Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663

https://www.czasopisma.uni.lodz.pl/foe/
3(368) 2024
DOI: https://doi.org/10.18778/0208-6018.368.04

Klaudia Lenart https://orcid.org/0000-0001-8135-9362

University of Economics in Katowice, Department of Statistics, Econometrics and Mathematics, Katowice, Poland, klaudia.lenart@edu.uekat.pl

Applications of Google Trends as a Data Source for Statistical Models

Abstract: As technology advances, there is a growing number of potential data sources that can provide an alternative to traditional surveys. An example of this is the real-time search popularity data made available through Google Trends. This type of data makes it possible to study public opinion, behaviour and attitudes in society or forecast economic phenomena.

A definite advantage of using search popularity data is the immediate availability and low cost of obtaining such data. Also of significance is the fact that the Google Trends tool allows for direct research into the behaviour of Internet users, and not just their declarations as in the case of a survey. This can make a difference if respondents consider one of the answers to be more morally correct. Nevertheless, the use of Google Trends requires selecting correct search topics and terms to be included in the study and an awareness of the fact that the research sample is limited to Google search engine users. The paper will present the advantages and disadvantages of Google Trends and review its usefulness as a data source especially in times of higher market volatility.

Keywords: Google Trends, statistical analysis, prognosis

JEL: C8, C51, C53

1. Introduction

When developing a model, one of the most difficult challenges a researcher faces is finding data that can be incorporated into it. The difficulty is caused not only by the availability of data, but also by aggregation, the frequency of publication, the accuracy of the measurement and, in the case of surveys, the honesty of the respondents.

The Internet is used not only as a source of information but also as a place to shop or publish adverts, which means that the search engine query data contain a wide spectrum of information about behaviours and attitudes such as job hunting, an intention to buy a particular product or even political views. As a result, Google Trends can be used as a data source for modelling macroeconomic variables. It is particularly utile when a variable is measured officially, but the values are reported with a delay. For example, the Polish Central Statistical Office announces the unemployment rate for a given month in the following month. Such post-factum information may not be sufficient, which presents statisticians with a challenge referred to in the literature as nowcasting, that is, predicting the value of a statistic in the present, very near past or future (BańBura, Giannone, Reichlin, 2012). As Google Trends data are available almost in real time, they can be used to predict the current value of a given variable. It is important to note that the ability to access the data for the current period can be especially valuable when nowcasting variables during periods of higher market volatility or when rare or unprecedented events occur. In the empirical study conducted, models of unemployment rate and number of apartments sold utilising Google Trends data have been estimated and compared with their benchmark models. In both cases, the precision of the ex post forecast during the periods of higher uncertainty caused by the pandemic has been tested.

2. Google Trends as a source of data

Google Trends was launched in 2006. However, it should be noted that the data were not updated in real time from the beginning. Through this website, Google provided statistical data on search popularity. In 2008, Insights for Search was created, which provided much better data analysis capabilities, and in 2012 the two platforms were merged. Initially, the use of search popularity data was seen mainly in marketing, but the usefulness of such data in prediction and estimation was quickly discovered. Even a year before the launch of Google Trends, the use of search query popularity data in the prediction of the unemployment rate was considered by Ettredge, Gerdes, and Karuga (2005).

The first year for which data are available via Google Trends is 2004, but it should be noted that the exact method of measurement has been changed several times since then. The selected time frame influences the degree of aggregation of the data generated by the website, in the case of long periods, it will be monthly or weekly data. It should also be noted that the degree of aggregation depends on how recent the data we are interested in are. For the last hour, we can get an observation for each minute, but data from a few weeks ago will be aggregated into days. Google Trends also makes it possible to track the popularity of searches both worldwide and in a selected country. In the case of countries, data are available on the first-level administrative division (voivodeships in Poland) and cities, but in the case of the latter, only selected locations are available. It needs to be stressed that the data available through Google Trends are not simply the number of all searches for a given topic. The huge number of search queries that are entered into Google every day means that it would be impossible to quickly process a dataset of this size. Due to this fact, only a sample of search queries is examined. To make working with the data easier, search topics were created. They include all searches within a given scope, regardless of the typos or vocabulary used.

