4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15093

Simulating the inconsistencies of Google Trends data

Eduardo Cebrián, Josep Domenech
Department of Economics and Social Sciences, Universitat Politècnica de València, Spain.

Abstract

Google Trends (GT) allows users to obtain reports on the evolution of the popularity of searches made through the Google Search engine. Its main output is the Search Volume Index (SVI), a relative measure of the popularity of a term, which is computed using a sample of the searches. Due to the sampling error, the reports are not completely consistent, as the same query produces different time series that can widely change from day to day. This paper simulates the process of generating the SVI time series in the same way as GT does. By doing this, it has been shown that the sampling error could be an important issue if the popularity of the term under study is relatively low. Averaging multiple extractions from GT can only partially alleviate this.

Keywords: *Google Trends, Consistency, Measurement Error, Online Data*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

Google Trends (GT) is a freely available tool developed by Google that allows users to obtain reports of the evolution of the popularity of searchers made through the Google Search engine. In the last decade, GT has become popular in the scientific literature because its reports can be used to measure the population's interest on any topic. Moreover, this data can be easily accessed and is constantly updated.

The main output of GT reports are time series data representing the Search Volume Index (SVI), a relative measure of the popularity of a term. To compute the SVI, Google does not consider the whole set of searches they received in a given time period, but a sample with unknown characteristics. Due to the sampling error, the reports are not completely consistent, as the same query can produce different time series which change from day to day (Choi and Varian, 2012).The importance of these inconsistencies is often minimized (Choi and Varian, 2012; Dilmaghani, 2019) although Cebrián and Domenech (2022) report that variations in GT data may be significant enough to hinder the interpretability and reproducibility of the models estimated with them.

To understand the inconsistencies of GT data, this paper proposes a simulation model of the GT data generating process. This model is then used to analyze how a typical time series with seasonality is distorted due to the sampling process and how the averaging of extractions can mitigate the error, but only partially.

2. Related Work

The inconsistencies of GT time series have long been described, although most researchers do not consider them relevant enough to affect their results (Choi and Varian, 2012; Preis et al., 2013). Other research works identify these inconsistencies as an important source of error and average multiple GT requests of the same time series on different days, or using some tricks to force a new sample. This way, the time series are smoothed, thus reducing the sampling error.

However, the number of extractions which are averaged widely vary across the literature. On the one hand, D'Amuri and Marcucci (2017) take 24 different extractions for the search term "jobs" and report cross-correlations of at least 0.99 between extractions. On the other hand, Cebrián and Domenech (2022) extract queries related to Austrian cities on 6 different occasions and find correlations between 0.79 and 0.94, while Carrière-Swallow and Labbé (2013) use the average standard deviation to measure the sampling error and report values above 15% for the term "Chevrolet" after 50 extractions.

To the best of our knowledge, there is no method for determining how many extractions should be averaged to alleviate the sampling error, or which factors may affect it. To

understand the intricacies of how the SVIs are produced and the effects of averaging multiple extractions of the same GT time series, this paper proposes a simulation model to generate the SVIs (and its sampling error) in the same way as Google Trends does.

3. Google Trends sampling

The process to compute SVIs is illustrated in Figure 1. It starts with the whole set of searches that Google has received from 2004. From this set (Total Searches), GT draws a random sample that is replaced over time. This introduces an unknown sampling error because the parameters of this sampling, such as the coverage or how often the sample is replaced, are not disclosed by Google. When a user requests the GT report for a given term and time period, the sample is filtered to keep only those rows matching the request so that frequencies by time period can be computed (Raw Popularity). Finally, the time series are normalized by setting the SVI in the period with highest frequency to 100 and scaling the frequencies in other periods proportionally (and rounding them to integers).

The sampling error introduced in GT reports is also illustrated in Figure 1. In the example provided, the term a is reported with an SVI of (0, 100, 33), but if it were computed with the Total Searches set, the result should have been (67, 100, 67).

