CEWP 20-05

Predicting the COVID-19 Pandemic in Canada and the US

Ba M. Chu and Shafiullah Qureshi

Carleton University & Ottawa-Carleton GSE

May 2020; revised 5 May 2020

CARLETON ECONOMICS WORKING PAPERS

Department of Economics

1125 Colonel By Drive
Ottawa, Ontario, Canada
K1S 5B6

Predicting the COVID-19 Pandemic in Canada and the US

Ba M Chu*
Carleton University & Ottawa-Carleton GSE
Shafiullah Qureshi†
Carleton University & Ottawa-Carleton GSE

May 5, 2020

Abstract

Our proposed time series model with the quartic trend function predicts that the peak of confirmed coronavirus cases has passed in Canada and the US while the end period of the pandemic will come around June 2020 in the best scenario and till the end of 2020 in the worst scenario. Both the bootstrap distance-based test of independence and the XGBoost algorithm reveals a strong link between the coronavirus case count and relevant Google Trends features (defined by search intensities of various keywords that the public entered in the Google internet search engine during this pandemic).

1 Introduction

Predicting the potential spread of a pandemic like COVID-19 is difficult because we do not have many epidemiological data, such as the transmission mechanism, the contagiousness of the virus, or its mutation patterns, as well as other complex human factors, such as the level of compliance with social distancing measures. Many models recently developed by infectious disease scientists [e.g., the Imperial College model Imai, Dorigatti, Cori, Donnelly, Riley, and Ferguson (2020) and The Reich Lab (2020)] can produce vastly different predictions as they are constructed based on various assumptions that may not be close to reality (such as the actual level of compliance with social distancing may be much higher than what is assumed in the model, or the infection rates can vary across different regions and groups of people, which cannot be easily captured by any model). In this paper, we propose a time series model with quartic trend functions to model the trajectories of the coronavirus pandemic in Canada and the US. We also demonstrate that Google Trends (GT) data collected for search intensities (normalized relative to the maximum of 100) of many different keywords that the public has entered into the Google internet search engine during the period of coronavirus outbreak can predict the number of reported cases very well. The potential reason why GT data on search frequencies can be used as good predictors for the case count is that internet search intensities indicate the people's interest in or anxiety about certain events surrounding the pandemic, and

*Department of Economics, B-857 Loeb, 1125 Colonel By Dr., Carleton University, Ottawa, Canada. Email: ba.chu@carleton.ca Tel: +1 613-520-2600 (ext. 1546)

†Department of Economics, A806 Loeb, 1125 Colonel By Dr., Carleton University, Ottawa, Canada; and Department of ORIC, NUML, Sector H-9, Islamabad, Pakistan. Email: shafiullah.qureshi@carleton.ca Tel: +1.613.520.2600 (ext. 3778)

the information provided by internet searches can also enhance the public's understanding of the threat of the coronavirus and its severe impact on various social and economic aspects. This understanding can make people more compliant with social distancing and other virus containment measures, thus leading to a reduction in coronavirus case counts.[1] Therefore, our approach is data-driven – we are trying to fit all the data available as much as possible using modern econometric tools.

Our model predicts that the number of coronavirus cases reported can reach its peak around late April 2020 while it will slow down, then eventually reach zero at around the end of May 2020 in Canada and the US. Since the confidence interval is so wide, the prediction of the peak time in the worst scenario can be the middle of May 2020 while that of the end time can carry over till the end of 2020. There are a few papers that have attempted to predict the coronavirus pandemic. Linton (2020) has used the quadratic trend equation to forecast the predicted peak for the various countries. Kuniya (2020) applied Susceptible-Exposed-Infected-Recovered (SEIR) compartmental model to predict peak for Japan. Similarly, Zahiri, RafieeNasab, and Roohi (2020) makes use of Susceptible, Infected, and Recovered(SIR) model to know the peak of the COVID-19 for Iran.

The remainder of the paper is organized as follows. Section 3.3 explains our main methods, including time series models with the quartic trend functions to predict the coronavirus pandemic, the bootstrap distance-based test of independence to statistically verify the link between the coronavirus case count and various GT features, and the XGBoost to fit the causal relationship of the case count to GT features. Section 3 presents a description of the data being used and our main empirical findings. Section 4 concludes this paper. The list of all the GT search keywords is given in an appendix at the end of this paper.

2 Prediction of COVID-19 Case Counts

Let Y_t represent a time series of COVID-19 case counts. As Y_t can be zero at several points in time, we define the logarithmic transform of Y_t as $y_t = \log(1 + Y_t)$. The (transformed) series y_t admits the following decomposition:

$$y_t = f(t) + \eta_t^{(y)}, \qquad t = 1, \ldots, T, \tag{2.1}$$

where T is the time horizon, $f(t)$ is a concave quartic (deterministic) trend function defined by $f(t) := a + bt + ct^2 + dt^4$ with $c \leq 0$ and $d < 0$, and $\eta_t^{(y)}$ is a stationary stochastic process centered at zero. As seen in Figure 1, it is most likely that the trend component $f(t)$ dominates the random component $\eta_t^{(y)}$ because the observations under our study have a little random variation. Also, a pandemic often increases slowly to the peak level once it starts (especially in large areas with big populations), then it slows down pretty fast and disappears eventually (when containment measures, such as social distancing and shutdown of nonessential services, start being put in place). Therefore, the trend function $f(t)$ must have some asymmetric concave shape. Linton (2020) employs the standard symmetric concave quadratic function for the trend component (which does not allow for the asymmetric recovery path of a pandemic). We have done some experiments with both the quadratic function and the quartic function and found that

[1] We have observed in our GT data sample that the rapid growths in the Google searches for 'COVID fever', 'soar throat', and 'WHO covid19' occurred at least ten days prior to the report date of actual coronavirus cases. It may well be that those who were searching for these terms had got infected with the virus. Therefore, we can also associate the search intensities of these keywords with the coronavirus case counts.

the proposed quartic function can be fitted to our data sets very well by using the simple unconstrained nonlinear least squares (NLS) method. We can then use this quartic function to predict the peak times and end times of the pandemic in Canada and the US.