The data are subjected to the following transformations (Google_Trends_Data, 2023):

1) observations that are identified as automated attempts to manipulate or distort the data collected by Google are removed (however, it is sometimes the case that they are deliberately not removed from the data available via Google Trends so as not to reveal the detection of such incidents);
2) searches on unpopular topics, repeated searches by the same person and searches containing special characters, such as apostrophes, are not included;
3) data are transformed to take values between 0 and 100.

For Google Trends data to be successfully used in a model, it is crucial to identify the search topics and terms that are connected with the dependent variable. For example, in the case of nowcasting the unemployment rate using Google Trends, it is assumed that an increase in unemployment results in more people searching for job offers online, thus increasing the popularity of searches on this topic. However, it should be stressed that entering a specific keyword in Google is not equivalent to a direct answer to a question. It is much less straightforward, as an employed person can also search for job offers looking for a better-paying position. Internet users can also search for something just out of curiosity since browsing car sales is not tantamount to expressing a desire to buy a car.

When comparing Google Trends to traditional surveys, an evident advantage of the former is the fact that this tool is costless and allows one to obtain new data instantaneously. It is also important to note that respondents tend to choose untrue but more socially acceptable answers on many issues perceived as embarrassing or related to morality, even if the survey is anonymous. This can lead, for example, to overestimation of voter turnout, as people who are not interested in voting may declare their intent to vote. One of the biggest disadvantages of Google Trends is hidden in its very name. The data available through the platform are limited to information only about web search queries made in the Google search engine. While Google is the most popular search engine in many, especially Western, countries, its position on the market is not that good in every country. Due to this fact, Google Trends data would be useless for making predictions about countries such as China. Another worth noting concern relating to the representativeness of the sample is the fact that Google Trends data do not include information about people who do not use the Internet.

3. Google Trends data applications

One of the best-known attempts to use search data for nowcasting was a project created by Google itself to estimate the number of patients who report flu symptoms. Google Flu Trends was intended as an alternative to the U.S. Centres for Disease Control and Prevention (CDC) reports, published with a one- to two-week delay, which are the main source of data on the subject. The idea behind using search popularity data to predict the number of sick people is fairly intuitive, as people with flu symptoms are much more likely to type in search terms related to the disease and its symptoms. Ginsberg et al. (2009), however, went one step further while creating Google Flu Trends. Using data from 2003 to 2007, they tested 50 million searches to see which ones correlated most closely with the number of patients reported by the CDC. The final model contained 45 independent variables. Testing it on data from 2007–2008 yielded very promising results: the predictions made with it were 97% accurate. Unfortunately, the model started failing as early as 2009 when it significantly underestimated the number of cases of swine flu during the epidemic. At the time, a change in the typical searches performed by people sick with swine flu was cited as the reason for the significant deterioration in the model's performance, but time quickly verified the initially reported accuracy of the model's prediction. Between 2011 and 2013, it significantly overestimated the number of cases of flu, sometimes even doubling it (Butler, 2013).

The way search terms were selected was partly responsible for the failure of Google Flu Trends. In Ginsberg et al. (2009), the creators of the model mentioned that some searches indicated as potential explanatory variables by the algorithm they used were clearly not related to influenza, but only correlated with the number of cases due to their seasonality. For this reason, searches about, for example, high school basketball were excluded. Although Google Flu Trends failed, the idea of using search data to estimate the number of cases of a disease was not forsaken, becoming especially topical during the COVID–19 pandemic (Saegner, Austys, 2022).

The increasing popularity of online shopping makes Google Trends data an increasingly valuable source of information about household expenditures. As with the CDC-reported number of cases of flu, the official reports about household expenditures are published with considerable delay. Using surveys to estimate these figures is also not a perfect solution, as respondents often state in their answers their expectations and plans instead of the actual situation. Due to this fact, it can be useful to incorporate Google Trends data in models forecasting private consumption (Vosen, Schmidt, 2011). The delay in data reporting may be an especially relevant issue in emerging markets, where data are often reported with a longer lag than in developed countries. Carrière-Swallow, and Labbé (2013) have managed to successfully utilise Google Trends data to nowcast automobile sales in Chile. Google Trends was a particularly useful source of information about this sector partially because buying a car is an important decision, often requiring a lot of research prior to the purchase. The authors also point out that the industry is dominated by a relatively small number of brands, which allowed for an easier identification of keywords for which popularity indexes were aggregated. Tourism is another sector for which data about popularity of search queries can be a valuable source of information about consumer interest. Li et al. (2017) have constructed a composite search popularity index using the generalised dynamic factor model (GDFM). It is important to note that, because the authors' were forecasting the tourist volume in Beijing, they used Baidu search trend data instead of Google Trends, as the former represents a more popular search engine in China. For the purpose of selecting the keywords that can be included in the index, the thought process of a Chinese tourist planning a trip was considered. Google Trends data has also been used to predict global oil consumption by Yu et al. (2019).

