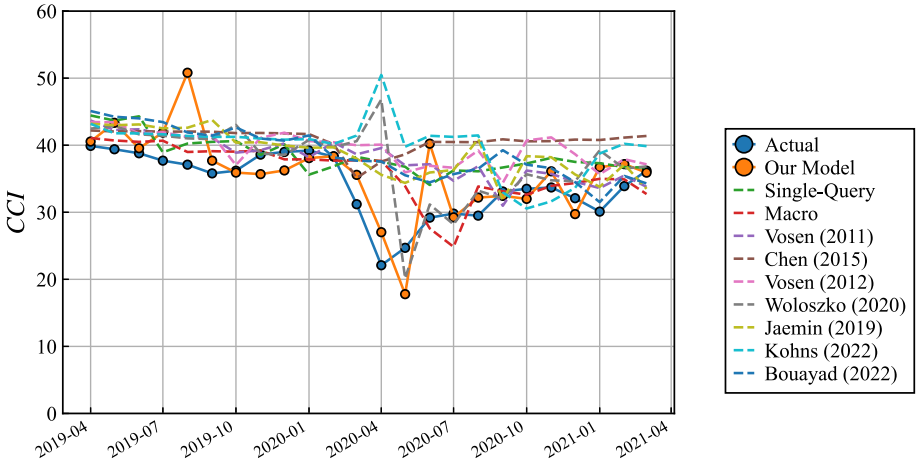

Fig. 4. Actual and nowcasted values of the CCI from April 2019 to March 2021.

search query. For example, the query in Table 6 indicating the famous comedian named *Kinoshita* means that it was selected as an important query for modeling CCI in 22 months of the simulation. Note that the queries listed in the tables are not actual economic indicators but queries which are selected in a fully data-driven approach. For instance, the query "koyojoseikin" was chosen. It does not indicate the subsidies for the employment index itself, but shows that users searched for the query "koyojoseikin" and the proposed model determined that the signal was related to the indices.

Analysis of the queries that were used to create the models revealed that different kinds of queries were important: not only those related to economic activities, such as "subsidies for employment," "credit bureau," and "shopping," but also those related to fundamental human activities, such as "comedy" and "adult entertainment."

4. Discussion

4.1. Comparison of model performance

As can be seen in Table 4, the *Macro* model was slightly more accurate than *Our Model* Before COVID-19. In contrast, *Our Model* outperformed during the Under COVID-19 period. The difference in MSE between *Our Model* and *Macro* Before COVID-19 is small compared to Under COVID-19 period.

This result can be interpreted due to the following background: in the case of the Macro model, it is originally necessary to use statistical indicators published two months later, but in practice, as shown in Figs. 3 and 4, it is necessary to use statistics showing values two months ago that are currently published if one wants to nowcast. The Single-Query model and models based on human knowledge tend to overestimate because it fails to capture social diversity.

Table 5
Top 20 queries used for CI nowcasting from April 2019 to March 2021 among 24 models.

Search query	Query meaning	Category	Query utilization(%)
syokuhinsokuho	a food news site	Others	75.0
Bikkukamera	an electronics store	Others	66.7
Tocco	a shopping site	Others	62.5
hanikamu	smiling	Others	62.5
Yuuji Kouji	name of famous comedy group	Comedy	62.5
koyojoseikin	subsidies for employment	Economic Activities	62.5
Roll chan	a food name	Others	58.3
tegaki	handwriting	Others	58.3
mondomuyo	an adult video site	Libido	54.2
nogyohojin	agricultural company	Economic Activities	54.2
Ishida Akira	name of famous comedian	Comedy	54.2
Audrey	name of famous comedy group	Comedy	54.2
Yafuu Kizzu	a website for children	Others	50
nifty.com	a web service provider	Others	50
comic	-	Others	45.8
2009nen	2009	Others	45.8
Sangyo koyo antei sentar	Employment Stabilization Center	Economic Activities	45.8
Sushida	a typing game	Others	45.8
dejitarufotofuremu	digital photo frame	Others	45.8
Minorin	a Youtuber	Others	45.8

However, looking at the Under/Before in Table 4, Our Model, based on multiple queries, has a lower deterioration rate when comparing Before COVID-19 and Under COVID-19, compared to the other methods. It means that our model can nowcast economic indices with the same performance between Before COVID-19 and Under COVID-19.