Next, we study the question of whether there is a statistically significant link between the COVID-19 case counts and the Google search intensities of relevant keywords (reported in a GT data set). Fetzer, Hensel, Hermle, and Roth (2020) points out that the initial arrival of the coronavirus has led to a substantial surge in the internet searches of topics that reflect people's worries about the pandemic and economic anxiety. The spike in internet searches also indicates that the public is trying to enhance their understanding of the threat and contagious nature of the coronavirus as well as its economic consequences. Public education surrounding the coronavirus via the internet can increase the effectiveness of measures to contain the spread of the coronavirus, which in turn decreases the number of the infected population. We employ two main statistical methods to study the relationship between the COVID-19 case counts and a selected subset of the GT features: the distance-based test of independence and the XGBoost algorithm. The distance-based test of independence proposed by Chu (2020) tests if two non-Gaussian vectors of stationary time series are independent. We will perform the bootstrap version of this test statistic, which is proven to have good sizes and powers.

The XGBoost is a very popular machine-learning algorithm proposed by Chen and Guestrin (2016) as a penalized version of the well-known gradient boosting [see, e.g., Friedman, Hastie, and Tibshirani (2000)]. This original paper on the XGBoost has been cited over 5000 times, indicating the wide applicability of the method. The main idea of the XGBoost can be briefly described as follows. For a given data set of size T with N features, say $(y_t^*, \boldsymbol{x}_t^*)_{t=1}^T$ with $\boldsymbol{x}_t^*$ being a vector of N features, the model predicting the output y_t^* using the features $\boldsymbol{x}_t^*$ as predictors is a weighted additive model of the form: $y_t^* = \sum_{i=1}^M \alpha_i f_i(\boldsymbol{x}_t^*) + \epsilon_t$, where $f_i(\cdot)$ for $i = 1, \ldots, M$ are independent base learners (often characterized by regression trees), α_i are weights, and ϵ_t is a random error term. The XGBoost estimates both the weights α_i, $i = 1, \ldots, M$, and the associated base learners $f_i(\cdot)$ by *sequentially* minimizing a penalized differentiable convex loss function of $y_t^* - \sum_{i=1}^M \alpha_i f_i(\boldsymbol{x}_t^*)$ (with respect to both α_i and $f_i(\cdot)$, $i = 1, \ldots, M$) over M boosting iterations. The purpose of penalizing the complexity of the model (i.e., the regression trees) is to avoid overfitting so that the algorithm is more likely to select a simple model with good prediction power. Technical details about various boosting algorithms can be found in several good monographs [e.g., Friedman, Hastie, and Tibshirani (2009) and Schapire and Freund (2012)].

Let $\boldsymbol{x}_t$ denote a vector of N GT features observed at time t. Similar to (2.1), the multivariate time series $\boldsymbol{x}_t$ admits the following decomposition:

$$\boldsymbol{x}_t = \boldsymbol{g}(t) + \boldsymbol{\eta}_t^{(x)}, \qquad t = 1, \ldots, T, \tag{2.2}$$

where $\boldsymbol{g}(t) \coloneqq (g_1(t), \ldots, g_N(t))$ is a vector of the asymmetric and possibly concave quartic (deterministic) trend functions defined by $g_i(t) = a_i + b_i t + c_i t^2 + d_i t^3 + e_i t^4$, $i = 1, \ldots, N$, because many GT features soared around mid-March and then cooled down back to normal afterwards, and $\boldsymbol{\eta}_t^{(x)}$ is a vector of stationary stochastic processes centered at zero. To study the link between y_t and $\boldsymbol{x}_t$, we remove the trend components $f(t)$ and $\boldsymbol{g}(t)$ from y_t and $\boldsymbol{x}_t$ respectively and focus on $\eta_t^{(y)}$ and $\boldsymbol{\eta}_t^{(x)}$. We first apply the bootstrap distance-based test of independence to two time series $\eta_t^{(y)}$ and $\boldsymbol{\eta}_t^{(x)}$ to explore the strength of the link between

these variables. We then apply the XGBoost to study the extent to which random variations in $\boldsymbol{x}_t$ can lead random variations in y_t. We set $y_t^* \coloneqq \eta_t^{(y)}$ and $\boldsymbol{x}_t^* \coloneqq \boldsymbol{\eta}_t^{(x)}$ in our XGBoost model defined above.

3 Results

3.1 Data

We downloaded the data on the number of COVID-19 cases using the `R` package 'COVID19' and the GT data using the package 'gtrendsR'. The Google search keywords are listed in Table 4 (in the appendix). We use 53 search terms related to the coronavirus. We also use WTI crude oil prices and the CBOE volatility index (VIX) gauging the forward-looking expectation of investors about the future market condition. We have 100 observations (from 13/1/2020 to 21/4/2020) for Canada and 100 observations (from 16/1/2020 to (24/4/2020) for the US. Note that the first coronavirus case ever reported in Canada and the US is around January 2020; thus, all the case counts before that take values of zero. The reason why we need 100 observations is that, to implement the bootstrap distance-based test of independence, we need to fit the GT data to a dynamic conditional correlation (DCC) model of Engle (2002) using the `R` package 'rmgarch' of Ghalanos (2012) which requires at least 100 observations to run. However, we remove all observations with zero case counts when estimating the quartic trend function $f(t)$ because doing so can improve the NLS estimates.

