Economics Bulletin

Volume 40, Issue 3

Predicting the COVID-19 pandemic in Canada and the US

Ba Chu
Carleton University

Shafiullah Qureshi
Carleton University

Abstract

We propose a time series model with the quartic trend function to make short-term forecasts of the COVID-19 confirmed cases in Canada and the U.S. Our one- to seven- days ahead out-of-sample forecast exercise demonstrates that the quartic trend model can produce very competitive short-term forecasts relative to the benchmark Susceptible, Infected, and Recovered (SIR) model. 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). Moreover, dynamic linear panel data models suggest a statistically significant relationship between the coronavirus case count and people's mobility trend provided by Google Mobility Reports (GMR) during the pandemic period.

Citation: Ba Chu and Shafiullah Qureshi, (2020) "Predicting the COVID-19 pandemic in Canada and the US", *Economics Bulletin*, Volume 40, Issue 3, pages 2565-2585
Contact: Ba Chu - ba.chu@carleton.ca, Shafiullah Qureshi - shafiullah.qureshi@carleton.ca.
Submitted: May 05, 2020. **Published:** September 24, 2020.

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).

This paper makes three main contributions. First, we propose a time series model with quartic trend function to model the coronavirus pandemic trajectory in Canada and the US. We then use this quartic trend function to construct short-term forecasts of COVID-19 confirmed cases. Second, we demonstrate that there is a strong association between the pandemic and Google Trends (GT) 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. The potential reason for this strong association is that internet search intensities indicate the people's interest in or anxiety about certain events surrounding the pandemic, and 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] Third, we employ dynamic panel data models to study the relationship between COVID-19 confirmed cases and people's mobility patterns at the province/state level in Canada and the U.S. A strong association between these variables would suggest that people tended to respond well to lockdown/social distancing measures taken to contain the virus. The Google Mobility Reports (GMR) provides data on various mobility trends across different crowded places (such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential). A positive GMR mobility index implies that the mobility trend is higher than that of a typical day while a negative index indicates that the mobility trend is lower than that of a typical day. Therefore, our approach is data-driven – we are trying to fit all the data available as much as possible using modern econometric tools.

As the case count data are highly persistent due to the contagious nature of the coronavirus (i.e., an infected individual is likely to transmit the virus to individuals that they make contact with), our quartic trend model can provide a very good fit for the confirmed case data as well as competitive short-term out-of-sample forecasts of the logarithmic case count relative to the benchmark Susceptible, Infected, and Recovered (SIR) model. We find a strong association between case counts and GT search intensities or GMR mobility indices. This association has two important implications: (1) the public was really concerned about this pandemic (which may be translated to negative sentiment about the economy) and (2) a lockdown measure to restrict people's mobility during the pandemic was effective in containing the spread of the virus (because this measure would not have been so effective if there was no correlation between case counts and GMR mobility index). We also mention at this point that a few papers have attempted to predict the coronavirus pandemic. Linton (2020) proposes the quadratic trend equation to forecast the

[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.

predicted peak for the various countries. Kuniya (2020) applies Susceptible-Exposed-Infected-Recovered (SEIR) compartmental model to predict peak for Japan. Similarly, Zahiri, RafieeNasab, and Roohi (2020) makes use of SIR model to predict the peak of the COVID-19 for Iran.

The remainder of the paper is organized as follows. Section 2 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, the XGBoost to fit the causal relationship of the case count to GT features, and the dynamic linear panel data models and their estimators used to model the relationship between case counts and aggregate GMR mobility indices. 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 Econometric Methods

2.1 Prediction of the Trend of COVID-19 Case Counts

Let Y_t represent a time series of cumulative COVID-19 case counts (so that a case count reported in a single day is actually the total number of all confirmed cases reported until that day). 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, $\eta_t^{(y)}$ is a stationary stochastic process centered at zero, and $f(t)$ is a concave quartic (deterministic) trend function defined by

$$f(t) \coloneqq a + bt + ct^2 + dt^4 \text{ with } c \leq 0 \text{ and } d < 0. \tag{2.2}$$

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 the proposed quartic function can be fitted to our data sets very well by using the simple unconstrained nonlinear least squares (NLS) method.