Other methods manually select direct queries related to the economy, but this idea is only acceptable when the economy continues to be an extension of the existing situation. Thus, it does not work in discontinuous situations such as COVID-19.

Our method, on the other hand, is data-driven and selects a model of the economy from the elementary level that constitutes that of the query. Regardless of the situation, the fact that the results of people's lives and behavior are reflected in the economic indices is constant, and as a result, the proposed method is able to nowcast with high accuracy in discontinuous situations of COVID-19 without affecting its performance. As long as human behavior continues to reflect economic conditions, this method will remain effective indefinitely.

Moreover, *Our Model* can capture the diversity of needs while still being able to quickly capture the changes in society with a high degree of rapidity. Although the queries used in the creation of this model all exclude multicollinearity, it is clear from the actual queries shown in Tables 5 and 6 how diverse needs this approach captures.

In a semi-steady state Before COVID-19 situation, it is half natural that the macro indicators of two months ago are not much different from those of today, and that models based on these indicators are likely to be highly accurate.

4.2. Query analysis from the proposed model

Figure 5 shows the time series analysis through Before COVID-19 and Under COVID-19 in the proportion of query categories used in *Our Model*. Queries related to economic activity are found in both CI and CCI. In particular, CI has a larger proportion of queries related to economic activity when compared with CCI, and this proportion increases even more when the timing of economic changes is significant. It seems to be attributed to the fact that the CI is generated by a combination of economic statistics, while the CCI is generated by consumer surveys.

We also would like to note that the CI model from October 2019 to December 2019 is composed of 100% of economy category query. Although usually Our models are composed of about 30 selected queries each month, the best results in this case were obtained with a model consisting of only one word related to the economic indicator ("employment adjustment subsidy"; *koyochoseijoseikin* in Japanese).

We noticed that two query categories; Libido and Comedy, are quite distinctive, though they are not common in top queries used as seen in Tables 5 and 6 Libido category includes adult entertainment services and adult entertainment celebrities. Comedy category includes the names of comedians, and their well-known performances. Libido category outnumbers Economy category in the case of CCI modeling. Moreover, the presence of Libido in CCI increases Under COVID-19 period, when the economy experienced a big discontinuity. In the case of CI modeling, the ratio of queries related to Comedy doubled from Before COVID-19 to Under COVID-19.

Table 6
Top 20 queries used for CCI nowcasting from April 2019 to March 2021 among 24 models.

Search query	Query meaning	Category	Query utilization(%)
Kinoshita	name of famous comedian	Comedy	91.7
M2	a car name	Others	89.3
jukujyonoshiro	adult entertainment	Libido	79.2
Toriyosegurume	gourmet food delivery site	Others	75.0
Hotto Ishikawajima	a food company	Others	75.0
harima jyukogyo	company in heavy industry	Others	75.0
note pc	laptop computer	Others	70.8
Niigata fuzoku	adult entertainment	Libido	66.7
Tokyoshokorisaachi	credit bureau	Economic Activities	58.3
2 syotto	a type of shot	Others	58.3
Soul Eater	a manga name	Others	54.2
Niconico	a popular video site	Others	50.0
desukutoppu pc	desktop computer	Others	45.8
La Boheme	a cafe name	Others	41.7
Akimoro Junko	a singer name	Others	41.7
kokkoro	a character name	Others	41.7
flex	a car resale company	Others	41.7
tip	-	Others	41.7
Unicity	a consumer product company	Others	41.7
Minorin	a Youtuber	Others	41.7
musyuseigamitai	adult video genre	Libido	41.7
hanabatake	flower garden	Others	41.7
Kenta	a fast food company	Others	41.7
plaza	a variety store	Others	41.7
tennki	numeric keypad	Others	41.7
Hosyokyokai	guarantee association	Economic Activities	41.7

The results of these categorical analyses suggest what human motivations are behind the economic indicators from this multiple query based modeling approach. We believe that these results pave the way for the quantitative study of the human needs and psychological aspects behind the economy and business climate, which has been quite challenging until now.

