Google Trends search and the coronavirus contagion

Lincon Bilibio[1], Rodolfo Coelho Prates[2]

[1]Researcher at Pontifical Catholic University of Paraná, Curitiba, PR, Brazil.
[2]Professor at Graduate Program in Health and Environment, University of Joinville Region – UNIVILLE, Joinville, SC, Brazil.

Abstract— *Since December 2019, the world has been facing the advance of coronavirus. As of early April, over two million people have been infected and almost 80 thousand died.As we are dealing with a pandemic, it is expected that people seek information on different themes related to the disease. One of the most usual ways to seek information is through the Google search engine. Using time series econometric procedures, the aim of this paper is to analyze the existence of causality between the number of Google searches on the word coronavirus and the number of cases of infected people.The results we found show that the spread of the disease Granger-causes Google searches on coronavirus. On the other hand, Google searches do not impact the number of infected people.*

***Keywords*— coronavirus, google trends, causality, time series.**

I. INTRODUCTION

Since December 2019, the world has been facing the advance of coronavirus. As of early April, over two million people have been infected and almost 80 thousand died. These numbers show that the disease has both a high speed of transmission and high lethality. During this short time, the reaction of governments and people is extremely varied. Countries like China and South Korea, for example, have taken urgent measures to contain the spread of the disease. On the other hand, countries like Italy and Spain were less cautious, leading to a higher number of infected and deceased people. A World Health Organization report from April 1st reveals that every country has at least one case of coronavirus infection. Therefore, we are facing a disease that has quickly spread throughout the world.

As we are dealing with a pandemic, the media—through its many outlets—has been dedicating countless news and feature stories on coronavirus, whose purposes are varied. One such purpose is to track the evolution of the disease in different countries. A second one is to analyze public policies that countries have been implementing to fight the disease. A third is related to the economic impacts of the disease. Beyond these, the media also has the duty of raising awareness for people to adopt measures to avoid infection.

In this complex scenario, it is expected that people seek information on different themes related to the disease. One of the most usual ways to seek information is through the Google search engine. Besides offering websites on search results, Google also keeps a record of all the search terms and makes that data available through Google Trends, where it is possible to verify how much a particular term has been searched on a scale of 0 to 100.

Employing Google Trends in research is becoming increasingly popular. Jun, Yoo, and Choi (2018), for instance, have analyzed 657 research papers that made use of Google Trends. These authors indicated that the research themes are quite broad, including information technology, health, medicine, and economics.

Althouse, Ng, and Cummings (2011) used Google Trends to predict dengue incidence in Singapore and Bangkok. Butler (2013) compared the growth of influenza levels with influenza-related searches on Google Trends. Tkachenko et al. (2017) noted that it is possible to improve surveillance of Type 2 diabetes with Google Trends.

However, Carvellin, Comelli, and Lippi (2017) showed that Google Trends is less reliable to predict the epidemiology of relatively common illnesses with low media coverage and relatively rare illnesses. Therefore, according to the authors, illnesses that get more media coverage have a bigger impact on Google Trends.

On the specific case of coronavirus, Strzelecki and Rizun (2020) analyzed the relationship between the spread of the disease and searches on Google Trends, and noted that Google Trends had predicted the contagion worldwide. Hu et al. (2020) showed a slightly positive correlation between Covid-19 on Google Trends and the daily number of people infected with SARS-CoV-2 on most of the analyzed countries. Additionally, Husnayain, Fuad, and Su (2020) claim that Google Trends could potentially define the right time and place for practicing appropriate risk communication strategies to the coronavirus-affected population.

By making use of daily data, our goal is to analyze the existence of causality between the number of Google searches on the word *coronavirus* and the number of cases of infected people. We hypothesize that a higher number of infected people would spark an interest on the disease. On the other hand, we also hypothesize that better-informed people would be more careful and avoid getting infected.

II. DATA AND METHODOLOGY

The number of people infected by day was obtained from Our World in Data (https://ourworldindata.org/coronavirus-data). We analyzed the period from January 1st 2020 to April 16th 2020. Additionally, we downloaded data of the volume of coronavirus search queries from Google Trends. Google Trends has its own metric of data, in which the numbers are equivalent to the relevance of searches. A value of 100, for instance, is equivalent to the peak popularity of a term. A value of 50 means that the term has half that popularity. A value of 0 means that there was not enough data about the searched word.

