Vol.65: e22210648, 2022
https://doi.org/10.1590/1678-4324-2022210648
ISSN 1678-4324 Online Edition

Article - Human and Animal Health

Correlations between Web Searches and COVID-19 Epidemiological Indicators in Brazil

Marcelo Sartori Locatelli[1*]
https://orcid.org/0000-0002-0893-1446

Evandro L. T. P. Cunha[1,2]
https://orcid.org/0000-0002-5302-2946

Janaína Guiginski[1]
https://orcid.org/0000-0003-0590-4538

Ramon A. S. Franco[3]
https://orcid.org/0000-0002-2653-9835

Tereza Bernardes[1]
https://orcid.org/0000-0001-7199-3888

Pedro Loures Alzamora[1]
https://orcid.org/0000-0001-9599-0198

Daniel Victor F. da Silva[1]
https://orcid.org/0000-0001-7662-6737

Marcelo Augusto S. Ganem[1]
https://orcid.org/0000-0003-0842-4732

Thiago H. M. Santos[1]
https://orcid.org/0000-0001-6784-0002

Anne I. R. Carvalho[1]
https://orcid.org/0000-0002-6533-5229

Leandro M. V. Souza[1]
https://orcid.org/0000-0002-1150-2546

Gabriela P. F. Paixão[1]
https://orcid.org/0000-0003-1821-3912

Elisa França Chaves[1]
https://orcid.org/0000-0002-8176-586X

Guilherme Bezerra dos Santos[1]
https://orcid.org/0000-0002-1979-1522

Rafael Vinícius dos Santos[1]
https://orcid.org/0000-0002-6534-5785

Amanda Cupertino de Freitas[1]
https://orcid.org/0000-0001-8600-4077

Matheus G. Flores[1]
https://orcid.org/0000-0002-1654-5852

Rachel F. Biezuner[1]
https://orcid.org/0000-0002-9542-090X

Rodolfo Lins Cardoso[1]
https://orcid.org/0000-0002-0139-107X

Rodrigo Machado Fonseca[1]
https://orcid.org/0000-0001-6125-642X

Ana Paula Couto da Silva[1]
https://orcid.org/0000-0001-5951-3562

Wagner Meira Jr[1]
https://orcid.org/0000-0002-2614-2723

[1]Universidade Federal de Minas Gerais (UFMG), Departamento de Ciência da Computação, Belo Horizonte, Minas Gerais, Brasil; [2]Universidade Federal de Minas Gerais (UFMG), Faculdade de Letras, Belo Horizonte, Minas Gerais, Brasil; [3]Universidade Federal do Oeste da Bahia (UFOB), Centro das Ciências Exatas e das Tecnologias, Barreiras, Bahia, Brasil.

Editor-in-Chief: Paulo Vitor Farago
Associate Editor: Paulo Vitor Farago

Received: 07-Oct-2021; Accepted: 08-Nov-2021.

*Correspondence: marcelosartlocatelli@gmail.com; Tel.: +55-31-998006733 (M.S.L.).

HIGHLIGHTS

- Google Trends data could be useful for predicting COVID-19.
- High correlations (>=0.7) were found between keywords and indications when using a lag.
- ARIMAX model could help predict COVID-19 cases and deaths per week.

Abstract: COVID-19 rapidly spread across the world in an unprecedented outbreak with a massive number of infected and fatalities. The pandemic was heavily discussed and searched on the internet, which generated big amounts of data related to it. This led to the possibility of attempting to forecast coronavirus indicators using the internet data. For this study, Google Trends statistics for 124 selected search terms related to pandemic were used in an attempt to find which keywords had the best Spearman correlations with a lag, as well as a forecasting model. It was found that keywords related to coronavirus testing among some others, such as “I have contracted covid”, had high correlations (≥0.7) with few weeks of lag (≤4 weeks). Besides that, the ARIMAX model using those keywords had promising results in predicting the increase or decrease of epidemiological indicators, although it was not able to predict their exact values. Thus, we found that Google Trends data may be useful for predicting outbreaks of coronavirus a few weeks before they happen, and may be used as an auxiliary tool in monitoring and forecasting the disease in Brazil.