4.3. Model selection

We experimented with four modeling techniques: Multiple Regression, Random Forest, Ridge Regression, and XGboost. Table 7 shows the predictive results during the "Before COVID-19" and "Under COVID-19" periods when we adapted the different modeling approaches. Machine learning methods such as Random Forest and XGboost performed well in the "Before COVID-19" period. However, their performance dropped significantly during the "Under COVID-19" period, suggesting a high probability of overfitting. Conversely, Multiple Regression, while slightly less powerful than the other approaches, showed no deterioration in predictive performance during the "Under COVID-19" period, indicating greater stability in its predictions. Therefore, we adopted Multiple Regression as our proposed approach for our final modeling.

4.4. Prospective application

This series of analyses demonstrate the effectiveness of models utilizing Search Big Data in estimating under disruptive conditions or when rapid reporting is necessary. We also found that when using search query data, it is more effective to build a model based on multiple queries rather than a single query, which can improve forecasting performance in both normal and emergency situations.

In addition to the impact of COVID-19 examined in this study, unforeseen events such as the Great East Japan Earthquake of 2011 can occur in any economy, and the ability to forecast economic indicators during such events is invaluable. Furthermore, the instantaneity of query data aggregation allows for nowcasting on a daily as well as a monthly basis. Given the dynamic nature of such situations, we believe the proposed method has the potential to play a crucial role in a wide range of decision-making processes.

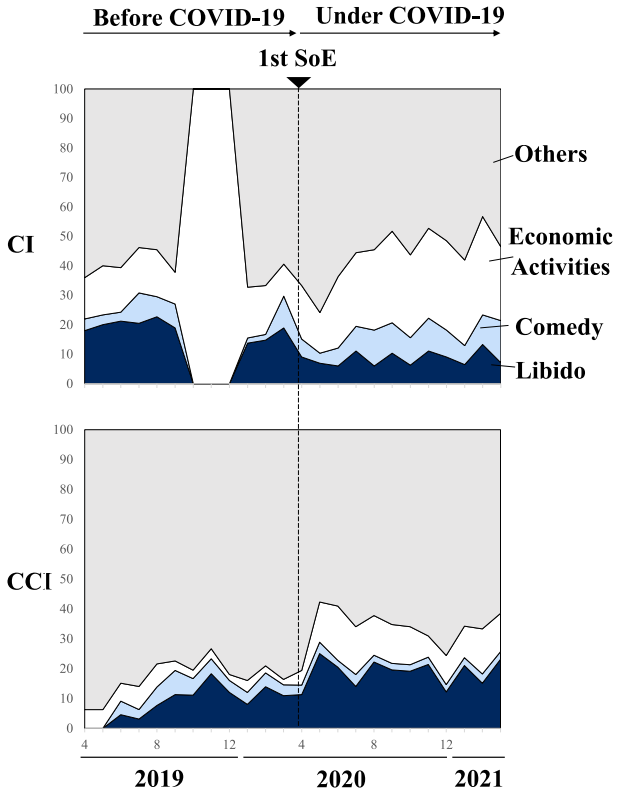

Fig. 5. Proportions of query genres used for CI and CCI nowcasting during Before COVID-19 and Under COVID-19 periods.

Table 7
Comparison of models estimating CI and CCI values Before and Under COVID-19 period, and their Deterioration Rates.