Figure 1 clearly demonstrates that the logarithmically transformed (log-transformed) case count tends to increase with the search intensities of the five keywords ('corona', 'isolation', 'quarantine', 'stock market', and 'unemployment') forming a minimal subset of all the keywords (*or* GT features) used to estimate our XGBoost model. And the log-transformed case count curve tends to flatten as the search intensities decline. We select this small subset of keywords to aid the visualization of the data as joint plots of many GT features are very difficult to read.

Figure 2 shows some interesting patterns. The case count residuals $\widehat{\eta}_t^{(y)} \coloneqq y_t - \widehat{f}(t)$, where $\widehat{f}(t)$ is the (unconstrained) NLS estimate of $f(t)$ in (2.1), tends to move alongside the residuals, say $\left(\widehat{\eta}_{i,t}^{(x)}\right)_{i=1}^{5}$, of the aforementioned five GT features obtained from the NLS fit of $\boldsymbol{g}(t)$ in (2.2) to the GT data. It is also quite clear that $\widehat{\eta}_t^{(y)}$ leads $\left(\widehat{\eta}_{i,t}^{(x)}\right)_{i=1}^{5}$ for the most part of the sample period. This 'leading' patterns implies that the GT features have some predictability implication for the coronavirus case counts, which justifies our application of the XGBoost algorithm in Section 3.3 below.

3.2 The Peak and End Periods of the Pandemic

Figure 4 presents fitted trends of the log-transformed case counts. The quality of fit using the quartic trend function is excellent. The plot shows a clear asymmetric shape of the pandemic recovery paths. The curves flatten out in Canada and the US at around the end of April 2020. The predicted end period of the pandemic is around June 2020 in the best scenario (i.e., if the future observations still follow our estimated trend functions), and it can extend till the end of 2020 in the worst scenario. The confidence bands become wider for longer forecast horizons, indicating that it is difficult to make good predictions without epidemiological information, such as the coronavirus transmission mechanism or the contagiousness of the virus, and information about the effectiveness of measures to contain this virus.

3.3 GT Features as Good Predictors

First, we conduct the bootstrap distance-based test of independence to confirm that there is indeed a strong link between y_t and $\boldsymbol{x}_t$. As mentioned in Section 3.3 above, the test verifies if the two sequences of residuals $\widehat{\eta}_t^{(y)}$ and $\left(\widehat{\eta}_{i,t}^{(x)}\right)_{i=1}^{5}$ are independent. This procedure requires fitting the two sequences separately to time series models. Since the sequence $\widehat{\eta}_t^{(y)}$, $t = 1, \ldots, T$, is not very noisy as shown in Figure 4, a simple autoregressive process of order one could easily give an excellent fit. Meanwhile, as shown in Figure 2 the sequence $\left(\widehat{\eta}_{i,t}^{(y)}\right)_{i=1}^{5}$, $t = 1, \ldots, T$, exhibits a lot of random variations throughout the whole sample period. Therefore, we fit this sequence of residuals to the DCC model of order $(1, 1)$. We have done some experiments and found that this DCC model can provide the best fit. Also, augmented Dickey-Fuller (ADF) tests show that the two sequences of residuals are stationary.

The results of the bootstrap distance-based test are reported in Figures 5 and 6. In these figures, each value (h) of the bandwidth means that the sequence $\widehat{\eta}_t^{(y)}$ can lead or lag the sequence $\left(\widehat{\eta}_{i,t}^{(x)}\right)_{i=1}^{5}$ at most h periods. All the p-values are equal to zero; thus, the null hypothesis of independence is strongly rejected. Therefore, we can conclude that y_t can either lead or lag $\boldsymbol{x}_t$ over many periods. Finally, we implement the XGBoost to estimate the predictive relationship between $\eta_t^{(y)}$ and $\boldsymbol{\eta}_t^{(x)}$ (as predictors). Figure 7 shows that the residuals of all the GT features listed in Table 4 can predict the residuals of the log-transformed case count almost perfectly. Note that all these residuals themselves are quite small, so are the XGBoost errors [thus, they remain invisible on the two plots].

4 Conclusion

We can predict the peak and end periods of the coronavirus pandemic in Canada and the US by using a time series model with the quartic trend function. It seems that the height of coronavirus pandemic has almost passed based on the sample under our study. However, we may have to wait till the end of this year for this pandemic to end ultimately. Our predictions may change significantly if all lockdown measures are lifted early. Both the bootstrap distance-based test of independence and the XGBoost algorithm confirms that there is a strong link between the coronavirus case count and internet search intensities of relevant keywords.

References

Chen, T., and C. Guestrin (2016): "XGBoost: a scalable tree boosting system," *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Chu, B. (2020): "A distance-based test of independence between two weakly dependent stationary multivariate time series," mimeo, URL: `https://www.dropbox.com/s/5wzjbbjukkl9r92/test_independence_v13.pdf?dl=0`.

Engle, R. (2002): "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models," *Journal of Business & Economic Statistics*, 20(3), 339 – 350.

Fetzer, T., L. Hensel, J. Hermle, and C. Roth (2020): "Coronavirus perceptions and economic anxiety," mimeo, URL: `https://arxiv.org/pdf/2003.03848`.

Friedman, J., T. Hastie, and R. Tibshirani (2000): "Additive logistic regression: a statistical view of boosting," *Annals of Statistics*, 28(2), 337–407.

——— (2009): *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer.

Ghalanos, A. (2012): "rmgarch: multivariate GARCH models," mimeo, URL: `http://www.vps.fmvz.usp.br/CRAN/web/packages/rmgarch/`.

Imai, N., I. Dorigatti, A. Cori, C. Donnelly, S. Riley, and N. M. Ferguson (2020): "Estimating the potential total number of novel coronavirus (2019-ncov) cases in wuhan city, china," Imperial College London COVID-19 Response Team.

Kuniya, T. (2020): "Prediction of the epidemic peak of Coronavirus Disease in Japan, 2020," *Journal of Clinical Medicine*, 9(3), 789.

