Experimental Evaluation of Scale, and Patterns of Systematic Inconsistencies in Google Trends Data

Philipp Behnen[1], Rene Kessler[2], Felix Kruse[2], Jorge Marx Gómez[2], Jan Schoenmakers[1], and Sergej Zerr[3]

[1] HASE & IGEL GmbH, Oldenburg Germany
{jan.schoenmakers,philipp.behnen}@haseundigel.com
[2] VLBA, Oldenburg University, Germany
{rene.kessler,felix.kruse,jorge.marx.gomez}@uni-oldenburg.de
[3] L3S Research Center, Leibniz University Hannover, Germany
szerr@l3s.de

Abstract. Search analytics and trends data is widely used by media, politicians, economists, and scientists in various decision-making processes. The data providers often use sampling when calculating the request results, due to the huge data volume that would need to be processed otherwise. The representativity of such samples is typically assured by the providers. Often, limited or no information about the reliability and validity of the service or the sampling confidence are provided by the services and, as a consequence, the data quality has to be assured by the users themselves, before using it for further analysis.
In this paper, we develop an experimental setup to estimate and measure possible variation in service results for the example of Google Trends. Our work demonstrates that the inconsistencies in Google Trends Data and the resulting contradictions in analyses and predictions are systematic and particularly large when analyzing timespans of eights months or less. In our experiments, the representativity claimed by the service was disproved in many cases. We found that beyond search volume and timespan, there are additional factors for the deviations that can only be explained by Google itself. When working with Google Trends data, users must be aware of the marked risks associated with the inconsistencies in the samples.

Keywords: Google Trends · Service Reliability · Sampling

1 Introduction

Web applications and services are being developed and extensively used around the globe since the start of the Internet revolution at the end of the last century. The user interaction data with those services has been turned into a valuable source of information not only for improving the services themselves but also for third party market analytics. Media, politicians, economists, and scientists are

widely using search analytics in various decision-making processes. Such data is provided by big web companies with a large number of users, with analysts relying on the quality of the services assured by the providers. On the example of Google Trends, our work shows that caution and careful pre-processing are required when using the data in the decision-making processes.

Since its introduction in 2006, Google Trends [4](GT) service has established itself as a tool for investigation, research and forecasting with a broad range of use cases ranging from forecasting epidemics [16], to indicating movements in the stock market [9] or identifying consumer trends and demand [17]. Inconsistencies in GT data may have considerable implications because of the service's regular and widespread use in politics, journalism, economy, and science. Especially during the corona crisis various media are using these resources, basing their research and reports on data from GT[5]. For this service, Google is using ad-hoc samples from the total of searches in its database and assures that the sample sizes are sufficiently large for the data to be representative for all searches on Google.

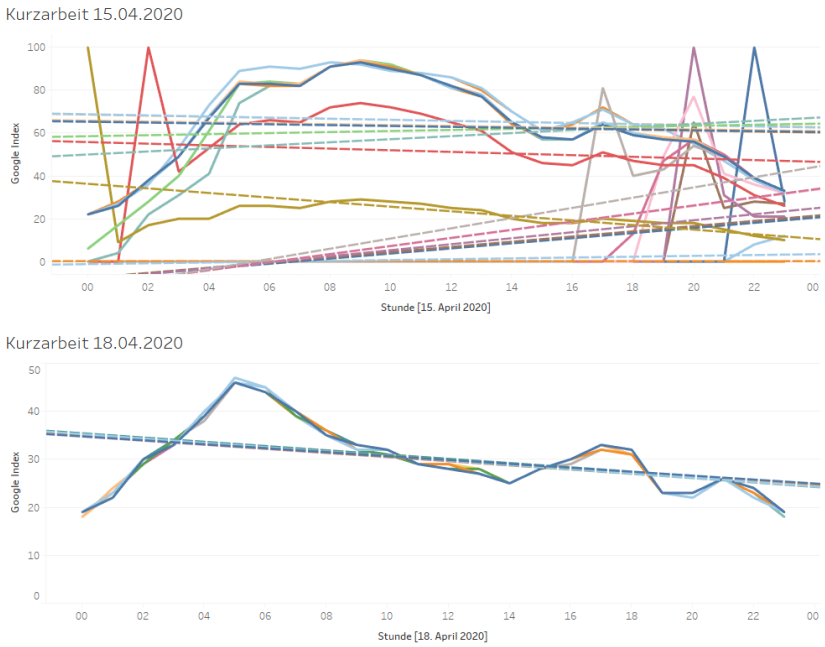

Fig. 1: Google scores and linear regression for "kurzarbeit" for April 15 and 18, 2020. Every request has its own regression.