We also compare quartic trend forecasts generated by (2.2) with forecasts by the classic epidemiological SIR model [see, e.g., Smith and Moore (2001)]. Let S_t represent the number of people that are susceptible to the virus (at a particular point in time t), I_t be the number of infectious patients, and R_t denote the number of patients that have either recovered or died. The total population at time t is then defined by $N_t = S_t + I_t + R_t$. The SIR model specifies the (deterministic) dynamics of S_t, I_t, and R_t through the following system of differential equations:

$$\frac{dS_t}{dt} = -\beta^* I_t \frac{S_t}{N_t},$$

$$\frac{dI_t}{dt} = \beta^* I_t \frac{S_t}{N_t} - \gamma^* I_t,$$
$$\frac{dR_t}{dt} = \gamma^* I_t,$$

where β^* is the time-invariant contact rate governing the transition rate from S_t to I_t, and γ^* is the combined recovery and death rate of an infected individual. Then $\beta^* \frac{I_t}{N_t}$ is the average number of contacts that a susceptible person makes with an infectious person at each point in time. Given initial conditions, $R_1 = 0$ and $I_1 = y_0$, and specific values of S_1, β, and γ, one can just use the `R` package 'deSolve' to solve for the time paths S_t, I_t, and R_t. However, we need to estimate S_1, β, and γ from observations, $y_1, \ldots, y_T$. The sum of squared residuals can then be defined as $RSS\left(S_1, \beta, \gamma\right) \coloneqq \sum_{t=1}^{T} \left(y_t - (I_t + R_t)\right)^2$. The least squares estimates (LS) of S_1, β, and γ can be found by minimizing $RSS\left(S_1, \beta, \gamma\right)$, which is implemented by using the `R` routine 'optim'.

2.2 Relationship between COVID-19 Case Counts and Google Search Intensities

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 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: $y_t^* = \sum_{i=1}^M \alpha_i f_i\left(\boldsymbol{x}_t^*\right) + \epsilon_t$, 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 $y_t^* - \sum_{i=1}^M \alpha_i f_i\left(\boldsymbol{x}_t^*\right)$ (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:

$$\boldsymbol{x}_t = \boldsymbol{g}(t) + \boldsymbol{\eta}_t^{(x)}, \qquad t = 1, \ldots, T, \tag{2.3}$$

where $\boldsymbol{g}(t) \coloneqq \big(g_1(t), \ldots, g_N(t)\big)$ 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.

2.3 Dynamic Linear Panel Models of COVID-19 Case Counts and Google Mobility Trends

Let $\widetilde{y}_{i,t} \coloneqq y_{i,t} - \widehat{f}_i(t)$, where $y_{i,t}$ is the logarithmic case counts in province/state i at time t and $\widehat{f}_i(t)$ is the (unconstrained) NLS estimate of the quartic trend function $f(t)$ defined by (2.2) in province/state i at time t, represent the de-trended logarithmic case count. Let $x_{i,t}$ denote the total GMR mobility index (aggregated across all mobility indices for the six crowded places) in province/state i at time t. Since the GMR mobility indices $x_{i,t}$ across all the states declined during March 2020, then kept rising slowly from the beginning of April 2020 due to improving weather condition and relaxed lockdown measures [as shown in Figure 9.(c) and 10.(c)], we eliminate this trend non-stationarity from the data by de-trending these indices by fitting them to a cubic trend function, $g(t) = a_g + b_g t + c_g t^2 + d_g t^3$. We represent all the de-trended GMR mobility indices by $\widetilde{x}_{i,t} \coloneqq x_{i,t} - \left(\widehat{a}_g + \widehat{b}_g t + \widehat{c}_g t^2 + \widehat{d}_g t^3\right)$. We model the relationship between people's mobility trends and the number of confirmed COVID-19 cases via two dynamic linear panel data models: (1) a first-order autoregressive panel data model with an exogenous covariate and serially independent errors, and (2) a panel data regression model with serially correlated errors. The first panel data model is defined as

$$\widetilde{y}_{i,t} = \alpha_i + \beta \widetilde{x}_{i,t} + \gamma \widetilde{y}_{i,t-1} + \epsilon_{i,t}, \tag{2.4}$$