Linton, O. B. (2020): "When will the Covid-19 pandemic peak?," mimeo, URL: `http://covid.econ.cam.ac.uk/linton-uk-covid-cases-predicted-peak`.

Schapire, R. E., and Y. Freund (2012): *Boosting: Foundations and Algorithms*. MIT Press.

The Reich Lab (2020): "The COVID Forecast Hub," Web application, URL: `https://github.com/reichlab/covid19-forecast-hub`.

Zahiri, A., S. RafieeNasab, and E. Roohi (2020): "Prediction of Peak and Termination of Novel Coronavirus Covid-19 Epidemic in Iran," mimeo, URL: `https://www.medrxiv.org/content/10.1101/2020.03.29.20046532v1`.

Appendix

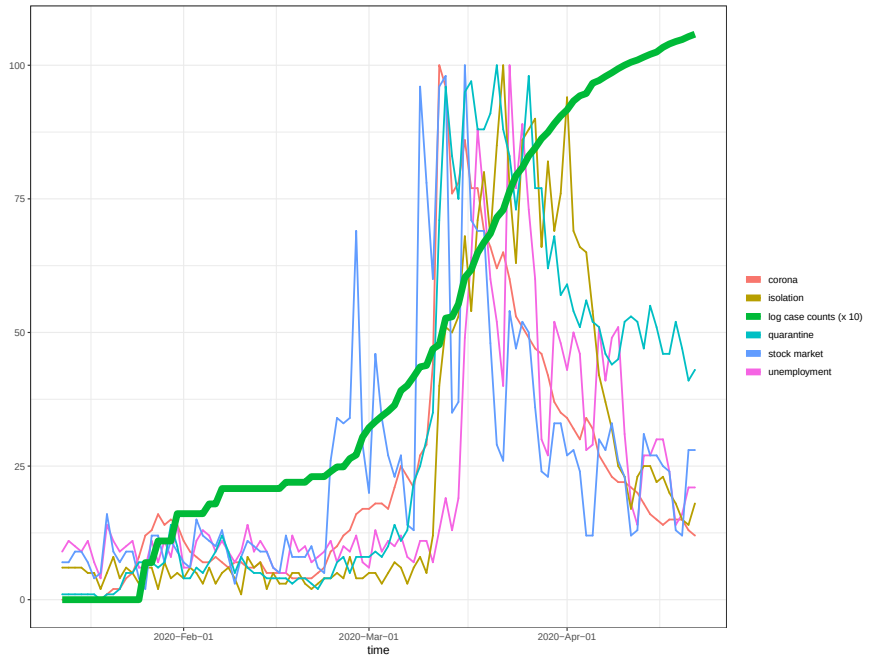
Table 1: Google search term for COVID-19

social distance	virus	Coronavirus	corona
corona virus	corona virus cases	corona virus update	corona virus canada
canada corona virus update	coronavirus update canada	coronavirus update	corona virus update ontario
canada covid19	covid19 cases	covid19 update	quebec covid19
covid19 Canada	covid19 in canada	death	unemployment
benefit	isolation	self isolation	quaran
Coronavirus quarantine	grocery	parks	flights
travelling	shopping	Lysol	prepping
Cancel trip	Carnivorous	Dog coronavirus	cat coronavirus
Contagion	Netflix_stock	handwashing	facemask
fever	Sorethroat	Shortnessofbreath	Lossofsmell
Lossoftaste	stockmarket	testing	WHO
WHOcovid19	mask	fear	hunger
handwash			

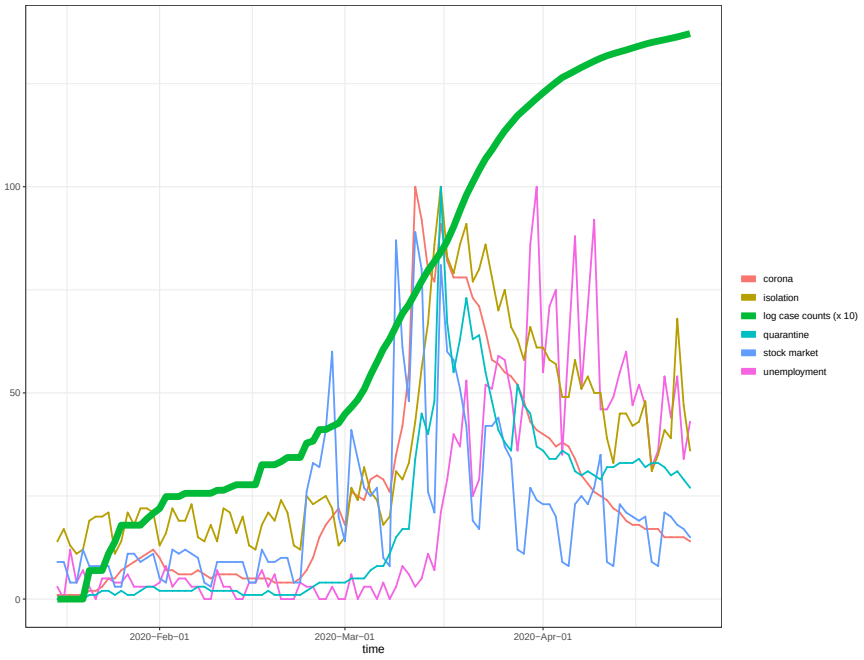
Table 2: Google Search Categories and Search frequency

Category	frequency	Category	frequency
canada covid19	100	coronavirus update	100
covid19 cases	53	corona virus canada	77
covid19 ontario	50	canada corona virus update	75
covid19 news	26	coronavirus update canada	38
alberta covid19	26	corona virus update ontario	25
covid19 update	24	update on corona virus	24
covid19 bc	24	bc corona virus update	22
quebec covid19	23	corona virus ontario	21
covid 19 canada	23	corona virus bc	21
covid19 in canada	23	corona virus news	21
covid19 quebec	23	corona virus update in canada	19
covid19 in ontario	16	corona virus in canada	19
covid19 symptoms	16	corona virus update alberta	15
covid19 cases canada	16	corona virus update live	13
covid19 toronto	16	covid update	13
covid19 map	15	corona virus world update	12
usa covid19	15	covid 19 update	12
covid19 world	13	corona virus update world	12
covid19 italy	13	corona virus update china	11
who covid19	13	corona virus china update	11