Another area in which Google Trends data can assist or even to some extent replace survey research is opinion polling, for example, concerning issue salience. Which issues the voters care about the most is of particular importance during election campaigns. The surveys traditionally used for this purpose leave much to be desired; unlike Google Trends data, they do not allow for continuous monitoring. The idea behind the use of Google Trends data for this purpose is simple: searching the Internet for information on a given topic usually indicates interest in that topic (Mellon, 2014). Monitoring the popularity of selected keywords may also help to preliminarily examine society's reactions and attitude towards new policies or arising difficult situations. During lockdown, Brodeur et al. (2021) used popularity of keywords such as boredom, sadness and loneliness to assess the impact of the policies implemented during the pandemic on the well-being in nine Western European countries.

There are also areas where the use of surveys is impossible for practical reasons. In the case of investments in stock markets, often hours or even minutes can determine large gains and losses. Providing access to present-time data, Google Trends is, therefore, a good alternative, which can serve as a source of information on the behaviour of other investors in the market as well as reflecting the current economic situation, and has been used as such by Hu et al. (2018) in forecasting the direction of the S&P 500 index price. Unlike in the examples mentioned above, Google Trends data were used to forecast future prices of securities and not to predict the current value, as the prices on the stock markets are reported in real time. Google Trends data is particularly useful when the market is characterised by high volatility. An example would be cryptocurrencies, whose value fluctuates very dynamically and is at the same time highly dependent on the interest in a particular cryptocurrency. In this case, Google Trends is an especially valuable source of information. In particular, an increase in the popularity of searches on a given cryptocurrency is usually highly correlated with an increase in its value (Zhang, Wang, 2020).

4. Nowcasting the unemployment rate in Poland

To provide an estimate of the registered unemployment rate in Poland before the official value is published by the Polish Central Statistical Office, ARIMAX models were estimated. This variable will be denoted as RUR throughout this paper. In order to verify the usefulness of Google Trends data for nowcasting the unemployment rate, ARIMAX models using the search popularity data will be compared with exponential smoothing, ARIMA and SARIMA models, within this study. The ARIMAX models were estimated using the function auto.arima(), which is part of the forecast package in R (Hyndman, Khandakar, 2008; Hyndman et al. 2024). This function automatically tests the stationarity of the explanatory variable by applying the KPSS test. The values of the parameters are then selected based on the AIC criterion, which takes into account both the accuracy of the fit to the data and the complexity of the models. The benchmark ARIMA and SARIMA models were estimated using the same method.

Table 1. Search terms and topics included in the study

Variable	Searches	Type	Correlation with RUR (2009–2021)
GT1	Oferty pracy (job offers)	Topic	0.5449
GT2	Jak znaleźć pracę (how to find a job)	Term	0.6607
GT3	Praca od zaraz (job straight away)	Term	0.7102
GT4	Praca za granicą (working abroad)	Term	0.8633
GT5	Gdzie szukać pracy (where to find a job)	Term	0.4411
GT6	Pracuj.pl (popular portal with job offers)	Topic	0.7102
GT7	Rozmowa kwalifikacyjna (job interview)	Topic	0.7052
GT8	Zasiłek dla bezrobotnych (unemployment benefit)	Term	0.5294
GT9	Urząd pracy (employement office)	Topic	0.6888
GTsr	Mean of GT1-GT9	–	0.9020

Source: author’s own work

The initial selection of search terms and topics is presented in Table 1, where two types of searches can be seen. The terms and topics GT1–GT7 are connected with job hunting as more unemployed people means more people looking for a job. It is important to keep in mind that not every person who is looking for a job qualifies to be registered as unemployed in Poland (for example: a university student) or even is unemployed. More and more often people search for new opportunities and better pay even if they have a stable job already. The remaining searches are connected with the registration as an unemployed person in Poland.