[4] https://trends.google.de

[5] https://www.finanznachrichten.de/nachrichten-2020-04/49427722-google-trends-as-a-proxy-for-covid-19-incidence-and-deaths-378.htm (accessed 17.07.2020)

During the computation of trends using Google data, however, we observed large deviations between results of identical API requests[6]. While, for example, the values for the (German) term "kurzarbeit" for April 15, 2020, differ so wildly by the time the request was initiated that they sometimes show completely contradictory trends, there is hardly any contradiction in the data for the same search term for April 18 as displayed in Figure 1. Upon our contacting them, Google posited that such deviations may happen, but should only occur for requests with small search volumes and should be marginal. We tested these hypotheses. This paper aims to provide first insights into our investigation of the observed inconsistencies, with a focus on implications for practitioners using GT data.

2 Related Work

Search analytics and especially GT data have been widely used in research, including computer science, sociology, economics, and medicine. Especially in areas missing official statistics on some subjects, GT is frequently used as a proxy [1]. Whether the conclusions based on GT are sound, stands and falls with the credibility of the data provided.

The research has a high societal impact. Aguilera et al. [1] employed GT to access interest in burnout and models based on GT data outperformed traditional autoregressive approaches in forecasting touristic demand. [7]. GT was reported as a useful tool to acquire evidence for social hierarchy impacting income inequality and racial bias by Connor et al. [5].

In the domain of economics, Xu et al. [18] employed GT as a proxy for event impact to link to US macroeconomic variables. There is a body of work on applicability of GT for monitoring and forecasting of stock markets [8, 14, 2], new products [4] and cryptocurrencies [15] development.

In the medical domain, search statistics gained especial popularity as official health statistics are often not available for some geographic regions, however, the users tend to develop a certain level of trust for "Dr.Google" [13]. GT was adopted to monitor search interest in epilepsy surgery [11], for monitoring and forecasting deceases outbreaks [3], in particular influenza [12], respiratory syncytial virus [6] and, recently COVID19 [10].

Some criticism concerning GT was reported in research, reflecting anticipated challenges when using it as a data source. This includes the obscure score calculation, irregularly missing data [6], and the fact that user properties behind Google searches can not be identified [5]. Often, Goggle Trends tends to underestimate the real value of observation when the general public has poor knowledge of a given term. For example, the timely popularity of diseases and regional media coverage has more impact on the index as their real spread [3].

[6] There is a discussion thread at Google support: https://support.google.com/google-ads/thread/8389370?msgid=26184434 (accessed 17.07.2020)

Recent studies agree that the services should be used only to estimate the public interest for a particular keyword. To the best of our knowledge, none of the works mentioned above report any pre-processing, or data cleaning steps when using GT. In our work, we do the first step towards the systematic evaluation of the data quality provided.

3 Evaluation Setup

In this section, we provide a setup to a) evaluate the overall reliability of GT services, b) access the correlation between the reliability and search volume, and c) evaluate the representativeness of data samples provided by Google.

3.1 GT Service

GT is a service provided by Google free of charge, which can be used to extract time series of index values indicating the search intensity and trends for freely selectable keywords and topics worldwide. The analytics timespan can be chosen at liberty from between a few hours or days up to a time series from the year 2004 until today. It is also possible to filter for specific countries or regions. A single data point is a score that reflects the search popularity of the keyword, compared to the total amount of searches in the same region and timespan. The index score ranges from 0 to 100 with 100 being the data point with the highest search intensity for the selected keyword within the selected timespan. The aggregation granularity level is defined by the service as one of hour/day/week/month, depending on the length of the requested period. The data can be displayed as a graph and exported as a CSV file from a web dashboard.

Table 1: A motivated choice of keywords (German) that were used in our experiments

Keyword	**Search VolumeØ 01.2015-03.2020**	**Reason**
dachdecker	39,800 / 40,500	Medium sized, relatively constant long-term demand with some seasonal peaks.
kurzarbeit	20,900 / 27,100	Single big peak during corona crisis. Before there was a small search volume, after the peak it was medium sized.
sofa	197,000 / 201,000	Volume increases constantly and on a long-term basis. High volume overall with regular seasonal fluctuations.