4. Simulation and Results

This section provides some simulations of the GT process described in Section 3 to check how the SVI time series change depending on the popularity of the search term and what the effect of averaging multiple extractions is.

For illustrative reasons, it has been assumed that the total number of searches of the term *y* follows a function with linear trend and a seasonal component, modeled with a sinusoidal function with a period of 12 time units, as defined in Equation 1.

$$Y_t = \beta_1 t + \beta_2 \sin\left(\frac{2\pi}{12} t\right) \quad (1)$$

where Y_t is the number of searches of the term *y* at time *t*, β_1 is the parameter defining the strength of the linear trend, and β_2 defines the strength of the seasonal component.

For each time period t, the presence of the term *y* in the sample (y_t) follows a binomial distribution with parameters *n* equals the sample size, and *p* equals the proportion of searches of term *y* among all the searches received by Google at that time period.

Therefore, the expected number of occurrences in the sample of term *y* at time t is:

$$E[Y_t] = n * p_t \quad (2)$$

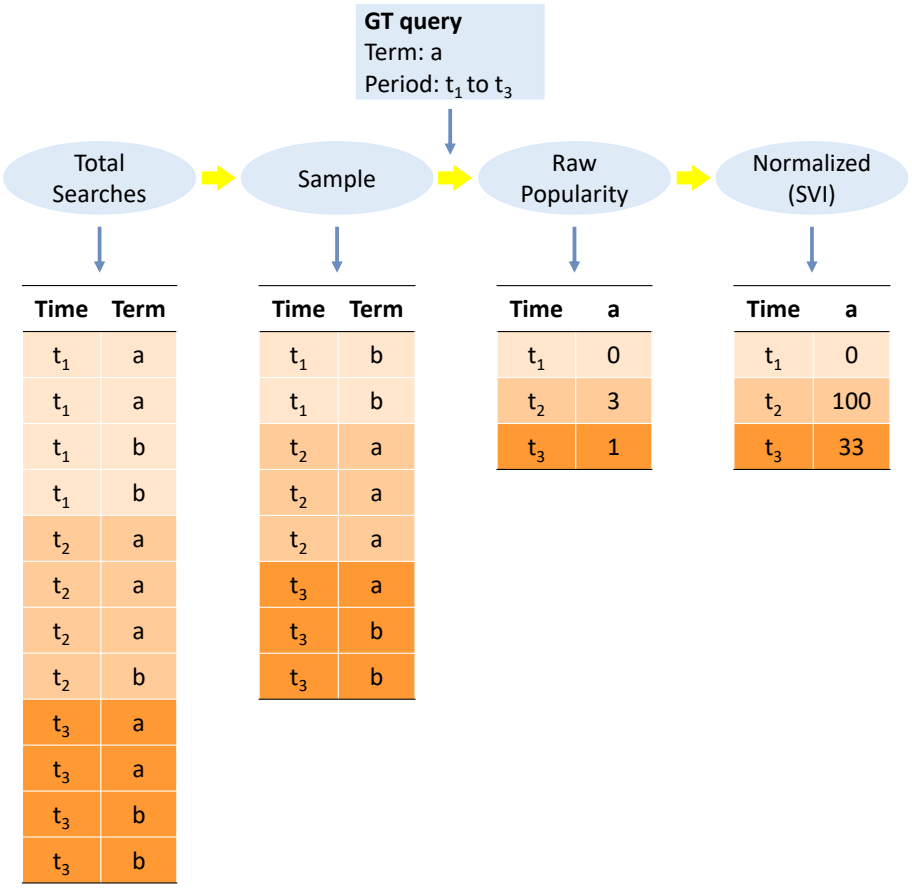

Figure 1: Process GT follows to compute an SVI time series. It is illustrated with an example with three time periods (t_1, t_2 and t_3) and two terms (a and b). Only the GT report for term a is requested.

Notice that p_t varies in time, being this variation the change in popularity of the term.