ARIMAX models were estimated using data from 2009–2018, each including one of the variables presented in Table 1. The results are shown in Table 2.

Table 2. ARIMAX models containing Google Trends variables (2009–2018)

Variable	Search	Model (p,d,q)	RMSE
GT9	Urząd pracy (employement office)	(2,1,1)	0.1639
GT4	Praca za granicą (working abroad)	(2,1,3)	0.1669
GT3	Praca od zaraz (job straight away)	(2,1,1)	0.1697
GTsr	Mean of GT1-GT9	(2,1,1)	0.1717
GT8	Zasiłek dla bezrobotnych (unemployment benefit)	(2,1,1)	0.1788
GT6	Pracuj.pl (popular portal with job offers)	(2,1,1)	0.1798

Variable	Search	Model (p,d,q)	RMSE
GT1	Oferty pracy (job offers)	(2,1,1)	0.1801
GT7	Rozmowa kwalifikacyjna (job interview)	(2,1,1)	0.1852
GT5	Gdzie szukać pracy (where to find a job)	(2,1,1)	0.1865
GT2	Jak znaleźć pracę (how to find a job)	(2,1,1)	0.1866

Source: author's own work

The ARIMAX models with the lowest RMSE values (from GT9 to GT6) were then compared with the benchmark models not using Google Trends. The results are shown in Figure 1.

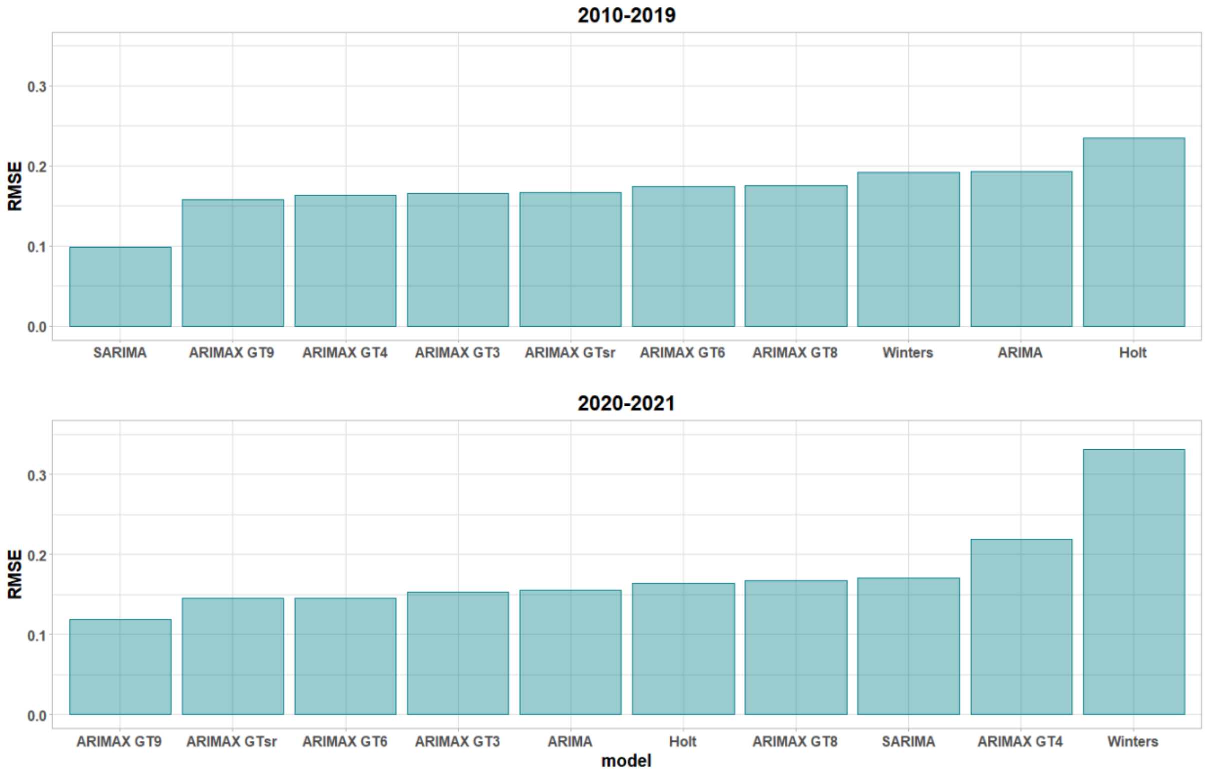

Figure 1. Comparison of RMSE values of the estimated models
Source: author's own work in the R program