Keywords: Google Trends; infodemiology; epidemiological predictions; digital health.

INTRODUCTION

COVID-19 was first identified in December 2019 in Wuhan, China, quickly spreading to the rest of the world in the following months [4, 24]. In Brazil, the first case of this contagious disease was reported on the 26th of February, 2020 [23]. A month later there were already as many as 2,915 confirmed cases and 77 deaths. As of the 18th of July 2021, the numbers grew much higher, to a total of 19,376,574 cases and 542,214 deaths in the country [21]. Considering the fast growth of the disease in this part of the world, it is crucial to explore possible methods that might aid the prediction of the epidemic, in order to allow for better preparation to handle it.

Web applications, such as online social networks and search engines, have become viable tools for the massive acquisition of data and the use of methods for monitoring and predicting events [1,2]. Several studies have shown that it is possible to provide information on the epidemiological evolution of certain diseases using these data, acting in a complementary way to traditional epidemiological surveillance [3].

The rapid spread of COVID-19 [4], coupled with the significant growth of the internet and social network usage [5], gave the epidemic the distinction of being heavily discussed and searched for on this medium [6], which in turn resulted in the generation of big amounts of data related to the disease.

One of the platforms that have such relevant information is Google Search, as it is the most used and most relevant search engine in Brazil, with 97,06% of the total search engine market share as of June 2021 [7]. As it has been shown that relative search volumes of COVID-19 related search terms can be used to predict epidemiological data in other countries [8,9], it is pertinent to analyze these statistics for Brazil to study the potential uses of search data for epidemiological prediction in the country.

The aim of this study is to evaluate the temporal correlation between searches for certain terms on Google Search and variations in the epidemiological indicators of COVID-19 in Brazil. In addition to contributing to a better understanding of the web search behavior of the Brazilian population, this work aims to analyze the feasibility of using data obtained from the web to forecast peak cases and deaths in a given region, motivating a better allocation of resources during the pandemic.

MATERIAL AND METHODS

Data

Web search data

Data related to web searches were collected using the Google Trends tool (https://trends.google.com), which provides information on the volume of searches for certain terms in different regions and time periods. 124 pre-selected search terms were considered, all of them related to various topics pertinent to COVID-19, including symptoms, testing, means of prevention, vaccines, and medications.

Google Trends does not provide absolute search numbers; it provides a measure from 0 to 100 that denotes the interest on a given keyword over time [10]. A score of 100 means peak popularity for the search term on a given timeframe, while other scores are in relation to this peak, meaning a measurement of X is X% as popular. If a keyword has a score of 0, it means that there was not enough data.

Epidemiological data

The epidemiological indicators evaluated here were the numbers of new cases and new deaths per week in Brazil. Since the data from the epidemiological indicators were in a very different scale when compared to the Google Trends data, for the sake of consistency they were normalized to a range from 0 to 100, where 100 denotes the peak for that indicator. This was done by dividing the values of each measurement by its maximum value during the period analyzed and then multiplying by 100. For this study, the data pertains to the weeks from February 23, 2020 to May 8, 2021. Searching these data for different time frames may yield different results due to the nature of Google Trends data.

Correlation

We decided to use a correlation-based method to find the search terms that best predict the tendency of growth of COVID-19 in Brazil. For this, for each keyword, there was an attempt to find the number of weeks of lag between itself and the official indicators that yielded the highest correlation. For most keywords that reached a high correlation, the lag used was lower than a month, with a median of 2 weeks, which suggested a real possibility of predicting future COVID-19 related data by using Google Trends data [8].

We used the Spearman correlation method, as Pearson correlation only identifies linear relationships between data, which does not always happen in the real world. Spearman correlation can be used to measure any monotonic relationship between two variables, assigning different rank values for each [11]. It is calculated as the Pearson correlation between the rank values of the two variables:

$$r_s = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}}, \tag{1}$$

where ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables, Cov(rgx, rgy) is the covariance of the rank variables, σ rgx and σ rgy are the standard deviations of the rank variables. Its results range from -1 to 1, where a score with absolute value greater or equal to 0.7 is considered high. For the predictions, only positive highly correlated keywords were used in order to simplify the analysis and also because just one negative correlation was found.