3.2 Data Acquisition

We systematically retrieved trends data for our experiments, repeating for the same keywords and timespans over and over again for several weeks while limiting the region to Germany. We excluded empty request results (not enough data available) form the analysis.
GT does not provide information about the search volume. To test Google's claim that fluctuations are limited to low-volume keywords, we employed Google Ads data as a proxy. To this end, we used the tool "KWFinder" by Slovakian company Mangools which is well-established on the European market.[7]
We limited our research to German keywords (see Table 1). The keywords used in this report are "dachdecker" ("roofer"), "kurzarbeit" ("short-time work") and "sofa", as they show a range in search volume and volatility. Depending on the length of the timespan, Google automatically aggregates the data on an hourly, daily, weekly, or monthly basis. For the 16 timespans analyzed, the granularity of data is shown in Table 2. We executed requests at different times of day, with varying time intervals between requests and on different days of the week.

3.3 Evaluation Framework

Reliability: To measure the deviation within the GT results, we employ the standard deviation of the scores obtained from different samples of the same query(requested keyword, time period and geographic area), executed ad different times. We additionally employed relative standard deviation (the percentage of deviation from the mean) to make values between different keywords and time spans comparable.
Correlation: According to Google, small deviations may occur for keywords with low search volumes. We use the Spearman correlation coefficient to test this claim, as we are interested in rank correlation. We additionally employ R^2 to measure the amount of explainable variance.
Representativeness: Google emphasizes that the samples used in GT are representative. A representative sample is one that accurately represents and reflects the underlying data distribution. Thus, any two independent representative samples drawn from the same population will not significantly differ. To check the overall representativeness, we employed a Mann-Whitney-U test to measure the proportion of pairs of samples (from all available) that is coming from the same distribution.

[7] https://mangools.com/blog/kwfinder-top-questions/ A direct retrieval from Google Ads was not possible for us since Google only provides very rough figures such as "10,000-100,000" by default - only larger advertisers receive more precise data.

Table 2: Used timespans and granularity of data supplied by GT

Timespan	**Granularity of GT data**
01/2010 - 04/2020	monthly
01/2015 - 04/2020	monthly
01/2019 - 04/2020	weekly
01/2020 - 04/2020	daily
Q1, Q2, Q3 und Q4 (2019)	daily
January, February, and March (2020)	daily
15.04, 16.04, 17.04, 18.04 and 19.04 (2020)	hourly

4 Experimental Results

4.1 Descriptive Figures

Figure 2 shows an example of retrieved data. Every column represents the index value as supplied by GT for a single request (keyword "kurzarbeit" – "short-time work" for the timespan of January 2020 and region Germany). Every row is expected to contain only slightly varying or even the same values. In our examples, however, we observe large variations. To further identify patterns in observed in-

Kurzarbeit

Date	2020-04-25 11:15:15	2020-04-25 12:17:44	2020-04-25 22:36:05	2020-04-26 16:23:18	2020-04-26 17:30:05	2020-04-26 21:04:46	2020-04-26 21:34:43
01.01.20	14	16	23	0	0	9	9
02.01.20	6	14	27	14	12	18	18
03.01.20	12	14	23	11	12	21	21
04.01.20	6	11	14	9	6	11	11
05.01.20	6	11	17	29	0	15	15
06.01.20	50	49	37	21	12	27	27
07.01.20	44	55	59	69	69	88	88
08.01.20	50	49	50	50	64	57	57
09.01.20	39	56	40	55	23	51	51
10.01.20	29	47	51	27	42	56	56
11.01.20	0	8	18	18	27	19	19

Fig. 2: Heatmap for 7 requests for the keyword "kurzarbeit" with the timespan of January 2020, limited to the first 11 days to provide a short overview.

consistencies, we examined the values for the different keywords (Table 1) and timespans (Table 2). For the shortest timespans - five consecutive days from the April 15 to April 19 with values aggregated on an hourly basis - no clear patterns can be observed (Figure 4). For the keyword "sofa", which has the highest average search volume, GT returned no values for the vast majority of requests from the 15th to the 18th of April[8]. For April 19, however, a sufficient amount of data points could be retrieved.

[8] For April 16, 17 and 18, 2020 there are 35 query results each, which have the Google index value 0 for each hour. The corresponding timespans were, therefore, not considered in the analysis.