Estimation procedures

According to Mills (2019), consider the multivariate dynamic regression model

$$\boldsymbol{y}_t = \boldsymbol{c} + \sum_{i=0}^{p} \boldsymbol{A}_i \boldsymbol{y}_{t-i} + \sum_{i=0}^{q} \boldsymbol{B}_i \boldsymbol{x}_{t-i} + \boldsymbol{u}_t \qquad (1)$$

where $\boldsymbol{y}'_t = (y_{1,t}, y_{2,t}, \dots, y_{n,t})$ and $\boldsymbol{x}'_t = (x_{1,t}, xx_{2,t}, \dots, y_{k,t})$ are vectors of endogenous and exogenous variables, respectively. $\boldsymbol{c}' = (c_1, c_2, \dots c_n)$ is a vector of constants. The terms $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are sets of $n \times n$ and $n \times k$ matrices, respectively. $\boldsymbol{u}'_t = (u_{1,t}, u_{2,t}, \dots, u_{n,t})$ is a vector of innovations or errors, whose variances and covariances are serially uncorrelated, so that $E(\boldsymbol{u}_t, \boldsymbol{u}'_s) = 0$for $t \neq s$, where 0is an $n \times n$ null matrix.

If (1) does not contain exogenous variables—that is, if the model can be written as

$$\boldsymbol{y}_t = \boldsymbol{c} + \sum_{i=0}^{p} \boldsymbol{A}_i \boldsymbol{y}_{t-i} + \boldsymbol{u}_t \qquad (2)$$

then there is simply a p th order autoregression; the dependent variable is determined by its own lag operators. An alternative and perhaps simpler way of writing (2), proposed by Greene (2003), is through matrix notation of the form

$$\boldsymbol{y}_t = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{bmatrix} \boldsymbol{y}_{t-1} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

In this case, $\boldsymbol{y}'_t = (y_{1,t}, y_{2,t})$, $\boldsymbol{c}' = (c_1, c_2)$, $\boldsymbol{A}_i = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{bmatrix}$ e $\boldsymbol{u}'_t = (u_{1,t}, u_{2,t})$.

In both (2) and (3), it is assumed that all time series are stationary. These equations are pth order vector autoregressions (in (3), $p = 1$), or $VAR(p)$ for short. In (2), the presence of nonzero off-diagonal elements in $\boldsymbol{A}_i$ implies there is a dynamic relationship between the variables. Mills (2019) notes that such a dynamic relationship is known as Granger causality, which is seen as a prediction mechanism; in other words, when $y_{r,t-i}$ values are useful to forecast y_s present values. On the other hand, if $y_{s,t-i}$ values can also forecast y_r , these two variables are said to have feedback. We can apply a Wald statistics, based on the chi squared distribution, to test if $y_{r,t-i} \rightarrow y_s$ and $y_{s,t-i} \rightarrow y_r$.

Applying VAR requires determining the lag order p empirically, usually through sequential testing. Akaike's Information Criterion (AIC), Hannan and Quinn Information Criterion (HQIC), and Schwarz’s Bayesian Information Criterion (SBIC) are alternative ways to establish lag order.

As stated previously, the time series needs to be stationary; in other words, the time series cannot contain a unit root. If a non-stationary time series can become stationary after being differenced d times, such a time series is integrated of order d. The Augmented Dickey–Fuller test (ADF) verifies whether a time series is integrated or not, according to the following hypotheses:

- Null hypothesis: a unit root is present, or the time series is non-stationary, or it has a stochastic trend;
- Alternative hypothesis: the time series is stationary with a deterministic trend.

III. RESULTS

We analyzed a total of 107 observations, from January 1st through April 16th, as seen on Table 1. During this period, the mean value of infected people was almost 19 thousand, and the maximum value nearly reached 90 thousand.