where α_i, $i = 1, \ldots, N$, throughout this section represent the fixed effects, and $\epsilon_{i,t}$ is a mean-zero homoscedastic random error term, and β and γ are the slope and AR coefficients respectively to be estimated. The second panel data model is defined as in Han and Phillips (2010):

$$\widetilde{y}_{i,t} = \alpha_i + \varphi \widetilde{x}_{i,t} + u_{i,t}, \text{where } u_{i,t} = \rho u_{i,t-1} + \epsilon_{i,t},$$

which can be transformed to

$$\widetilde{y}_{i,t} = (1-\rho)\alpha_i + \varphi\left(\widetilde{x}_{i,t} - \rho \widetilde{x}_{i,t-1}\right) + \rho \widetilde{y}_{i,t-1} + \epsilon_{i,t}, \tag{2.5}$$

where α_i, $i = 1, \ldots, N$, are the fixed effects as above, and $\epsilon_{i,t}$ is still a homoscedastic random error term, φ is the slope coefficient and ρ is the AR coefficient (both of which need to be estimated). This model is nonlinear in its coefficients, ρ and φ.

A fixed effects (FE) approach is commonly used to estimate (2.4). The main idea of this approach is to estimate $\alpha_1, \ldots, \alpha_N$ jointly with $\boldsymbol{\theta} \coloneqq (\beta, \gamma)^\top$ by quasi-Gaussian maximum likelihood which is equivalent to the within-group LS estimator as described in detail by Dhaene and Jochmans (2016). When T is fixed and N tends to infinity, the estimates $\widehat{\boldsymbol{\theta}}$ generated by this approach will generally be inconsistent as a consequence of the incidental parameter problem first noticed by Nickell (1981). This inconsistency

explains poor small-sample performance when T is very small relative to N. However, in many panel data applications, T and N are of similar orders of magnitude (in other words, $T/N \to$ const.). In this case, the estimate $\widehat{\boldsymbol{\theta}}$ may be expanded as follows:

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \frac{B}{T} + O\left(\frac{1}{T^2}\right).$$

[See, e.g., Arellano and Hahn (2007).] When T is fixed, the first-order bias term B/T is obviously non-zero. The recent dynamic panel literature proposes various strategies to remove this first-order bias term, so that the resulting bias-reducing estimator has a bias of smaller order, $O\left(\frac{1}{T^2}\right)$ instead of $O\left(\frac{1}{T}\right)$. The popular bias-reducing methods include an analytical method and a delete-one jackknife technique both proposed by Hahn and Newey (2004). When T is moderately large such that T/N is not negligible like in the U.S. panel data under our study, these methods may improve the finite-sample performance of $\widehat{\boldsymbol{\theta}}$. An interesting property of the bias-reducing estimators is that bias reduction does not cause an increase in the asymptotic variance as long as T/N tends to a constant (e.g., Dhaene and Jochmans (2016)).

Han and Phillips (2010) proposes to estimate Model (2.5) by the first difference least squares (FDLS) approach. Since this approach is based on an exact identifying moment condition constructed by first-differencing data, the GMM framework can then be used to estimate the parameters of interest. Therefore, there will be no bias to be corrected at the cost of some loss of asymptotic efficiency. Another quite important advantage of the FDLS estimator is that it is also asymptotically valid regardless of how N and T grow, fixed or large T, or even when the level of persistence is quite high.

3 Results

3.1 Data

We downloaded the data on the number of COVID-19 cases and GMR mobility data by using Guidotti's (2020) `R` package '`COVID19`', and the GT data by using the `R` 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. To study the relationship between coronavirus case counts and GT search intensities, we use 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 (because GT data beyond April 2020 do not necessarily reflect people's interest/ anxiety about the pandemic). 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 $\widehat{\eta}_t^{(y)} := y_t - \widehat{f}(t)$, 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 (2.3) 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 2.1 below.

Figures 9.(a,b) and 10.(a,b) clearly show that the trend of COVID-19 case counts and the GMR mobility trend are highly correlated across provinces/states in both Canada and the U.S.: the case counts increase sharply during the period of March 1 - April 1, when social distancing measures started coming into effect – this in turn leads to a sharp drop in the GMR mobility trends during the same period. The GMR mobility trends tend to slowly increase from mid-April onwards as the pandemic curves started flattening.