Figure 1: Plot of GT search keywords and log-transformed case counts (x 10)

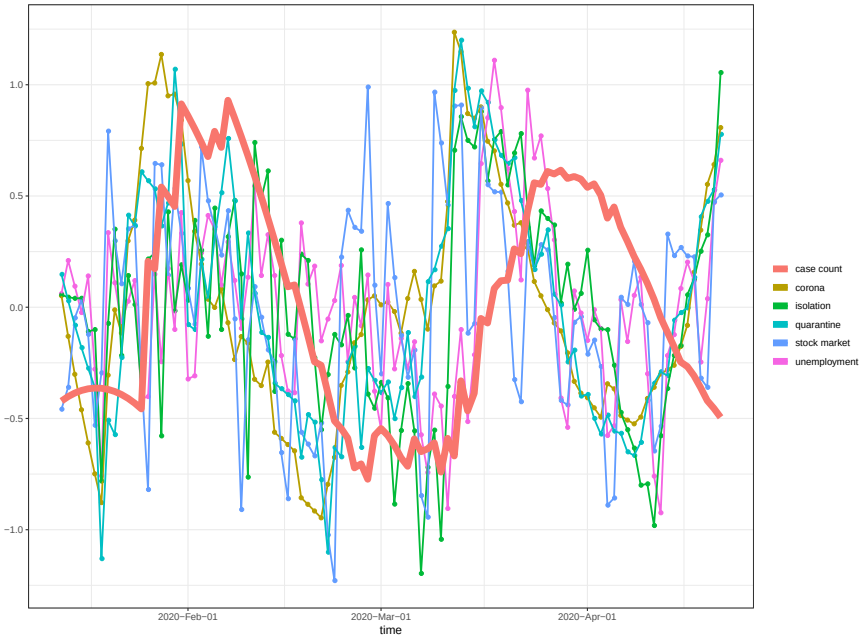


(a) Canada

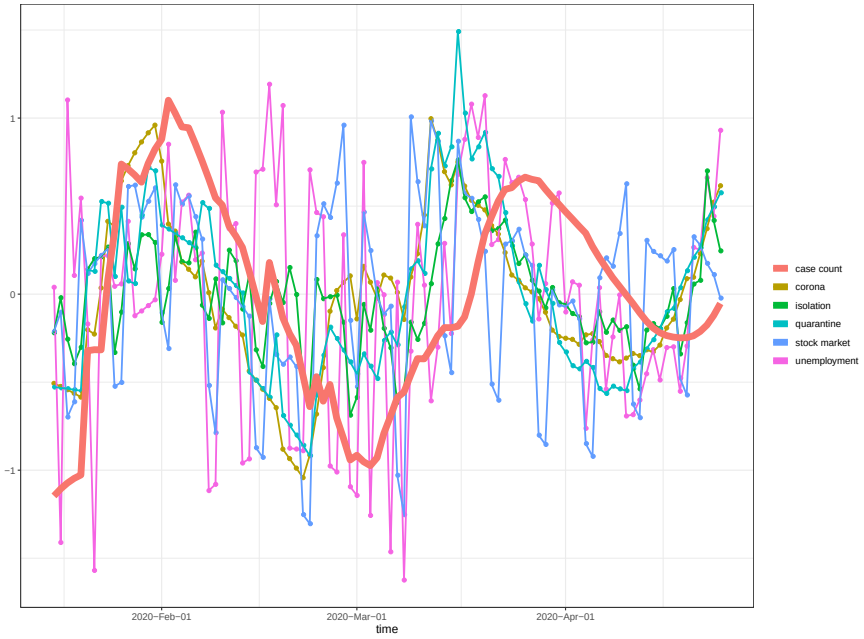


(b) US

Figure 2: Plot of residuals (obtained from the quartic equation fitting) of the GT search keywords vs. those (also obtained from the quartic equation fitting) of the log-transformed case count