Table 1: Descriptive statistics of the variables

Variable	N. observations	Mean	Std Deviation	Minimum	Maximum
People infected	107	18,971.06	29,260.86	0	89,349
Google Trends index	107	32.76168	31.40	0	100

Source: the authors.

The Google Trends index on *coronavirus* had a mean value of 32.7, on a scale of 0 to 100. We note that a 100 value means the most searches on a single day. A 50 value means 50% of that amount.

Figure 1 shows the progression of both the number of infected people and searches on Google. Both graphs, at first glance, behave similarly. The graph on the left side of the picture, which depicts the number of infected people, presents a small increase at the outset. In early February there is a 15-thousand cases surge, followed by a drastic return to the sub-five-thousand cases baseline. We imagined this could be an error; however, other data sources presented the same surge. From late February, we can see the continuous and fast rise in the number of cases. April 11th had the highest number of infected people in a single day. After this peak value, the curve fell slightly, but did not establish a trend.

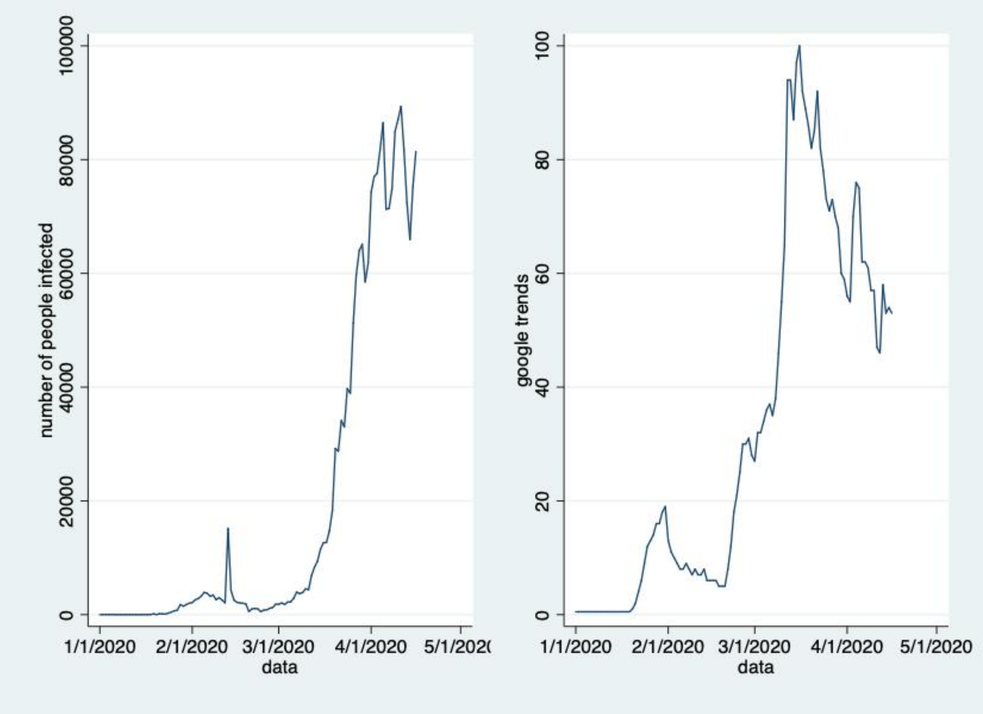

Fig.1: Number of infected people and Google Trends index

Source: the authors.

The graph on the right-hand side of the picture shows that, from the beginning of January until its third week, there were practically no Google searches on the coronavirus. From the last week of January, the number of searches increases until a peak on January 31st. Afterwards, we see a decrease until the last week of February, when the curve starts to show a clear rising trend. The maximum value occurred on March 16th and, from that day, there is a downward trend which nevertheless shows some fluctuation.

Figure 1 clearly shows that both time series are not stationary; that is, their statistical characteristics, such as mean, variance, autocorrelation, etc. are not constant throughout time. Non-stationary time series are also characterized by the presence of a unit root, show trends of growth or decline, as well as other patterns, such as seasonal cycles. Even though the graph points that the series are non-stationary, ADF test confirmation is necessary. Table 2 shows the results of that test. The second column presents the results for the original variable. The null hypothesis in the ADF means that the time series has a unit root and, therefore, is non-stationary. The values in the second column cannot reject the null hypothesis and, therefore, the series is non-stationary.