The RMSE calculated for the period of 2010–2019 is a measure of how well the observed values of the unemployment rate were replicated by the model. As shown in Figure 1, the SARIMA model's RMSE (0.0984) in that period was much lower than the RMSE observed for the ARIMAX models using the Google Trends data. The situation changes when we compare the RMSE values calculated for the data unused in the process of estimating the model parameters (2020–2021). Those values can be interpreted as a measure of precision of the ex post forecast. For the period of 2020–2021, the lowest RMSE was observed for the ARIMAX model using the popularity of the search topic Urząd pracy (employment office). It is important to note that the period for which the forecast precision was evaluated was not chosen randomly. In the years 2020 and 2021, the COVID–19 pandemic caused significant market turmoil. It is in the conditions of fast, unpredictable changes that the reported in real-time Google Trends data are an especially valuable source of information as shown by the estimated models.

5. Nowcasting the number of apartments sold

The second application of Google Trends data implemented in this paper refers to the prediction of the number of apartments sold. This type of data is reported by the Polish Central Statistical Office quarterly. The purpose of the models will be to estimate the number of apartments sold in Poland and a few chosen Voivodeships after two months of that quarter have passed. This means that the Google Trends data for each period will be taken only from the first two months of the quarter.

The form of the estimated models can be written as:

$$M_t = a_0 + a_1 GT_t + a_2 t + B_1 V_{1t} + B_2 V_{2t} + B_3 V_{3t} + r_t, \tag{1}$$

where:

M_t – number of apartments sold in the period *t*,
GT_t – popularity index of search topic ‘apartment’ in the period *t*,
V_{it} – seasonal variable, equal to 1 in the *i*-th quarter, –1 in quarter 4 and 0 in the remaining quarters,
r_t – residual in the period *t*.

The value of the parameters estimated using the method of least squares is shown in Table 3. When none of the seasonal variables were significant, those variables were excluded from the model.

Table 3. Estimated parameters and R^2 values of the models predicting the number of apartments sold

Region	a_0	a_1	a_2	B_1	B_2	B_3
Poland	5294.01	182.86	381.03	506.91	2225.13	– 2120.10
Śląskie	795.14	13.93	31.03	–	–	–
Łódzkie	411.26	4.75	38.14	–	–	–
Dolnośląskie	1267.12	17.33	55.68	160.14	128.13	– 291.89