3.2 Out-of-Sample Prediction of Trends

Figure 4 presents fitted trends of the log-transformed case counts using both the SIR model and the quartic trend model (described in Section 2.1 above) for the full-sample periods (03/01/2020 - 07/21/2020 for Canada and 02/15/2020 - 21/07/2020 for the U.S.) The reason that both the trend models provide good fit to the data is that case counts are highly persistent due to the contagious nature of the coronavirus. It is still difficult to make long shot forecasts because the spread of the disease also depends on people's behaviour as well as other epidemiological factors which are still not fully understood.[2] However, as we just mentioned, the case count data are highly persistent and they all show clear time trends, thus one can still make short-term forecasts (say, one- to seven- days ahead forecasts). These short-term forecasts are more reliable as it is quite unlikely that a 'black swan' event or any other significant event (like the arrival of an effective vaccine for Covid-19) can occur within such a short forecast window.

To compare the out-of-sample forecast performance of the SIR model and the quartic trend model, we apply a rolling-window strategy. To be specific, a window consisting of 100 observations is used to estimate the models. These estimated models are then used to predict the logarithmic case counts up to seven days ahead. We conduct the same forecast exercise by rolling this window until we are at the seventh observation from the end of the full-sample period. We can then calculate the mean absolute error (MAE) of one- to seven- days ahead forecasts for each forecast window. In total, we have 37 windows for Canada and 52 windows for the U.S. Figure 4 shows that the SIR model predicts trends much better than the quartic model for Canada while the quartic model performs much better for the U.S.

3.3 Link between COVID-19 Case Counts and GT Features

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 2.1 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 5, 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

[2] We thank a referee for suggesting us this point.

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 6 and 7. 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 8 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].

3.4 Link between GMR Mobility Indices and COVID-19 Case Counts

We study the relationship between people's mobility trends and the number of confirmed COVID-19 cases across 8 provinces in Canada and 56 states and territories in the U.S. over 154 days (2/15/2020 - 7/17/2020). (We removed Prince Edward Island, Yukon, Nunavut, Newfoundland and Labrador, and Northwest Territories because either Google does not provide mobility data for those provinces or their numbers of confirmed cases are pretty low and invariant, thus they contain very little information.) We then fit the data on the de-trended logarithmic case counts and the de-trended GMR mobility indices [shown in Figures 9.(b,d) and 10.(b,d)] to the dynamic linear panel data models defined by (2.4) and (2.5) in Section 2.3. Estimation results are reported in Table 3. All the estimators produce negative estimates of the slope coefficients β and φ for both Canada and the U.S., which are consistent with what is shown in Figures 9 and 10. An increase in the COVID-19 case counts is often contemporaneously associated with a decline in the mobility trends (due to lockdown measures); and the numbers of case counts started flattening and decreasing while the mobility trends increased slowly. It is important to note that, since lockdown measures have lagging effects, a decrease in the mobility trends takes some time to translate into a reduction in the number of case counts. Thus, this negative correlation suggests that the recent flattening pandemic curve is caused by low mobility trends observed earlier on.

The confidence intervals of β are quite narrow while those of φ are all negative, suggesting that the slope coefficients are statistically significant. The estimates of the AR coefficients are quite high, showing that the de-trended case count at the province/state level is still very persistent for both Canada and the U.S. All the confidence intervals are positive. Therefore, the AR coefficients in these dynamic linear panel data models are statistically significant.

4 Conclusion

Our time series model with quartic trend function can generate competitive short-term out-of-sample forecasts for logarithmic case counts relative to the classic epidemiological SIR model. Both the bootstrap distance-based test of independence and the XGBoost algorithm confirm a strong link between the coronavirus case count and internet search intensities of relevant keywords. Our application of dynamic linear panel models suggests a statistically significant relationship between COVID-19 case counts and people's mobility patterns, which then implies that people responded quickly to lockdown measures by restricting their travel/visit to crowded places during the pandemic period.

References

ARELLANO, M., AND J. HAHN (2007): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments," in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, ed. by R. Blundell, W. Newey, and T. Persson, vol. 3, pp. 381–409. Cambridge University Press.

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`.