Forecast and prediction

Given such a relationship, we also tested forecasting models, taking into account the lags aforementioned, in an attempt to find which ones could best be used for Google Trends data. The models tested were linear regression and ARIMA/ARIMAX. These models have been used in conjunction with Google Trends data in other studies [11-14]. Polynomial regression was also attempted but to no avail.

While the linear regression does not take the time factor into account, the ARIMA and ARIMAX regression do. The ARIMA model predicts time series by using: AR (auto-regressive), I (integrated), MA (moving-averages) [15]. These translate to the parameters (p,d,q) of the model. ARIMAX introduces explanatory variables, which have been shown to work with Google Trends for the prediction of other diseases, such as Zika [16]. The (p,d,q) parameters were found using the pmdarima auto_arima function for both models.

The metrics used to evaluate the performance of the models were Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

RESULTS

Lag correlations

The correlation calculations between the time series related to Google searches and the selected epidemiological indicators allowed us to identify which search terms are most related to the epidemiological data. We observed that, in some cases, the web search behavior predicts the temporal progression of the COVID-19 indicators: this is the case, for example, of the terms “taste and smell” and “covid symptoms”, whose peaks and valleys occur a few weeks before the peaks and valleys of new cases of COVID-19. This behavior, however, does not occur for all search terms related to symptoms, as evidenced by the term “tiredness”, which has a low correlation with the variation of indicators over the weeks. This demonstrates that the method employed has the potential to make predictions of the disease's behavior, but only if certain search terms are selected. Table 1 and Table 2 provide the lag correlation data of the keywords with the highest values, excluding those which are very similar, for the sake of variety, for example, as “pcr” was already in the table, “covid pcr“ was not shown. As it can be seen in Table 1 and Table 2, some terms have an extremely high correlation with the epidemiological indicators with the appropriate lag: for example, “I have contracted covid” with a correlation of 0.9 and “lockdown decree” with 0.86 in relation, respectively, to the new cases and new deaths per week.

Table 1. Lag correlation between Google Trends search terms and new cases per week. Values in bold are the ones of the optimal lag for the search term.

Weeks earlier	I have contracted covid	pcr	I have covid	covid examination	swab test
0	0.869	0.830	0.850	0.764	0.806
1	0.893	0.852	0.868	0.809	**0.830**
2	**0.906**	**0.875**	**0.871**	**0.838**	0.808
3	0.889	0.865	0.863	0.825	0.787
4	0.866	0.824	0.830	0.807	0.778
5	0.839	0.800	0.824	0.769	0.737

Table 2. Lag correlation between Google Trends search terms and new deaths per week. Values in bold are the ones of the optimal lag for the search term.

Weeks earlier	lockdown decree	covid	I have covid	oximeter	covid test
0	0.744	0.688	0.750	0.608	0.642
1	0.814	0.747	0.786	0.687	0.698
2	0.843	0.789	0.803	0.738	0.743
3	**0.860**	0.810	**0.823**	0.788	**0.767**
4	0.847	**0.829**	0.816	**0.806**	0.760
5	0.813	0.819	0.791	0.788	0.734

The high correlations found suggest the possibility of using Google Trends data to attempt to predict epidemiological indicators. In fact, the curves observed for many of the keywords with a correlation greater than 0.7 and an optimal lag of at most a month are very similar to the indicators, as it can be seen in Figure 1 and Figure 2. An interesting result was that most highly correlated search terms, such as “pcr“, “covid test”, “oximeter” and “igg reagent” were related to methods of testing for the presence of the coronavirus in an individual.