Source: author’s own work

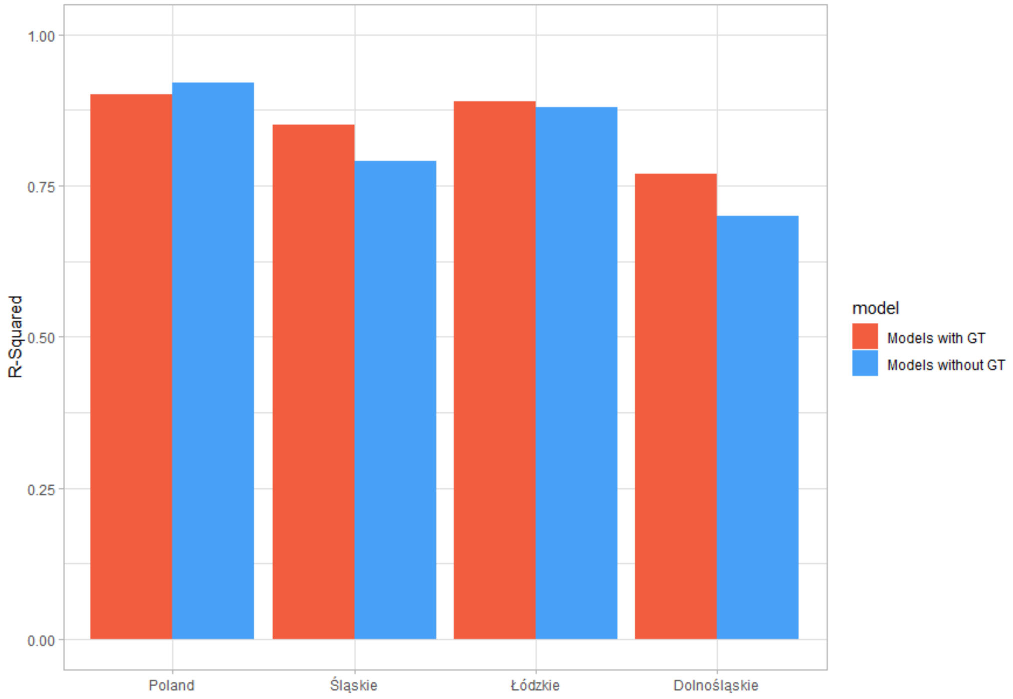

Figure 2. Comparison of R-Squared values of estimated models with and without Google Trends data (2010–2020)
Source: author's own work in the R program

As shown in Figure 2, the addition of the Google Trends variable did not significantly improve how well the models explained the variance of the independent variable. However, just like in the case of nowcasting the unemployment rate, the more important measure is the precision of the predictions made for the data that were not used during the process of model parameters estimation.

The values of MAPE estimated for 2021 data (which were not used to estimate the model parameters) are presented in Figure 3. As shown in Figure 3, in most cases, better accuracy was achieved by the models incorporating Google Trends data. The difference was especially noticeable for the model predicting the number of apartments sold in Poland.

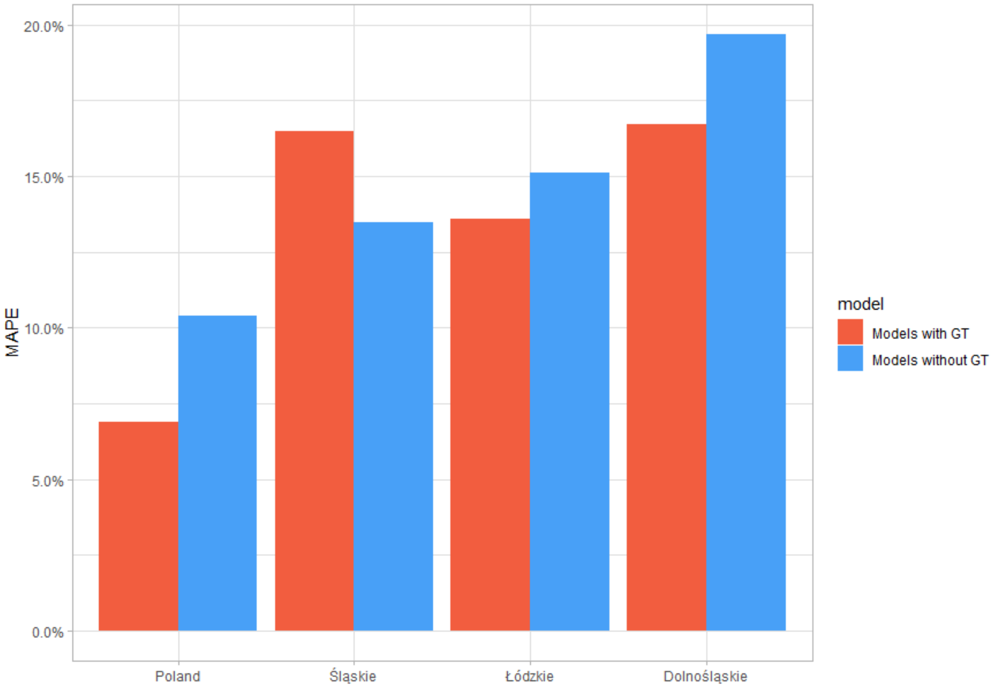

Figure 3. Comparison of MAPE values of estimated models with and without Google Trends data (2021)

Source: author's own work in the R program

A more specific reason for the lower values of MAPE observed in Figure 3 can be seen by comparing the predicted values with the actual number of apartments sold presented in Figure 4.