DHAENE, G., AND K. JOCHMANS (2016): "Bias-corrected estimation of panel vector autoregressions," *Economic Letters*, 145, 98–103.

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/`.

GUIDOTTI, E. (2020): *`R` package 'COVID19'*University of Neuchâtel.

HAHN, J., AND W. K. NEWEY (2004): "Jackknife and analytical bias reduction for nonlinear panel models," *Econometrica*, 72, 1295–1319.

HAN, C., AND P. C. B. PHILLIPS (2010): "GMM estimation for dynamic panels with fixed effects and strong instruments at unity," *Econometric Theory*, 26(1), 119–151.

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`.

NICKELL, S. (1981): "Biases in dynamic models with fixed effects," *Econometrica*, 49(6), 1417–1426.

SCHAPIRE, R. E., AND Y. FREUND (2012): *Boosting: Foundations and Algorithms*. MIT Press.

SMITH, D., AND L. MOORE (2001): "The SIR model for spread of disease," *Journal of Online Mathematics and Its Applications*, 3.

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

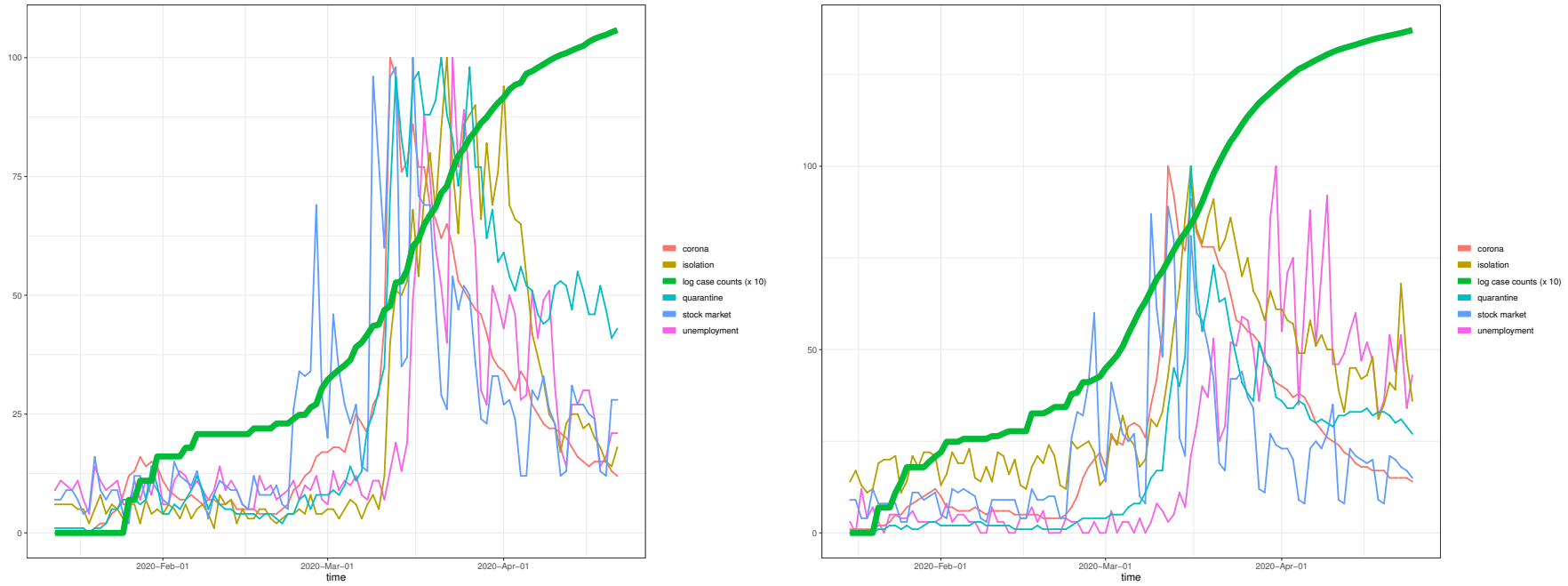
Category	frequency	Category	frequency
canada covid19	100	coronavirus update	100
covid19 cases	53	corona virus canada	77
covid19 ontario	50	canada corona virus update	75
covid19 news	26	coronavirus update canada	38
alberta covid19	26	corona virus update ontario	25
covid19 update	24	update on corona virus	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