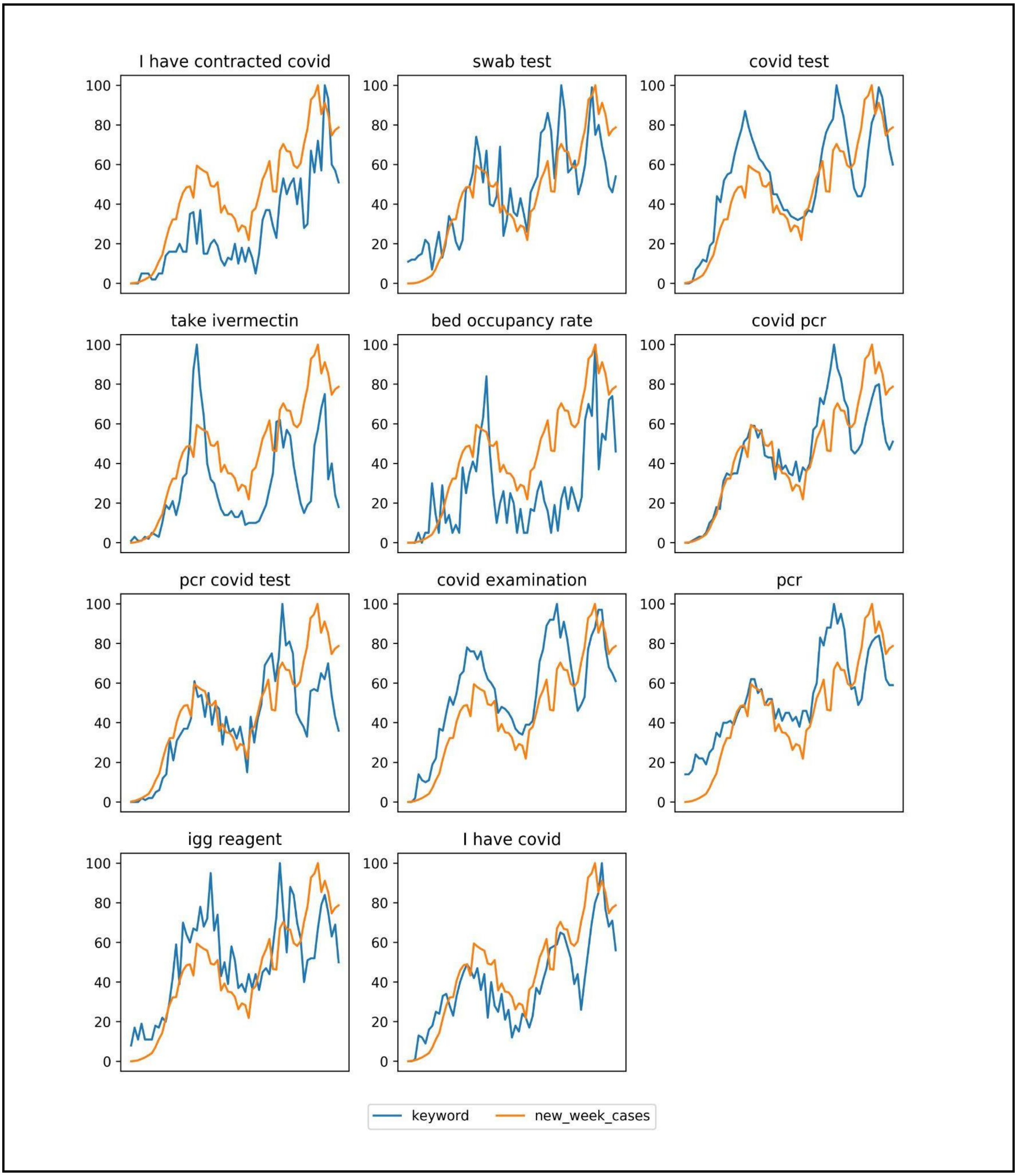


Figure 1. Keywords highly correlated to new cases per week in blue, new cases per week in orange. The x axis represents time in weeks, while the y axis represents % of popularity of the search term in relation to the peak.