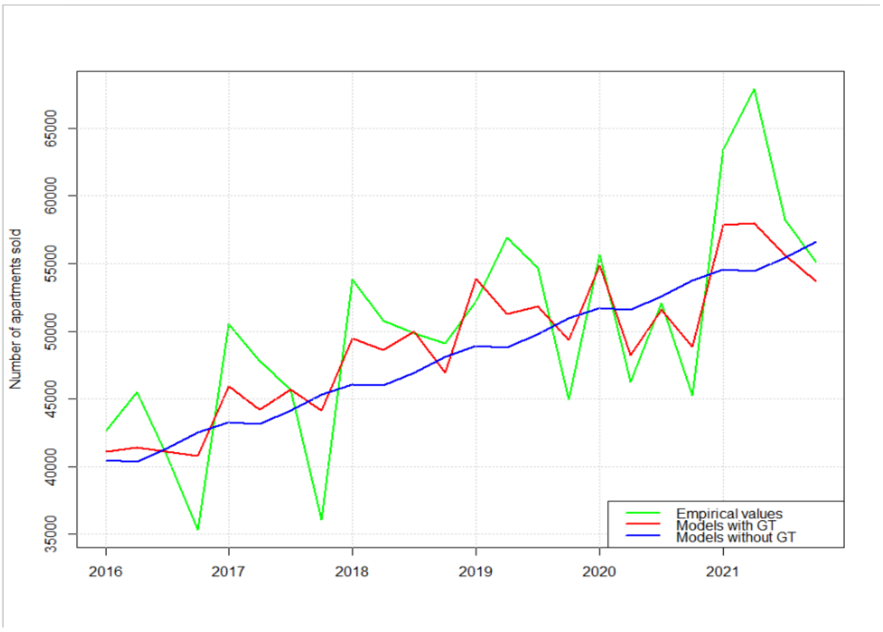

Figure 4. The number of apartments sold in Poland and the predictions of this variable

Source: author's own work in the R program

Figure 4 shows that a rapid increase of the number of apartments sold in Poland was recorded in the year 2021. Although the values predicted by the model that incorporated the Google Trends data failed to match the scale of that increase, the additional information about the search popularity allowed for a much more accurate prediction.

6. Conclusions

As shown by the models presented in this paper, incorporating Google Trends data may be especially helpful when the predictions are made in times of high variability and changing conditions. The wide range of uses for search popularity data makes Google Trends a source of data worthy of attention, particularly when faced with limited funding or when a frequent assessment of the situation is needed. It is also important to note that Google Trends may be a source of information on topics on which almost no other data are available. At the same time, the limitations of Google Trends must always be taken into account, as the data relate only to people using Google's search engine. Due to this fact, when researching countries such as China, data regarding other search engines should be utilised instead. When selecting searches whose popularity index can be incorporated into a model, one cannot be guided by the correlation coefficient alone. It is always necessary to establish a cause-and-effect relationship between the object of study and the increase or decrease in popularity of searches regarding a given topic. The selection of initial variables will therefore be a subjective process whose success or failure depends on the researchers' knowledge and assessment of Internet users' behaviours. The successful identification of correct keywords also depends on the nature of the research subject. When forecasting consumption or sales, search queries for purchases that are rare, important and expensive, such as cars or apartments or ones that require substantial research such as planning a trip, are more easily identifiable.

References

BańBura M., Giannone D., Reichlin L. (2012), *Nowcasting*, [in:] M.P. Clements, D.F. Hendry (eds.), *The Oxford Handbook of Economic Forecasting. Oxford Handbooks Online*, Oxford University Press, Oxford, pp. 193–224.

Brodeur A., Clark A.E., Fleche S., Powdthavee N. (2021), *COVID–19, lockdowns and well-being: Evidence from Google Trends*, "Journal of Public Economics", vol. 193, 104346.

Butler D. (2013), *When Google got flu wrong: US outbreak foxes a leading web-based method for tracking seasonal flu*, "Nature", vol. 494, pp. 155–157.

Carrière-Swallow Y., Labbé F. (2013), *Nowcasting with Google Trends in an emerging market*, "Journal of Forecasting", vol. 32, pp. 289–298.