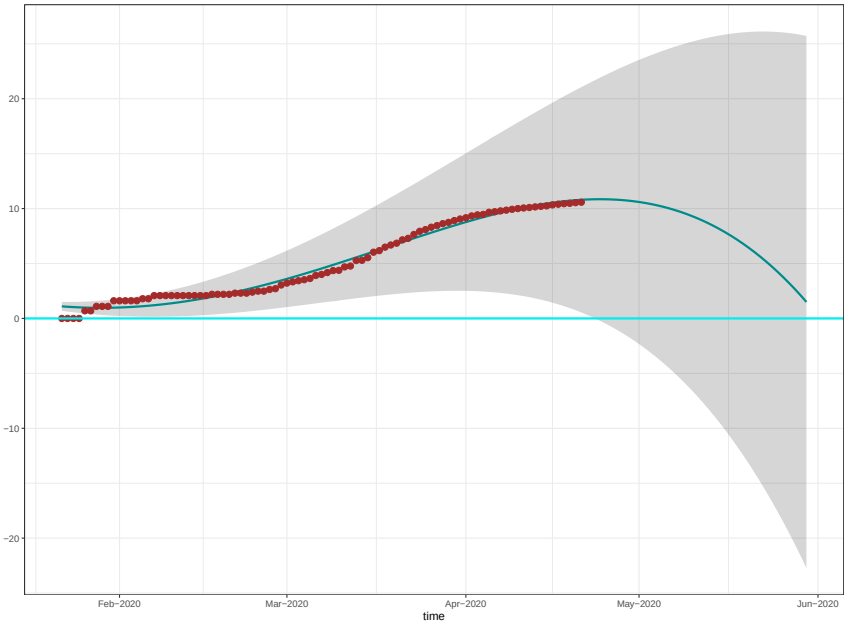


(a) Canada

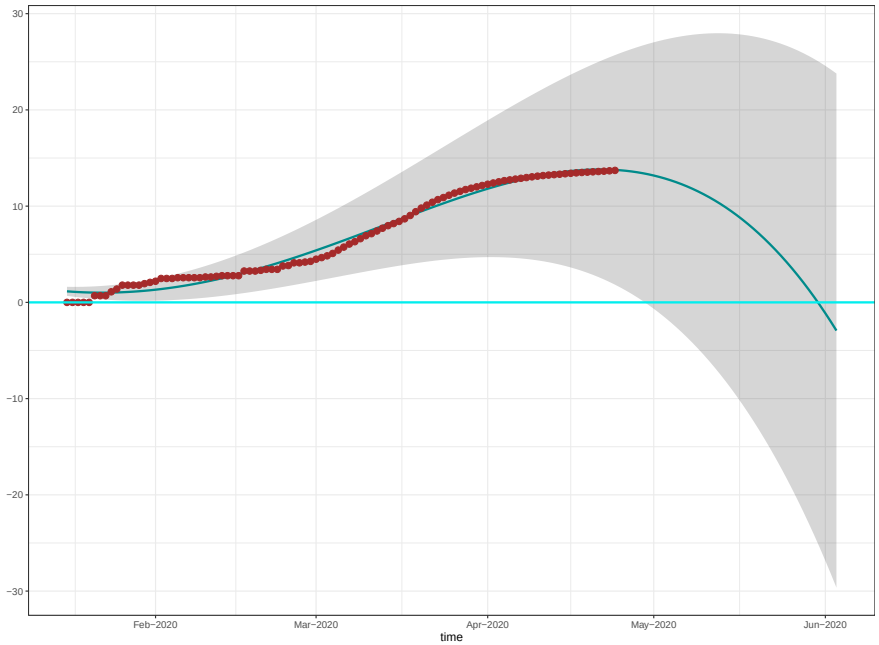


(b) US

Figure 3: The pandemic curves



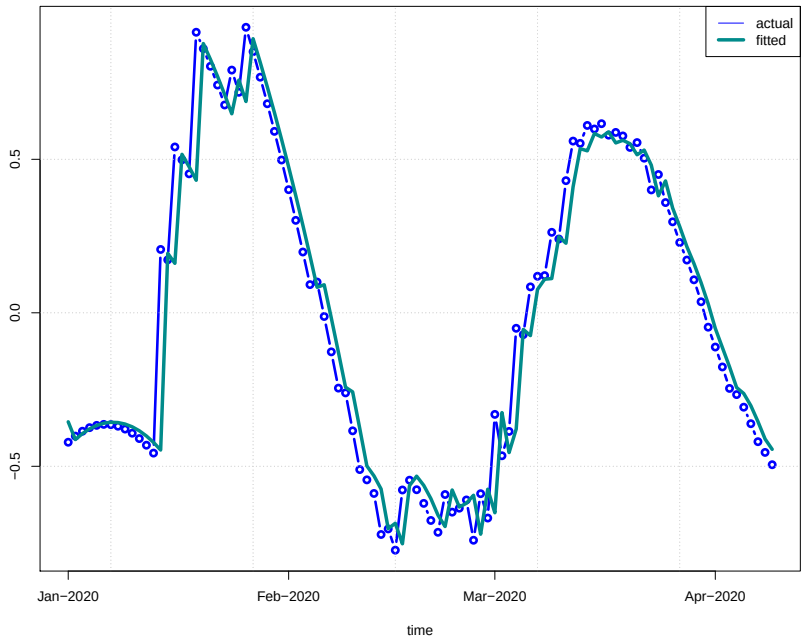
(a) Canada



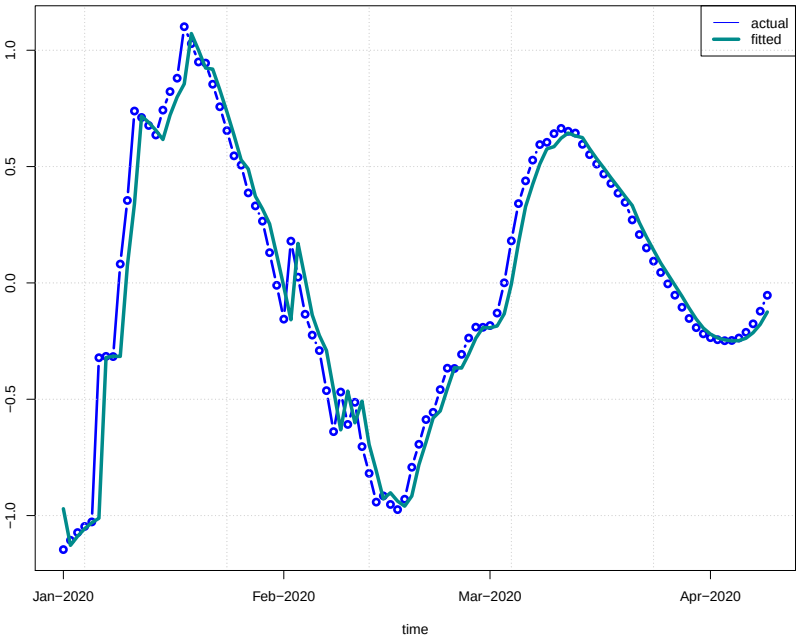
(b) US

[a] the brown dots are the actual log-transformed case counts
[b] the dark cyan line is the fitted curve using the quartic function $f(t) = a + bt + ct^2 + dt^4$
[c] the grey ribbon is the 95% confidence interval