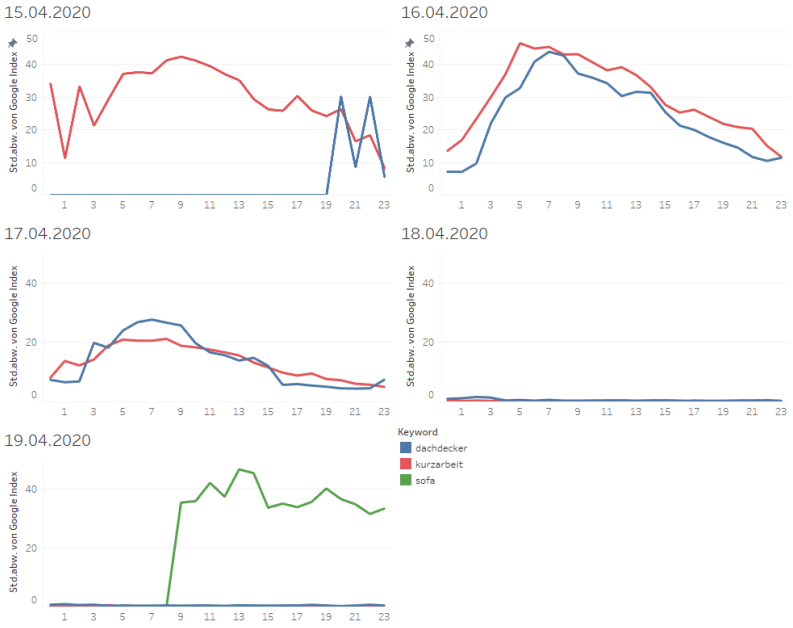

Fig. 3: Absolute standard deviations of hourly aggregated index values for the April 15. to 19., 2020

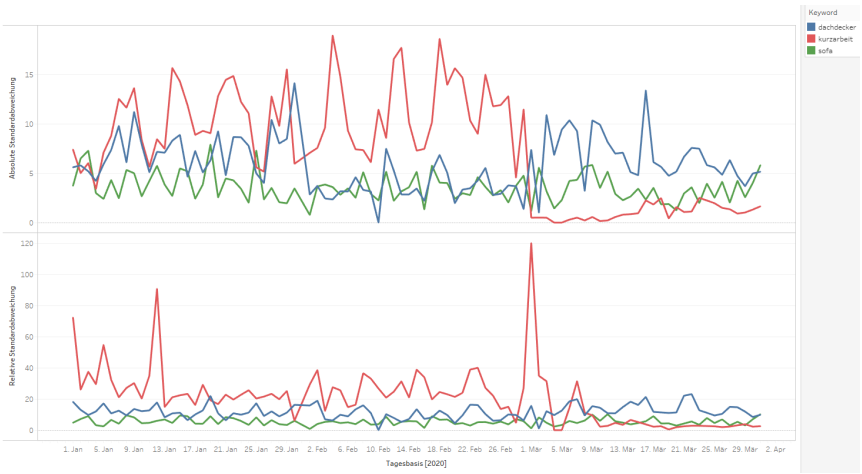

Fig. 4: Absolute standard deviations of hourly aggregated index values for the April 15 to 19, 2020

Although both other keywords ("kurzarbeit" and "dachdecker") have much smaller average search volumes, GT returned index scores for all the single days. Between April 15 and 17, those values showed a relative standard deviation of over 100%, which casts serious doubts on the reliability of the retrieved data. For April 18 and 19, however, the data is very coherent for "kurzarbeit" and "dachdecker" and seems fit for analytic use, while "sofa" displays high deviations despite being the "biggest" keyword.
For the next larger timespan, from January until March 2020, the data is automatically aggregated daily. Here, the values for the keyword "sofa" with the highest search volume is the least deviant, while the index values of the smallest keyword "kurzarbeit" diverge the most (see Figure 3). While this seems coherent at first glance, it should be noted that the difference in relative standard deviation between "sofa" and "dachdecker" is rather low, although the search volumes differ drastically. Larger search volumes seem to have some impact on the data quality ("sofa" vs. "kurzarbeit"), but do not explain all of the inconsistencies observed ("sofa" vs. "dachdecker"). Table 3 summarizes the results of the descriptive analysis. For the keywords "sofa" and "dachdecker" the hourly aggregated data (for timespans shorter than one week) shows deviations that are too high to be used in practice. Even for daily aggregated index values (timespans shorter than eight months) sometimes deviations of great size persist, depending on when the request is made. For timespans longer than 8 months (and therefore on the aggregation level of weeks or months), index values seem to be robust enough to justify practical use with some level of attention. For the smaller keyword "kurzarbeit" – which has been used by the German expert council for economy to determine the increase of short-time work in the country – the data is so deviant that it is questionable to use it at all. Although the fluctuations are often small in absolute terms, in percentage terms they can be very high, as can be seen in Table 3. Using these data may lead to different conclusions if only individual months are considered. These discrepancies can be explained by the frequent occurrence of null values. If values greater than null occur for the first time after a longer period, the relative standard deviation increases sharply.