Table 3: Estimation results for dynamic panel data models of the relationship between confirmed cases of COVID-19 and changes in visits to public places (provided by Google Mobility Reports)

	Canada			U.S.		
Coefficient	FE	Bias-corrected (BC) FE	BC FE confidence interval	FE	BC FE	BC FE confidence interval
β[a]	-0.014373	-0.011340	[-0.024218 , 0.001538]	-0.003823	0.000757	[-0.005998 , 0.007514]
γ[a]	0.939455	0.945665	[0.931801 , 0.959529]	0.940777	0.947231	[0.941952 , 0.952511]
	FDLS	FDLS confidence interval		FDLS	FDLS confidence interval	
φ[b]	-0.046909	[-0.064761 , -0.029056]		-0.116902	[-0.128460 , -0.105344]	
ρ[b]	1.366190	[1.185830 , 1.546560]		1.449330	[1.336280 , 1.562380]	

[a] β and γ are the coefficients of Model (2.4), estimated by using the *fixed effects* (FE) estimator and the bias-corrected (BC) FE estimator.
[b] φ and ρ are the coefficients of Model (2.5), estimated by using Han and Phillips's (2010) *first difference least squares* (FDLS) estimator.

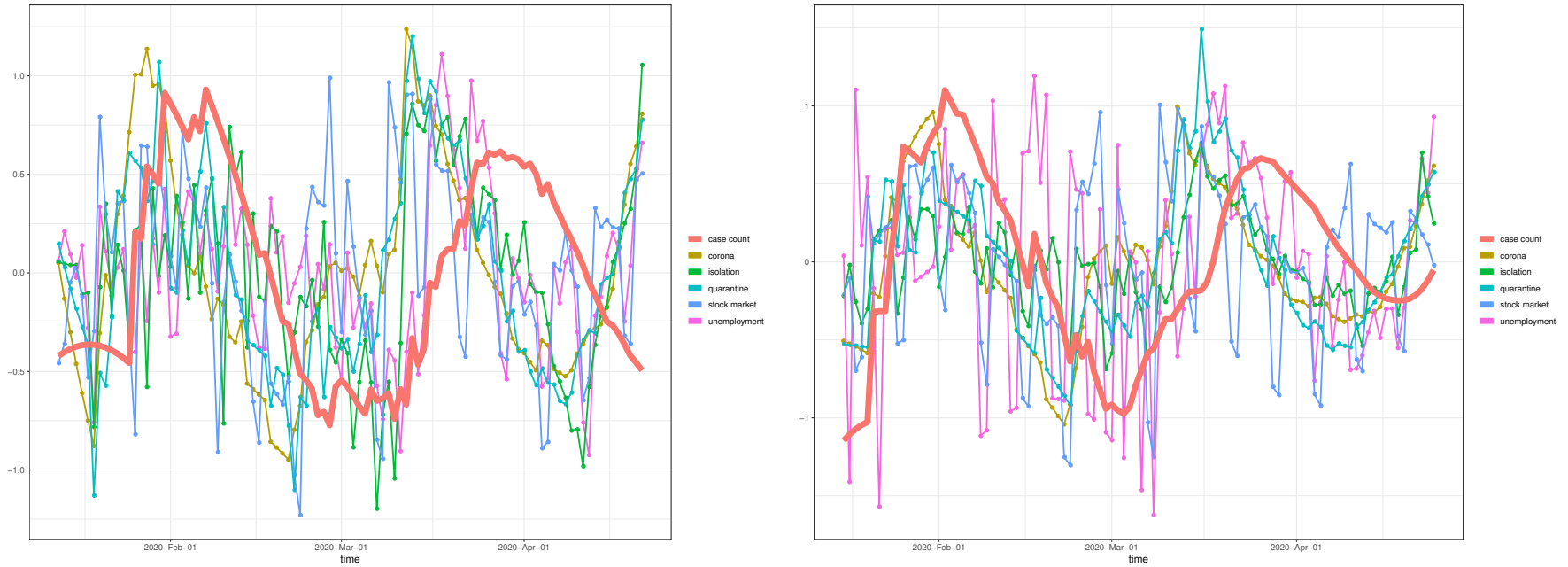
Figure 1: Plot of GT search keywords and log-transformed case counts (x 10)