Ettredge M., Gerdes J., Karuga G. (2005), *Using web-based search data to predict macroeconomic statistics*, "Communications of the ACM", vol. 48, pp. 87–92.

Ginsberg J., Mohebbi M.H., Patel R.S., Brammer L., Smolinski M.S., Brilliant L. (2009), *Detecting influenza epidemics using search engine query data*, "Nature", vol. 457, pp. 1012–1014.

Google_Trends_Data (2023), *FAQ about Google Trends data*, https://support.google.com/trends/answer/4365533?hl=en [accessed: 28.11.2023].

Hu H., Tang L., Zhang S., Wang H. (2018), *Predicting the direction of stock markets using optimized neural networks with Google Trends*, "Neurocomputing", vol. 285, pp. 188–195.

Hyndman R.J., Khandakar Y. (2008), *Automatic time series forecasting: the forecast package for R*, "Journal of Statistical Software", vol. 27, pp. 1–22.

Hyndman R.J., Athanasopoulos G., Bergmeir C., Caceres G., Chhay L., O'Hara-Wild M., Petropoulos F., Razbash S., Wang E., Yasmeen F. (2024), *forecast: Forecasting functions for time series and linear models. R package version 8.22.0*, https://pkg.robjhyndman.com/forecast/ [accessed: 4.03.2024].

Li X., Pan B., Law R., Huang X . (2017), *Forecasting tourism demand with composite search index*, "Tourism Management", vol. 59, pp. 57–66.

Mellon J. (2014), *Internet search data and issue salience: The properties of Google Trends as a measure of issue salience*, "Journal of Elections, Public Opinion & Parties", vol. 24(1), pp. 45–72.

Saegner T., Austys D. (2022), *Forecasting and surveillance of COVID–19 spread using Google trends: literature review*, "International Journal of Environmental Research and Public Health", vol. 19(19), 12394.

Vosen S., Schmidt T. (2011), *Forecasting private consumption: survey-based indicators vs. Google trends*, "Journal of Forecasting", vol. 30(6), pp. 565–578.

Yu L., Zhao Y., Tang L., Yang Z. (2019), *Online big data-driven oil consumption forecasting with Google trends*, "International Journal of Forecasting", vol. 35(1), pp. 213–223.

Zhang W., Wang P. (2020), *Investor attention and the pricing of cryptocurrency market*, "Evolutionary and Institutional Economics Review", vol. 17, pp. 445–468.

Zastosowanie Google Trends jako źródła danych w modelach statystycznych

Streszczenie: Wraz z postępem technologicznym rośnie liczba potencjalnych źródeł danych, które mogą stanowić alternatywę dla tradycyjnych badań ankietowych. Przykładem tego mogą być dane o popularności wyszukiwań, udostępniane w czasie rzeczywistym za pośrednictwem Google Trends. Dane tego typu pozwalają na badanie zachowań, postaw w społeczeństwie i opinii publicznej czy prognozowanie zjawisk ekonomicznych.

Zaletą wykorzystania danych o popularności wyszukiwań jest natychmiastowy czas i niski koszt ich pozyskania. Nie bez znaczenia jest też fakt, że Google Trends pozwala na bezpośrednie badanie zachowań użytkowników internetu, a nie jedynie ich deklaracji jak w przypadku ankiety. Może to mieć znaczenie, jeżeli ankietowani uważają którąś z odpowiedzi za bardziej moralnie słuszną. Korzystanie z Google Trends wymaga jednak trafnego dobrania uwzględnianych w badaniu wyszukiwań oraz świadomości ograniczenia próby badawczej do użytkowników wyszukiwarki Google. W ramach artykułu zaprezentowano wady i zalety Google Trends oraz zweryfikowano przydatność tego źródła danych, w szczególności w okresach zwiększonej zmienności na rynkach.

Słowa kluczowe: Google Trends, analizy statystyczne, prognozowanie

OPEN ACCESS	© by the author, licensee University of Lodz – Lodz University Press, Lodz, Poland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license CC-BY (https://creativecommons.org/licenses/by/4.0/)
	Received: 2024-03-04; revised: 2024-06-25. Accepted: 2024-10-01
C O P E Member since 2018 JM13703	This journal adheres to the COPE's Core Practices https://publicationethics.org/core-practices