4.2 Statistical Evaluation: Correlation with Search Volume

In this section we examine the dependence of the quality of the data on the search volume. To this end, we statistically examined the correlations between the google index (relative values), search volume (absolute values from Google Ads) and the standard deviation (as an indicator for the reliability of the data) by observing the four quarters of 2019. We repeated this calculation with the derivative (change) of the values in order to test whether data quality is affected by the fluctuations of search volume beyond their amount. For these calculations, we used the Spearman correlation measure. The resulting correlations were examined in a significance test and reported along with the proportion of the variance explained using the R^2 value. That way we can observe how strong the effect really is (see Table 4).

Table 3: Relative standard deviation for different index value aggregation levels

Keyword	**Index value aggregation level**	**Relative standard deviation range in %**	**Average relative standard deviation in %**
kurzarbeit	Hourly	0-181	38.4
sofa	Hourly	57-248	87.4
dachdecker	Hourly	0-134	33.3
kurzarbeit	Daily	0-170,8	34.2
sofa	Daily	0-13,4	5.8
dachdecker	Daily	0-36,1	12.8
kurzarbeit	Weekly	0-685,6	75.4
sofa	Weekly	0-3,7	1.7
dachdecker	Weekly	0-9,0	4.6
kurzarbeit	Monthly	0-888,8	100.9
sofa	Monthly	0-4,5	1.8
dachdecker	Monthly	0-5,5	2.4

Table 4: Correlations of the standard deviation, search volume, index volume as well as their derivatives for all keywords and Q1 until Q4 2019

(a) Correlations between the values

	Index	**STD**
STD	c= -0.49 r^2= 26.33% p<0.05	
Vol	c= 0.79 r^2= 51.39% p<0.05	c= -0.87 r^2= 53.22% p<0.05

(b) Correlations between value changes

	Index	**STD**
STD	c= 0.44 r^2= 18.16% p<0.05	
Vol	c= -0.06 r^2= 0.49% p= 0.69	c= -0.20 r^2= 4.13% p= 0.23

On average, the data quality does increase with the search volume. However, this only accounts for approximately a quarter (for the index value), respectively a half (absolute search volume) of the inconsistencies. Additionally, we found that changes in search volume positively correlate with the standard deviation, meaning that the discrepancy is higher when the search volume changes. This effect accounts for approx. $\frac{1}{5}$ of the inconsistencies. This clearly shows that Google's explaining the contradictions in GT data with low search volumes is not wrong in principle, but also far from sufficient: depending on which variable is used as a reference, a half to almost three-quarters of the discrepancies in GT data cannot be explained by the search volume of the respective search terms. Our analyses also reveal that the GT index is a very limited indicator for actual changes in search volume, since it is calculated concerning the total number of Google searches at any given time, and since this number seem to fluctuate massively, changes in actual search volume explain only half of the changes in the Google Index value. In fact, for 2 of the 3 terms examined, there was no

significant correlation between the development of the Google index value and the absolute search volume - at least judging from the data available to us.
Lacking further indications from Google on how the total number of searches has developed in the timespan observed, one should have fundamental doubts about the GT index's expressiveness.