Figure 4: Residuals (obtained from the quartic equation fitting) of the log-transformed case counts vs. their fitted autoregressive process of order one

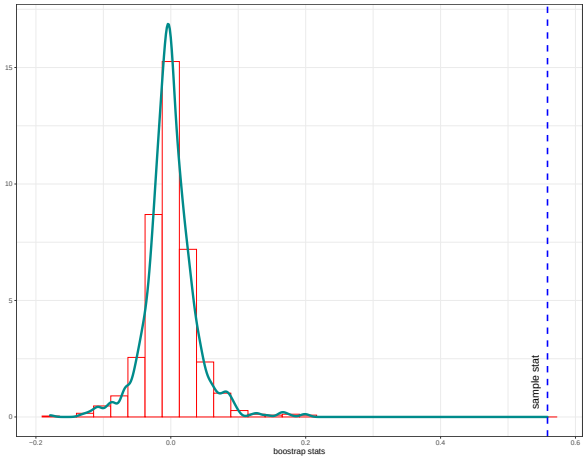


(a) Canada

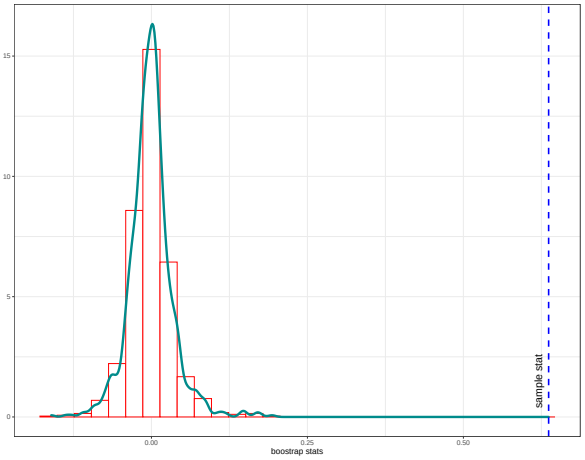


(b) US

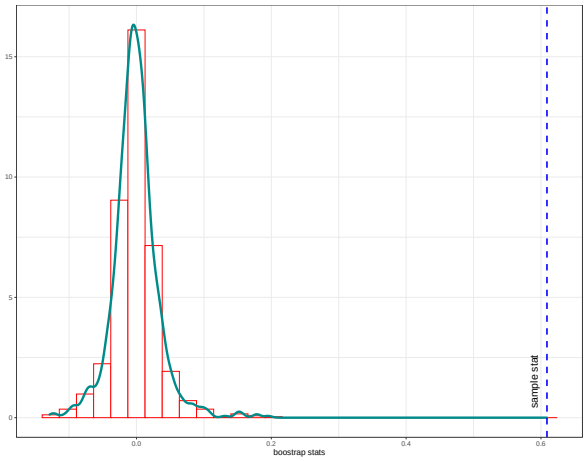
Figure 5: Histograms of the values of the bootstrap distance-based test for independence between the residuals (obtained from the quartic equation fitting) of the log-transformedcase counts and the residuals (also obtained from the quartic equation fitting) of the GT search keywords for Canada