Table 2: Augmented Dickey–Fuller test for unit root

Variable	Original variable	First difference
Number of infected people	0.8340 (0.9922)	-9.189*** (0.0000)
Google Trends index	-1.007 (0.7407)	-8.020*** (0.0000)

Source: authors’ estimation.

As we discussed briefly in the methodology section, time series models require that the series be stationary, and one way of ensuring that is through differencing. If the resulting time series remains non-stationary, it should be differenced again. We differenced the time series and tested them to check if the first differencing was able to render it stationary. The third column in Table 2 shows that the values are statistically significant and, therefore, we reject the null hypothesis that the series is non-stationary.

Once the time series is rendered stationary, the next step is to estimate the model for auto regression vectors. In order to do so, we need to identify the lag order. Table 3 shows different tests for this process.

Table 3: Lag order selection according to different criteria

lag	FPE	AIC	HQIC	SBIC
0	4.6e+08	25.6331	25.6546*	25.6862*
1	4.6e+08	25.6130	25.6774	25.7723
2	4.8e+08	25.6705	25.7778	25.9359
3	4.5e+08	25.5927	25.7431	25.9643
4	4.6e+08	25.6182	25.8112	26.0958
5	4.9e+08	25.6748	25.9109	26.2587
6	4.7e+08	25.6411	25.9201	26.3312
7	4.3e+08	25.5386	25.8605	26.3349
8	3.6e+08*	25.3742*	25.7391	26.2766
9	3.7e+08	25.3956	25.8034	26.4042

*indicates lag order selected by the criterion.

FPE: Final Prediction Error

AIC: Akaike Information Criterion

HQIC: Hannan and Quinn Information Criterion

SBIC: Schwarz’s Bayesian Information Criterion

Source: authors’ estimation.

The choice of a lag order is based on the lowest value in each test. The results in Table 3 reveal that the tests are inconclusive. Both HQIC and SBIC yield a preferred model with just an intercept and no lags (lag 0). With lag 0, we suggest that the VAR model might not fit the data adequately. On the other hand, both FPE and AIC yielded a preferred lag order of 8. Therefore, we estimated our model with lag 8: VAR(8).

The results of our estimation of the VAR model are presented on the top portion of Table 4 (Panel A). The bottom portion of the table (Panel B) shows the results of the Granger causality test. Through the VAR model, we estimated two equations. On the first one, the dependent variable is the number of infected cases, and the independent variables are the lagged values of infected cases (for 8 periods) and the lagged Google Trends index(GT). In this equation, we note that only some of the lagged cases variables are statistically significant, and none of the lagged GT variables are statistically significant.

Table 4: Estimated coefficients of the VAR model and Granger causality test

Panel A - VAR Model		
D_cases	$cases_t$	GT_t
_cons	-65.13593 (-0.16)	1.070754* (1.91)
$cases_{t-1}$	-0.1767816* (-1.84)	-0.0003578*** (-2.66)
$cases_{t-2}$	-0.1358221 (-1.47)	0.0001758 (1.36)
$cases_{t-3}$	-0.1771852* (-1.94)	-0.0001217 (-0.95)
$cases_{t-4}$	0.0431984 (0.44)	-0.0001734 (-1.26)
$cases_{t-5}$	0.0514191 (0.51)	-0.0002405* (-1.71)
$cases_{t-6}$	0.3347749*** (3.65)	-0.0001951 (-1.52)
$cases_{t-7}$	0.4634997*** (4.69)	0.0000194 (0.14)
$cases_{t-8}$	0.4141888*** (3.89)	0.0002497* (1.67)
GT_{t-1}	12.84894 (0.18)	0.327297*** (3.27)
GT_{t-2}	63.63167 (0.84)	-0.2278228** (-2.14)
GT_{t-3}	-91.37502 (-1.17)	0.2947302*** (2.69)
GT_{t-4}	-5.409178 (-0.06)	-0.2142141* (-1.78)
GT_{t-5}	29.45973	0.1752193

	(0.34)	(1.46)
GT$_{t-6}$	29.04265 (0.34)	-0.0606839 (-0.51)
GT$_{t-7}$	134.5903 (1.63)	0.060728 (0.53)
GT$_{t-8}$	100.8649 (1.26)	-0.2795573** (-2.49)
Panel B - Granger causality		
cases do not Granger-cause GT	19.442 ** (0.013)	
GT does not Granger-cause cases	11.069 (0.198)	

*Significance at the 10% level.