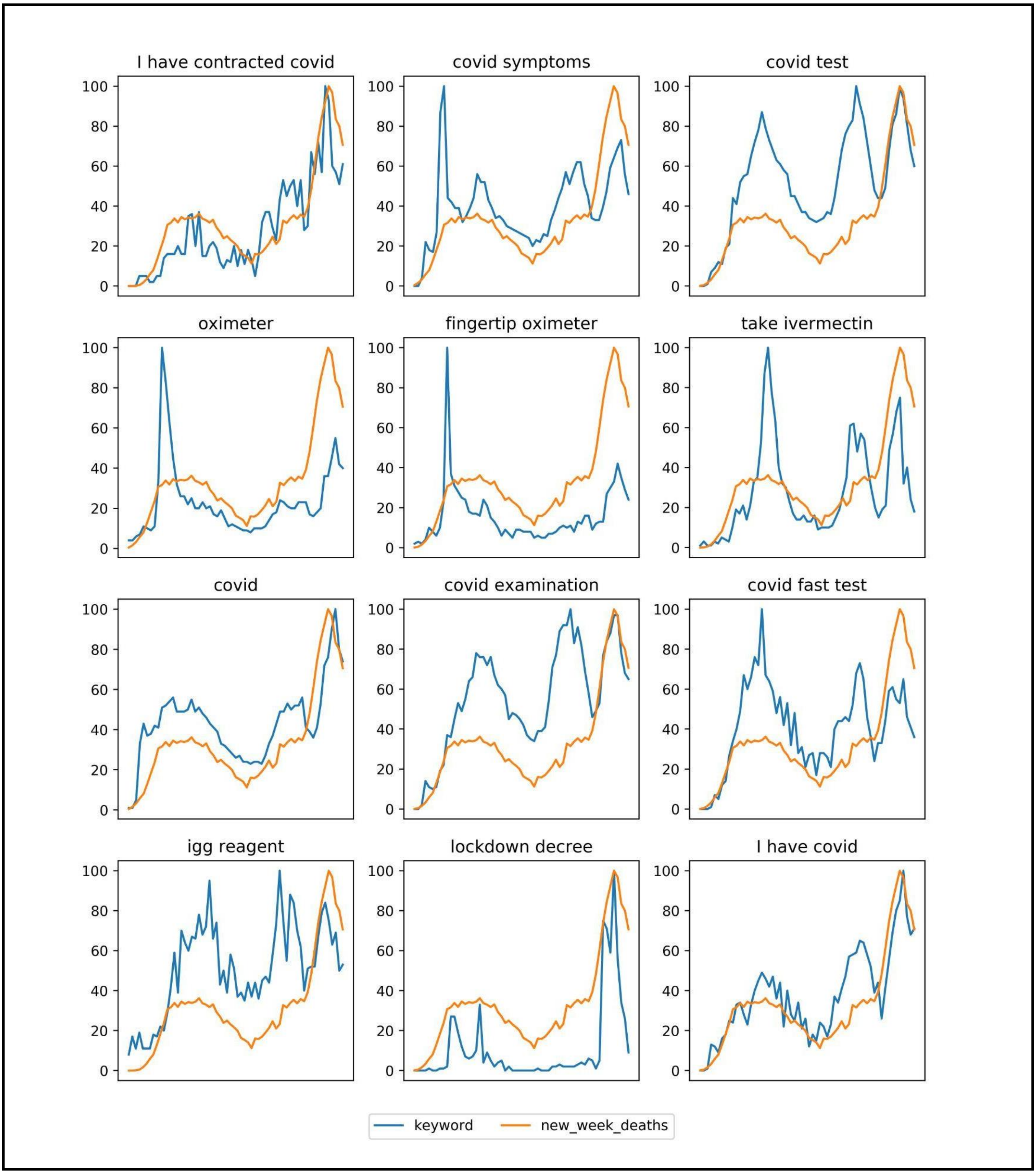


Figure 2. Keywords highly correlated to new deaths per week in blue, new deaths per week in orange. The x axis represents time in weeks, while the y axis represents % of popularity in relation to the peak.

Regressions

Evaluation metrics

As mentioned before, linear regressions, as well as ARIMA/ARIMAX models were used to attempt to predict official data by using GT. The evaluation metric of choice were the root mean squared error (RMSE), as well as the Mean Absolute Error (MAE) calculated as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}, \qquad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left|Y_i - \hat{Y}_i\right|, \tag{3}$$

where $\hat{Y}_i$ is the prediction and Y_i the true value. These metrics were chosen as both have been shown to be unambiguous and very useful when assessing models [17,18]. A split of 70% train and 30% test was used for the predictions.