	CI			CCI		
Model	Before COVID-19	Under COVID-19	Deterioration Rate	Before COVID-19	Under COVID-19	Deterioration Rate
Multiple regression	31.8	34.4	**1.1**	21.9	22.5	**1.0**
Random Forest	14.7	53.7	3.7	15.9	48.4	3.0
Ridge regression	7.2	59.8	8.4	11.7	16.7	1.4
XGBoost	6.7	42.9	6.4	10.4	35.6	3.4

4.5. Limitation

Our method has five limitations.

Initially, our proposed model exhibits a high degree of accuracy in approximating the actual data regarding the decline in economic indicators; however, it portrays a more optimistic recovery from disruptive events than what is observed in reality. Consequently, it is imperative to exercise caution when employing the model for policy making, as there exists a potential risk of overestimating the rate of economic recovery. From a practical standpoint, it is critical to respond quickly to rapid economic fluctuations, and our model's ability to predict the timing of these changes contributes to the field.

Secondly, in terms of estimating economic indicators, using data from periods of significant economic change can contribute to the creation of a robust model. Incorporating data from a longer time frame, including periods of change, may enhance the performance of the proposed model.

The third limitation is that we only evaluated the COVID-19 period as a discontinuous situation, although our dataset includes three major discontinuous situations: the Great Recession of 2008, the Great East Japan Earthquake of 2011, and the COVID-19 of 2020. Although we also attempted to do nowcasting in two additional cases (2008–2009 and 2011), the accuracy was not high. We suspect that this is largely due to the fact that less than 5 years of training data were available for the two cases, while 15 years of training data were available for COVID-19. We plan to expand our dataset, tackle predictions for other disruptive events that will occur, and evaluate the adaptability and robustness of our approach.

The fourth limitation pertains to the utilization of Google Translate to convert the characteristics of English research papers into Japanese. Since we used Google Translate, there is a possibility that the original language context may not be fully reflected when translated into Japanese queries, which could potentially affect the results. This process is a limitation when applying methods from other languages to Japanese. Note, however, that various translation tools, not limited to Google Translate, can be used. Since the process does not involve arbitrariness or individual judgments, it maintains reproducibility.

Finally, it is important to note that while this approach allows for the identification of economic trends prior to the release of various statistical data, it does not diminish the value of current indicators and methods that rely on statistical data and are released several months later.

5. Conclusion

We developed a fully data-driven methodology for nowcasting economic indicators, which are typically released one to two months later, using Japan's leading dataset of search query data (Search Big Data).

Existing methods of nowcasting models with economic variables and manually selected search queries directly related to the economy showed a dramatic drop in nowcasting performance in discontinuous situations such as COVID-19, whereas the proposed method did not show a drop in performance Before or Under COVID-19. This indicates that the proposed method is able to achieve highly accurate nowcasting without changing before and after discontinuous situations.

We believe this is because the data-driven approach to query selection is successful in extracting elements such as people's behavior and thoughts that make up the economy. As long as the principles of human behavior remain unchanged, this approach will be effective in nowcasting the economy.

Moreover, analysis of the key variables (search queries) in the generated model revealed a quantifiable link between “human interest” and economic indicators, suggesting the potential for utilizing search queries to gain insight into human activities related to economic indicators. Notably, we found that fundamental human needs, such as libido and desire for laughter, were correlated with economic indicators.

Funding

This research was conducted using the resources of Yahoo Japan Corporation.

Declaration of competing interest

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.

Acknowledgements

We deeply appreciate the editor and the anonymous reviewers for their invaluable feedback, which significantly improved the quality of this paper. We also thank Miho Kuroda and Minami Ueda for their precursor studies. We are grateful to Mayu Inaba, Daiki Yoshitsugu, and Souta Noguchi for their diligent research assistance. Of course, any remaining errors in the paper are our sole responsibility.

References

Aguilar, P., Ghirelli, C., Pacce, M., Urtasun, A., 2021. Can news help measure economic sentiment? An application in COVID-19 times. Econ. Lett. 199, 109730. https://doi.org/10.1016/j.econlet.2021.109730.