Since n and p_t are unknown (as they are not disclosed by Google), we have studied the SVIs of two terms with different popularity. Term H has an average frequency in the sample of 200 times, while term L is less popular and has an average frequency in the sample of 20 times through all the considered periods. Simulations are conducted considering GT requests for 60 periods. The random process of generating the SVI for each term has been repeated 20 times, each one representing one extraction from GT.

Figure 2 shows the simulation results of 1 (left), 10 (center), and 20 (right) SVI extractions. Light blue lines represent each individual extraction, while dark blue lines illustrate the

average of all the extractions in each plot. Plots in the top row refer to the most popular term (H), while plots in the lower row refer to the less popular term (L). Each plot includes the Pearson's correlation coefficient (r) of the time series in dark blue with the actual popularity of the term (defined by Equation 1).

Plots in the left part of Figure 2 evidence that a single extraction has significant noise. This noise, which is introduced by the sampling process, can be alleviated by averaging multiple extractions. After averaging 10 extractions, the curve for the term with high popularity (H) is smoother, as evidenced by the $r = 0.997$ value as well. However, the less popular term (L) requires more extractions to obtain a good approximation to the actual trend. Indeed, after averaging 20 extractions of the less popular term, the correlation coefficient r is still below the one of 10 extractions of the high popular term.

These simulation results highlight the relationship between the sharpness of the SVI and the absolute popularity of the term and, therefore, the need for averaging more GT extractions when studying less popular terms. However, as one can observe in Figure 2, there is a side effect related to the construction of a time series as the average of a number of extractions: the range of the SVI is reduced. In the case of the less popular term (L), the SVI takes values from 18 to 100 in a single extraction (bottom-left plot). When the average of 20 extractions is considered, SVI ranges from 27.1 to 81.3 (bottom-right plot). This implies that the value of a single extraction (for instance, when using GT for nowcasting purposes) cannot be directly compared to the series obtained after averaging multiple extractions.

5. Conclusions

Although Google Trends has become a very popular data source among researchers, its sampling error has not been intensively studied. This paper has replicated the process of generating the SVI time series in the same way as GT does. By doing this, it has been shown that the sampling error could be an important issue if the popularity of the term under study is relatively low, as the quantity of noise it introduces in the series is noticeable.

The technique of extracting GT data multiple times and using the average of the series has also been studied. Our results showed that it certainly alleviates some of the variability introduced in the random sampling, but the number of repetitions needed to smooth the curve heavily depends on the absolute popularity of the term. Moreover, this procedure changes the range and scale of the SVI time series, thus increasing the complexity of using GT data for nowcasting and forecasting, as additional transformations should be considered.

Acknowledgments

This work was partially supported by grants PID2019-107765RB-I00 and funded by MCIN/AEI/10.13039/501100011033.

References

Barreira, N., Godinho, P., & Melo, P. (2013). Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends. *NETNOMICS: Economic Research and Electronic Networking*, 14(3), 129–165.

Borup, D., Christian, E., and Schütte, M. (2022). In search of a job: Forecasting employment growth using Google Trends. *Journal of Business & Economic Statistics*, 40(1), 186–200.

Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with google trends in an emerging market. *Journal of Forecasting*, 32(4), 289–298.

Cebrián, E., & Domenech, J. (2022). Is google trends a quality data source? *Applied Economics Letters*, pages 1–5.

Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. *Google Inc*, 1, 1–5.

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88(s1), 2–9.

Dilmaghani, M. (2019). Workopolis or The Pirate Bay: what does Google Trends say about the unemployment rate? *Journal of Economic Studies*.

D'Amuri, F., & Marcucci, J. (2017). The predictive power of google searches in forecasting US unemployment. *International Journal of Forecasting*, 33(4), 801–816.

Preis, T., Moat, H. S., & Stanley, H. E. (2019). Quantifying trading behavior in financial markets using google trends. *Scientific Reports*, 3(1), 1684.

Saxa, B. (2015). Forecasting mortgages: internet search data as a proxy for mortgage credit demand. *National Bank of the Republic of Macedonia*, 107.