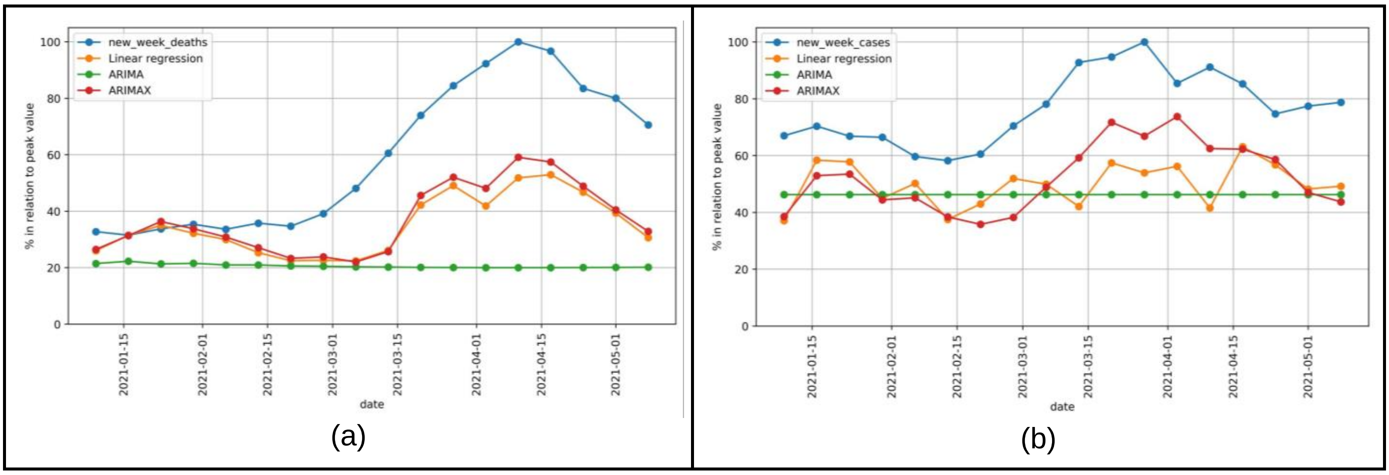
Comparing regressions

For the linear regression, a 5-fold cross validation was used to evaluate the model, obtaining a mean RMSE of approximately 12.1 and 8, and a mean MAE of 10.4 and 6.1, for the prediction of new cases and new deaths, respectively, across the folds, which is a significant error, considering the values range from 0 to 100.

As the ARIMA model uses no explanatory variables, it is suitable for univariate datasets. The usage of this model was to evaluate whether or not the data is independent of other variables. The root squared errors and mean absolute errors were slightly better than the linear regression, with RMSE of 6.14 and 3.7, and MAE of 4.4 and 2, for the prediction of new cases and new deaths, respectively. Though, as can be seen in Figure 3 and Figure 4, the forecast was atrocious, which rules out the possibility of predicting the indicator by using only its past values.

The ARIMAX model, using the keywords from GT, had great results. It had RMSE of 5.28 and 1.8 and MAE of 3.5 and 1.5, for the prediction of new cases and new deaths, respectively.

Figure 3 and Table 4 summarize the performance of the forecast for the models, while Table 3 shows the training errors.



(a) (b)

Figure 3. Comparison of the forecast models (% in relation to peak value X date). (a) Forecast of new cases per week; (b) Forecast of new deaths per week.

Table 3. Model errors (training)

Model	RMSE (new cases)	RMSE (new deaths)	MAE (new cases)	MAE (new deaths)
Linear Regression	12.15	8.06	10.49	6.14
ARIMA	6.14	3.74	4.42	2.03
ARIMAX	5.28	1.82	3.58	1.52

Table 4. Model errors (forecast)

Model	RMSE (new cases)	RMSE (new deaths)	MAE (new cases)	MAE (new deaths)
Linear Regression	29.31	29.89	25.31	24.50
ARIMA	32.72	46.30	30.28	38.6
ARIMAX	25.31	27.48	24.24	22.59

DISCUSSION

In the last few years, internet data has been increasingly useful for predicting real disease outbreaks around the world [9,19,20]. This is related to the fact that it has been observed that people who contract a disease are prone to search for their symptoms, or their illness, on the internet, before going to a healthcare facility[19]. Thus, in the case of Google Trends, the number of people searching for a coronavirus related keyword should increase before the official indicators do, especially for the search terms with high lag correlations.