Askitas, N., Zimmermann, K.F., 2009. Google Econometrics and Unemployment Forecasting.

Ataka, K., 2016. Revisiting the big data-driven forecast of Japan's 24th upper house election by Yahoo! Japan. Policy and Research 11, 75–82. https://doi.org/10.24561/00018267.

BańBura, M., Giannone, D., Reichlin, L., 2011. Nowcasting. In: Clements, M.P., Hendry, D.F. (Eds.), Oxford Handbook on Economic Forecasting. Oxford University Press, pp. 63–90.

BańBura, M., Giannone, D., Modugno, M., Reichlin, L., 2013. Now-casting and the real-time data flow. In: Elliott, G., Timmermann, A. (Eds.), Handbook on Economic Forecasting, 2A. North Holland, pp. 195–237.

Bouayad, I., Zahir, J., Ez-zetouni, A., 2022. Nowcasting and forecasting Morocco gdp growth using google trends data. IFAC-PapersOnLine 55 (10), 3280–3285. https://doi.org/10.1016/j.ifacol.2022.10.129, 10th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2022. https://www.sciencedirect.com/science/article/pii/S2405896322021395.

Bragoli, D., 2017. Now-casting the Japanese economy. Int. J. Forecast. 33 (2), 390–402. https://doi.org/10.1016/j.ijforecast.2016.11.004. https://www.sciencedirect.com/science/article/pii/S0169207016301297.

Bragoli, D., Fosten, J., 2018. Nowcasting Indian GDP. Oxf. Bull. Econ. Stat. 80 (2), 259–282. https://doi.org/10.1111/obes.12219.

Cabinet Office, 2021. Government of Japan, Indexes of Business Conditions. https://www.esri.cao.go.jp/en/stat/di/di-e.html. (Accessed 25 September 2021).

Carriero, A., Clark, T.E., Marcellino, M., 2015. Realtime nowcasting with a Bayesian mixed frequency model with stochastic volatility. J. Roy. Stat. Soc. 178 (4), 837–862.

Carriero, A., Clark, T.E., Marcellino, M., Giuseppe, M., 2020. Nowcasting tail risks to economic activity with many indicators. FRB of Cleveland Working. https://doi.org/10.26509/frbc-wp-202013r2. Paper No.20-13R2.

Cavallo, A., Rigobon, R., 2016. The billion prices project: using online prices for measurement and research. J. Econ. Perspect. 30 (2), 151–178. https://doi.org/10.1257/jep.30.2.151.