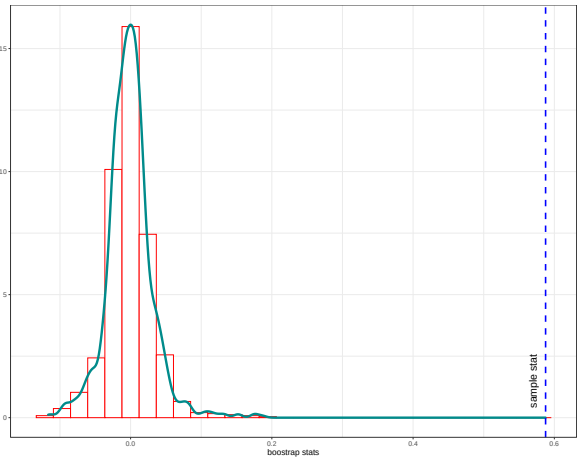
(a) bandwidth = 5



(b) bandwidth = 10

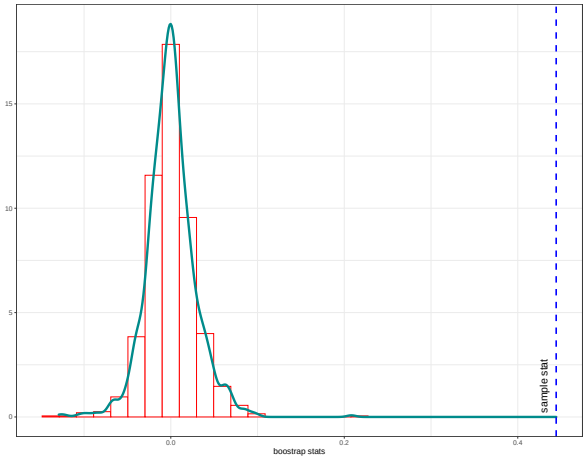


(c) bandwidth = 15

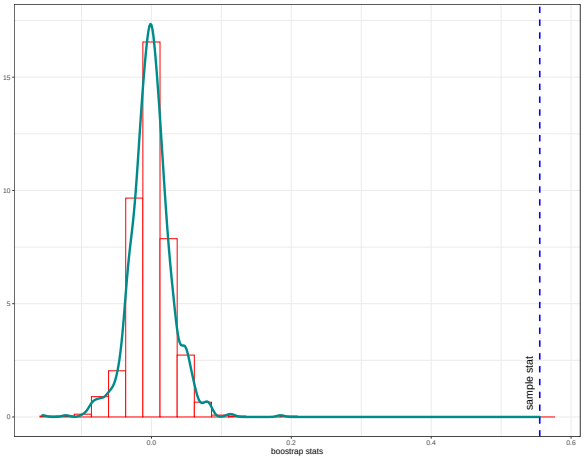


(d) bandwidth = 20

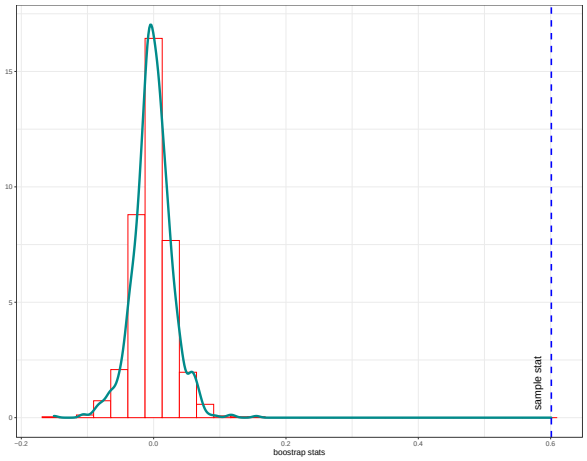
Figure 6: Histograms of the values of the bootstrap distance-based test for independence between the residuals (obtained from the quartic equation fitting) of the log-transformedcase counts and the residuals (also obtained from the quartic equation fitting) of the GT search keywords for the US



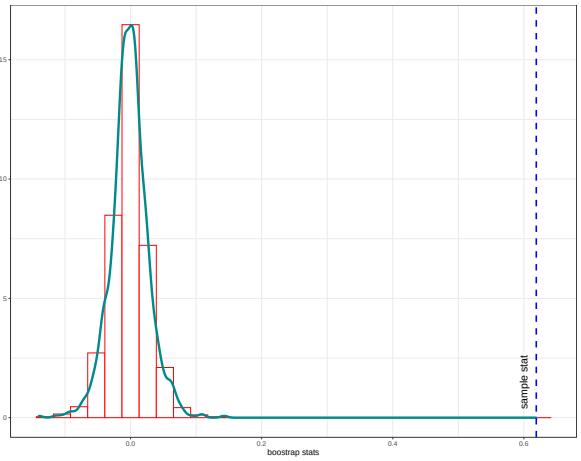
(a) bandwidth = 5



(b) bandwidth = 10

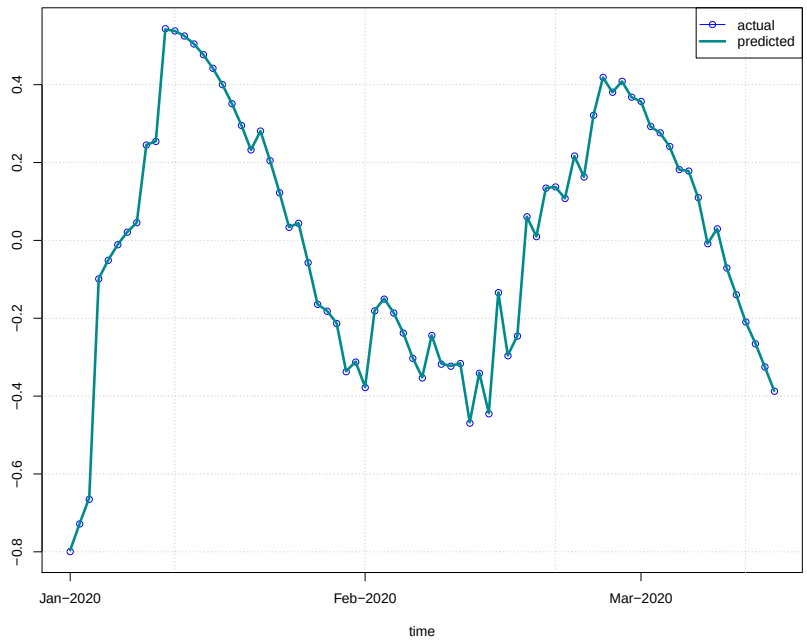


(c) bandwidth = 15

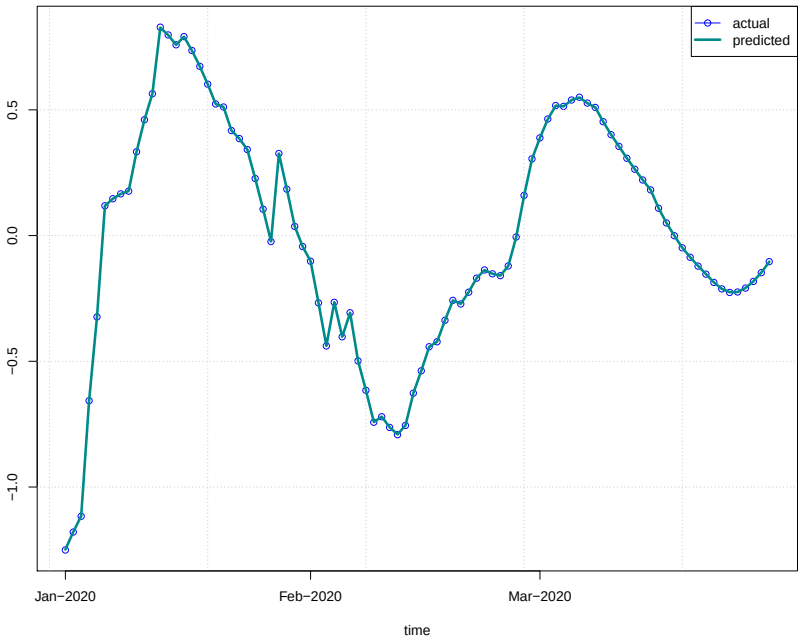


(d) bandwidth = 20

Figure 7: The XGBoost fit of the relationship between the residuals (obtained from the quartic equation fitting) of the log-transformed case counts and the residuals (also obtained from the quartic equation fitting) of the GT search keywords



(a) Canada



(b) US