Previous studies had found coronavirus and pneumonia as keywords with great correlation to the official indicators in India and China respectively [8,20]. In Brazil, the terms found by the analysis were mostly related to testing for COVID-19, with keywords such as “covid test”, “oximeter”, “igg reagent”, “swab test”, among others. This could suggest that people who contracted covid, start looking for ways to get tested. Some keywords similar to the ones found in previous studies were also identified with high correlations, such as “coronavirus symptoms“ and “I have contracted covid”. As the majority of optimal lags found are of 2 or 3 weeks, it opens the possibility to predict the increase and decrease of cases and deaths up to 21 days before it happens. It is also interesting to note that while most optimal lags in relation to new cases are of 1 or 2 weeks, most for new deaths are of 3 weeks, which is reasonable, as a person has to acquire a disease before passing away because of it.

As for the forecasts, the ARIMAX model had the best results overall, with much better predictions for new cases per week in relation to the other models. Linear regression was able to capture the general tendencies of the new deaths per week, but failed for the other indicator, while ARIMA could not predict either of them. Besides that, even the predictions for the best performing model had significant errors, and, thus, the value predicted has been shown to not be reflective of reality. What the ARIMAX model did succeed in was following the general trend of the curve, showing an increase when the official indicator is higher than its previous value and showing a decrease if the opposite happens, for most weeks.

It is important to note that, while Google Trends is useful for predicting real world occurrences, it should not replace traditional methods for predicting the spread of pandemics, but, rather, be used alongside it. Lastly, despite high correlations, GT data is susceptible to false alerts in case of an unusual event such as a drug recall for a popular cold or flu remedy [22].

For future studies, other models of forecasting could be attempted, such as those based on neural networks. Also, other internet data could also be used to make the system sturdier, such as Youtube, Twitter, and Facebook data. Lastly, the analysis could be done for each federative unit of Brazil individually, in an attempt to get more accurate results, as search engine results may vary depending on geographical location.

CONCLUSION

In a pandemic context, greater agility in providing information to public health institutions allows better allocation of resources, enhancing the capacity to contain the disease and resulting in a lower number of deaths. Given that, among the search terms evaluated in this study, some showed a high correlation with the selected epidemiological indicators, it demonstrates the potential of this method to analyze the space-time evolution of COVID-19 in Brazil and, eventually, to make epidemiological predictions with some days or even weeks in advance.

Though the predictions explored in this article should still only be used as a supplementary tool, the high lag correlations shown by some search terms, especially those related to coronavirus testing, as well as the forecasting model, can be further explored for the prediction of coronavirus official indicators in Brazil, such as new cases per week and new deaths per week.

Funding: This research was funded by the Brazilian National Council for Scientific and Technological Development (CNPq), the Coordination for the Improvement of Higher Education Personnel (CAPES) and the Minas Gerais Research Funding Foundation (FAPEMIG), and projects Covid Data Analytics (PRPq/UFMG/SESU/MEC), MASWEB, INCT-Cyber, and Atmosphere.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