Cepni, O., Güney, I.E., Swanson, N.R., 2019. Nowcasting and forecasting GDP in emerging markets using global financial and macroeconomic diffusion indexes. Int. J. Forecast. 35 (2), 555–572. https://doi.org/10.1016/j.ijforecast.2018.10.008.
Chauvet, M., Potter, S., 2013. Forecasting output. In: Elliott, G., Timmermann, A. (Eds.), Handbook on Economic Forecasting, 2A. North Holland, pp. 141–294.
Chen, T., So, E.P.K., Wu, L., Yan, I.K.M., 2015. The 2007–2008 U.S. recession: What did the real-time google trends data tell the united states? Contemp. Econ. Pol. 33 (2), 395–403. https://doi.org/10.1111/coep.12074.
Chikamatsu, K., Hirakata, N., Kido, Y., Otaka, K., et al., 2018. Nowcasting Japanese GDPs. Bank of Japan.
Choi, H., Varian, H., 2012. Predicting the present with google trends. Econ. Rec. 88, 2–9.
Comelli, M.F., 2015. Estimation and Out-Of-Sample Prediction of Sudden Stops: Do Regions of Emerging Markets Behave Differently from Each Other? International Monetary Fund.
Della, N., Penna, Huang, H., 2009. Constructing Consumer Sentiment Index for Us Using Internet Search Patterns, 26. Department of Economics, WP.
D'Amuri, F., Marcucci, J., 2010. 'google It!'forecasting the Us Unemployment Rate with a Google Job Search Index.
D'Agostino, A., Giannone, D., Lenza, M., Modugno, M., 2016. Nowcasting business cycles: a Bayesian approach to dynamic heterogeneous factor models. Adv. Econom. 35, 569–594.
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.
Evans, M.D., 2005. Where are we now? Real-time estimates of the macro economy. International Journal of Central Banking 1 (2), 127–176.
Hayashi, F., Tachi, Y., 2022. Nowcasting Japan's gdp. Empir. Econ. 1–37.
Hisada, S., Murayama, T., Tsubouchi, K., Fujita, S., Yada, S., Wakamiya, S., Aramaki, E., 2020. Surveillance of early stage COVID-19 clusters using search query logs and mobile device-based location information. Sci. Rep. 10 (1), 1–8. https://doi.org/10.1038/s41598-020-75771-6.
Jaemin, W., Owen, A.L., 2019. Forecasting private consumption with google trends data. J. Forecast. 38 (2), 81–91. https://doi.org/10.1002/for.2559 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.2559. https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2559.
Kohns, D., Bhattacharjee, A., 2023. Nowcasting growth using google trends data: a bayesian structural time series model. Int. J. Forecast. 39 (3), 1384–1412. https://doi.org/10.1016/j.ijforecast.2022.05.002. https://www.sciencedirect.com/science/article/pii/S0169207022000620.
McLaren, N., Shanbhogue, R., 2011. Using internet search data as economic indicators. Bank Engl. Q. Bull. 51 (2), 134–140. URL. https://EconPapers.repec.org/RePEc:boe:qbullt:0052.
Ministry of Internal Affairs and Communications, 2021. Key Points of the 2021 White Paper on Information and Communications in Japan. https://www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2021/key-points.pdf. (Accessed 16 March 2022).
Proietti, T., Giovannelli, A., 2021. Nowcasting monthly GDP with big data: a model averaging approach. J. Roy. Stat. Soc. 184 (2), 683–706.
Radinsky, K., Davidovich, S., Markovitch, S., 2008. Predicting the news of tomorrow using patterns in web search queries. In: 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 1. IEEE, pp. 363–367.
Richardson, A., van Florenstein Mulder, T., Vehbi, T., 2021. Nowcasting GDP using machine-learning algorithms: a real-time assessment. Int. J. Forecast. 37 (2), 941–948. https://doi.org/10.1016/j.ijforecast.2020.10.005.
Sasaki, W., Kawase, H., Miyahara, S., Tsubouchi, K., Okoshi, T., 2021. Nation-wide Mood: Large-Scale Estimation of People's Mood from Web Search Query and Mobile Sensor Data. arXiv preprint arXiv:2111.05537. https://doi.org/10.48550/arXiv.2111.05537.
Siliverstovs, B., 2020. Assessing nowcast accuracy of US GDP growth in real time: the role of booms and busts. Empir. Econ. 58 (1), 7–27. https://doi.org/10.1007/s00181-019-01704-6.
Vosen, S., Schmidt, T., 2011a. Forecasting private consumption: survey-based indicators vs. google trends. J. Forecast. 30 (6), 565–578.
Vosen, S., Schmidt, T., 2011b. Forecasting private consumption: survey-based indicators vs. google trends. J. Forecast. 30 (6), 565–578. https://doi.org/10.1002/for.1213 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.1213. https://onlinelibrary.wiley.com/doi/abs/10.1002/for.1213.
Vosen, S., Schmidt, T., 2012. A monthly consumption indicator for Germany based on internet search query data. Appl. Econ. Lett. 19 (7), 683–687. https://doi.org/10.1080/13504851.2011.595673 arXiv:
Woloszko, N., 2020. Tracking Activity in Real Time with Google Trends. OECD Economics Department Working Papers (1634). https://doi.org/10.1787/6b9c7518-en. https://www.oecd-ilibrary.org/content/paper/6b9c7518-en.