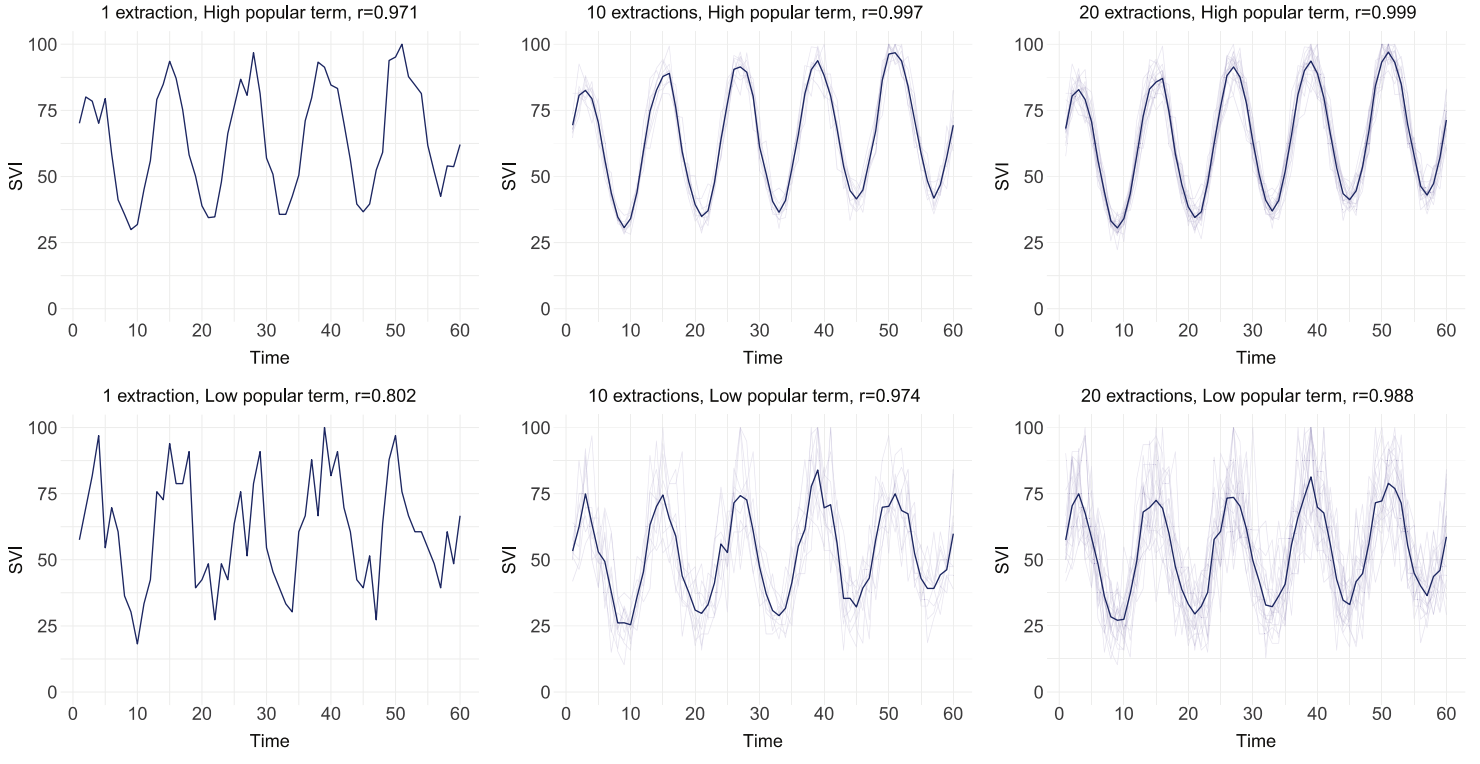

Figure 2: Simulation of GT data generation process of a search with a 12-period seasonality. Each individual extraction is shown in light blue. The dark blue line represents the average of all extractions in each plot.