**Significance at the 5% level.

***Significance at the 1% level.

Source: authors’ estimation.

The second estimated equation has GT as a dependent variable, and lagged GT and lagged number of cases as independent variables. For this second equation, both the lagged GT values and the lagged number of cases forecast the present behavior of the GT variable.

As for the Granger causality test (Panel B), we reject the null hypothesis that the number of cases does not cause GT at the 5% significance level. In other words, the number of cases raises awareness or sparks interests in coronavirus, and that leads people to search for information on Google.

On the other hand, we do not reject the hypothesis that GT does not cause cases. This means that searching for information on Google does not impact the number of people infected.

IV. CONCLUSION

As early discussed, coronavirus is spreading rapidly throughout the world. Up to the conclusion of this research paper, the maximum number of infected people in one day occurred on April 11th, with nearly 90 thousand new cases. After this peak, there has been a slight decrease accompanied by a lot of fluctuation.

Google searches about coronavirus have also grown, with the peak value occurring on March 16th, 26 days before the peak of new daily infected cases. We also noted that there was a stronger downward trend after the peak value on Google searches than on new coronavirus cases.

The results we found show that the spread of the disease Granger-causes Google searches on coronavirus. On the other hand, Google searches do not impact the number of infected people. Unlike the results of Strzelecki and Rizun (2020) and Husnayain, Fuad, and Su (2020), the results we obtained show that the Google Trends index is not a good predictor in the case of coronavirus, precisely because of the non-causality relationship. This might be related to the scope of the analysis, which in this case is global. However, the mismatch between the progression of the disease and Google Trends data might point to a certain degree of torpor on the part of the population, possibly due to the big exposition the coronavirus is getting in the media. Therefore, it would be more appropriate to explore specific themes, such as methods of coronavirus prevention. In this sense, a more accurate analysis of Google Trends data might define more efficient strategies of communication aimed at reducing the number of infected people.

REFERENCES

[1] Adjei, K. K., & Dei, D. J. (2015). Assessing implementation of knowledge management systems in banks, a case of Ghana. Journal of Information and Knowledge Management,5(1), 133–140

[2] Althouse, B.M., Ng, Y.Y., &Cummings, D.A. (2011). Prediction of dengue incidence using search query surveillance. PLoS Neglected Tropical Diseases, 5(8), e1258, 1-7.

[3] Butler, D., (2013). When Google got flu wrong. Nature 494, 155. Retrieved on: https://www.nature.com/news/when-google-got-flu-wrong-1.12413

[4] Cervellin, G. Comelli, I. & Lipp, G. (2017) Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. Journal of Epidemiology and Global Health, 7, 185-189.

[5] Greene, W.H. (2003).Econometric analysis. 5a edição, New York: MacMillan, 827p.

[6] Husnayain A, Fuad A, &Su EC-Yu. (2020). Applications of Google search trends for risk communication in infectious disease management: A case study of COVID-19outbreak in Taiwan. International Journal of Infectious Diseases.

[7] Jun, S., Yoo, H.S. & Choi S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. Technological Forecasting and Social Change, 130, 69-87.

[8] Mills, T. C. (2019) Applied Time Series Analysis: a practical guide to modeling and forecasting. London: Academic Press.

[9] Strzelecki, A., & Rizun, M. (2020) Infodemiological Study Using Google Trends on Coronavirus Epidemic in Wuhan, China. International Journal of Online and Biomedical Engineering, 16(4), 139-146.

[10] Tkachenko, N., Chotvijit, S., Gupta, N. (2017).Google Trends can improve surveillance of Type 2 diabetes. *Sci Rep***7,** 4993 (2017).