1. König V, Mösges RA. A model for the determination of pollen count using google search queries for patients suffering from allergic rhinitis. J. Allergy. 2014;2014:1-9.
2. Bousquet J, O'Hehir R, Anto J, D'Amato G, Mösges R, Hellings P, et al. Assessment of thunderstorm-induced asthma using Google Trends. J. Allergy Clin Immunol. 2017;140(3):891-893.e7.
3. Ram S, Zhang W, Williams M, Pengetnze Y. Predicting asthma-related emergency department visits using big data. IEEE J Biomed Health Inform. 2015;19(4):1216-1223.
4. World Health Organization. Listings of WHO's response to COVID-19 [Internet]. 2020 [updated 29 January 2021; cited 11 July 2021]. Available from: https://www.who.int/news/item/29-06-2020-covidtimeline.
5. Boyd DM, Ellison NB. Social network sites: definition, history, and scholarship. J Comput Mediat Commun. 2007;13(1):210-230.
6. Higgins TS, Wu AW, Sharma D, Illing EA, Rubel K, Ting JY. Correlations of online search engine trends with coronavirus disease (COVID-19) incidence: infodemiology study. JMIR Public Health Surveill. 2020;6(2):e19702.
7. StatCounter Global Stats. Search Engine Market Share Brazil [Internet]. 2021 [updated 2021 May 10; cited 2021 July 18]. Available from: https://gs.statcounter.com/search-engine-market-share/all/brazil.
8. Venkatesh U, Gandhi PA. Prediction of COVID-19 outbreaks using Google Trends in India: a retrospective analysis. Healthc Inform Res. 2020;26(3):175-184.
9. Walker A, Hopkins C, Surda P. Use of Google Trends to investigate loss-of-smell–related searches during the COVID-19 outbreak. Int Forum Allergy Rhinol. 2020;10(7):839-847.
10. Google News Initiative. Google News Initiative Training Center [Internet]. 2021 [updated March 1 2021; cited 30 June 2021]. Available from: https://newsinitiative.withgoogle.com/training/lesson/5748139575214080?image=trends&tool=Google%20Trends.
11. Prasanth S, Singh U, Kumar A, Tikkiwal VA, Chong PHJ. Forecasting spread of COVID-19 using google trends: a hybrid GWO-deep learning approach. Chaos Solitons Fractals. 2021;142:110336.
12. Lu Y, Wang S, Wang J, Zhou G, Zhang Q, Zhou X, et al. An epidemic avian influenza prediction model based on google trends. Lett Org Chem. 2019;16(4):303-310.
13. Fantazzini D. Short-term forecasting of the COVID-19 pandemic using Google Trends data: evidence from 158 countries. Appl Econometr. 2020;59:33-54.
14. Ayyoubzadeh S, Ayyoubzadeh S, Zahedi H, Ahmadi M, R Niakan Kalhori S. Predicting COVID-19 incidence through analysis of Google Trends data in Iran: data mining and deep learning pilot study. JMIR Public Health Surveill. 2020;6(2):e18828.
15. Conejo A, Plazas M, Espinola R, Molina A. Day-ahead electricity price forecasting using the wavelet transform and ARIMA models. IEEE Trans Power Syst. 2005;20(2):1035-1042.
16. Adebayo G, Neumark Y, Gesser-Edelsburg A, Abu Ahmad W, Levine H. Zika pandemic online trends, incidence and health risk communication: a time trend study. BMJ Glob Health. 2017;2(3):e000296.
17. Chai T, Draxler R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci Model Dev. 2014;7(3):1247-1250.
18. Willmott C, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res. 2005;30:79-82.
19. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis. 2009;49(10):1557-1564.
20. Li C, Chen LJ, Chen X, Zhang M, Pang CP, Chen H. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. Euro Surveill. 2020;25(10).
21. Ministério da Saúde. Painel Coronavírus [Internet]. 2021 [updated 27 January 2022; cited 18 July 2021]. Available from: https://covid.saude.gov.br/.
22. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457(7232):1012-1014.
23. Cavalcante JR, Cardoso-Dos-Santos AC, Bremm JM, Lobo AP, Macário EM, Oliveira WK, et al. COVID-19 in Brazil: evolution of the epidemic up until epidemiological week 20 of 2020. Epidemiol Serv Saude. 2020;29(4).
24. Santos A, Gaspar P, Hamandosh A, Aguiar E, Guerra Filho A, Souza H. Best practices on HVAC design to minimize the risk of COVID-19 infection within indoor environments. Braz Arch Biol Technol. 2020;63.

Post Scriptum

The following terms words were investigated in Portuguese:

Bed occupancy rate: taxa de ocupação de leitos

Covid examination: exame covid

Covid fast test: teste rápido covid

Covid symptoms: sintomas do covid

Covid test: teste covid

Fingertip oximeter: oximetro de dedo

Igg reagente: reagente igg

I have contracted covid: peguei covid

I have covid: estou com covid

Lockdown decree: decreto lockdown

Oximeter: oximetro

Pcr covid test: pcr exame covid

Swab test: exame cotonete

Take ivermectint: tomar ivermectina

Taste and smell: paladar e olfato

Tiredness: cansaço