4.3 Statistical Evaluation: Pairwise Representativeness

In this section, we analyze the pairwise differences between request results for the same search at different times to check the sample representativity. To this end, we employed the Mann-Whitney-U-Test, as we cannot assume a normal data distribution.
Depending on the timespan and keyword, up to 35 % of the compared sample pairs failed the test - and thus the requests cannot be considered representative. On an hourly level (single day timespan), a quarter of all samples fail the test, while on a daily level (timespan of weeks or months), the risk of being shown a non-representative sample remains just as high. From the weekly aggregation of the data (timespan of 8 months and longer), all samples for "sofa" and "kurzarbeit" pass the test and can, therefore, be considered representative. For "dachdecker" however, the proportion of non-representative samples is highest at the weekly level (35%) and falls significantly at the monthly level only (e.g. for periods of at least five years), although even on this scale still more than one in seven samples are not representative.

5 Conclusions and Outlook

In this work, we took a first step towards a comprehensive, systematic analysis of data retrieved from Google Trends. Our experiments show that GT data is risky to use for analysis and forecasts: requests for the same term and period at different times can return very different results. The discrepancies can be unexpectedly large and question the representativity of the samples. The patterns behind these contradictions in GT data are complex and can not be explained with insufficient search volume alone.
Consequently, analyses based on the data provided by the service should not be used in the decision-making process without careful pre-processing. In future work, we plan to analyze the deviation patterns further and include other publicly available web services.

6 Acknowledgements

This work is partly funded by the European Research Council under grant agreement 833635 (ROXANNE) and 832921(MIRROR) and by the Lower Saxony Ministry of Science and Culture under grant number ZN3492 within the Lower Saxony "Vorab" of the Volkswagen Foundation, supported by the Center for Digital Innovations (ZDIN).

References

1. Aguilera, A.M., Fortuna, F., Escabias, M., Di Battista, T.: Assessing social interest in burnout using google trends data. Social Indicators Research 2019
2. Alsmadi, I., Al-Abdullah, M., Alsmadi, H.: Popular search terms and stock price prediction. In: Big Data'19
3. Cervellin, G., Comelli, I., Lippi, G.: Is google trends a reliable tool for digital epidemiology? insights from different clinical settings. Journal of epidemiology and global health (2017)
4. Chumnumpan, P., Shi, X.: Understanding new products' market performance using google trends. Australasian Marketing Journal (2019)
5. Connor, P., Sarafidis, V., Zyphur, M.J., Keltner, D., Chen, S.: Income inequality and white-on-black racial bias in the united states: Evidence from project implicit and google trends. Psychological science (2019)
6. Crowson, M.G., Witsell, D., Eskander, A.: Using google trends to predict pediatric respiratory syncytial virus encounters at a major health care system. JMC'20
7. Höpken, W., Eberle, T., Fuchs, M., Lexhagen, M.: Google trends data for analysing tourists' online search behaviour and improving demand forecasting: the case of åre, sweden. Information Technology & Tourism 2019
8. Hu, H., Tang, L., Zhang, S., Wang, H.: Predicting the direction of stock markets using optimized neural networks with google trends. Neurocomputing (2018)
9. Huang, M.Y., Rojas, R.R., Convery, P.D.: Forecasting stock market movements using google trend searches. Empirical Economics'19
10. Husnayain, A., Fuad, A., Su, E.C.Y.: Applications of google search trends for risk communication in infectious disease management: A case study of covid-19 outbreak in taiwan. International Journal of Infectious Diseases (2020)
11. Kinney, M.O., Brigo, F.: What can google trends and wikipedia-pageview analysis tell us about the landscape of epilepsy surgery over time? Epilepsy & Behavior'20
12. Kondo, K., Ishikawa, A., Kimura, M.: Sequence to sequence with attention for influenza prevalence prediction using google trends. In: ICCBB'19
13. Lee, K., Hoti, K., Hughes, J.D., Emmerton, L.: Dr google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. JMIR'14
14. Preis, T., Moat, H.S., Stanley, H.E.: Quantifying trading behavior in financial markets using google trends. Nature Scientific reports (2013)
15. Smuts, N.: What drives cryptocurrency prices? an investigation of google trends and telegram sentiment. ACM SIGMETRICS'19
16. Verma, M., Kishore, K., Kumar, M., Sondh, A.R., Aggarwal, G., Kathirvel, S.: Google search trends predicting disease outbreaks: An analysis from india. Healthcare informatics research 2018
17. Vosen, S., Schmidt, T.: Forecasting private consumption: survey-based indicators vs. google trends
18. Xu, Q., Bo, Z., Jiang, C., Liu, Y.: Does google search index really help predicting stock market volatility? evidence from a modified mixed data sampling model on volatility. Knowledge-Based Systems 2019