(a) Canada

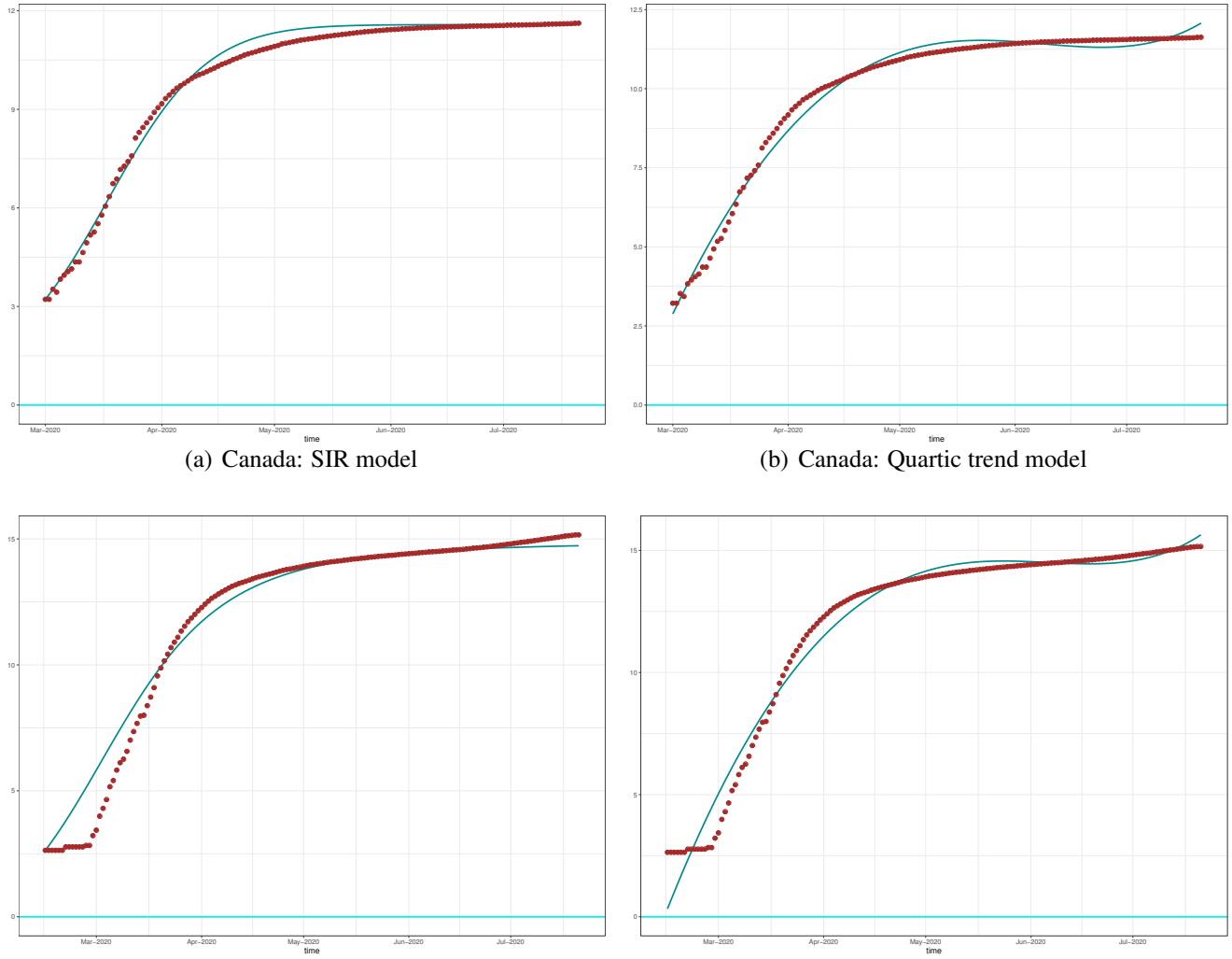
(b) US

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: Plots of logarithmic case counts versus their predicted values for the whole sample period (02/15/2020 - 07/17/2020)



(a) Canada: SIR model

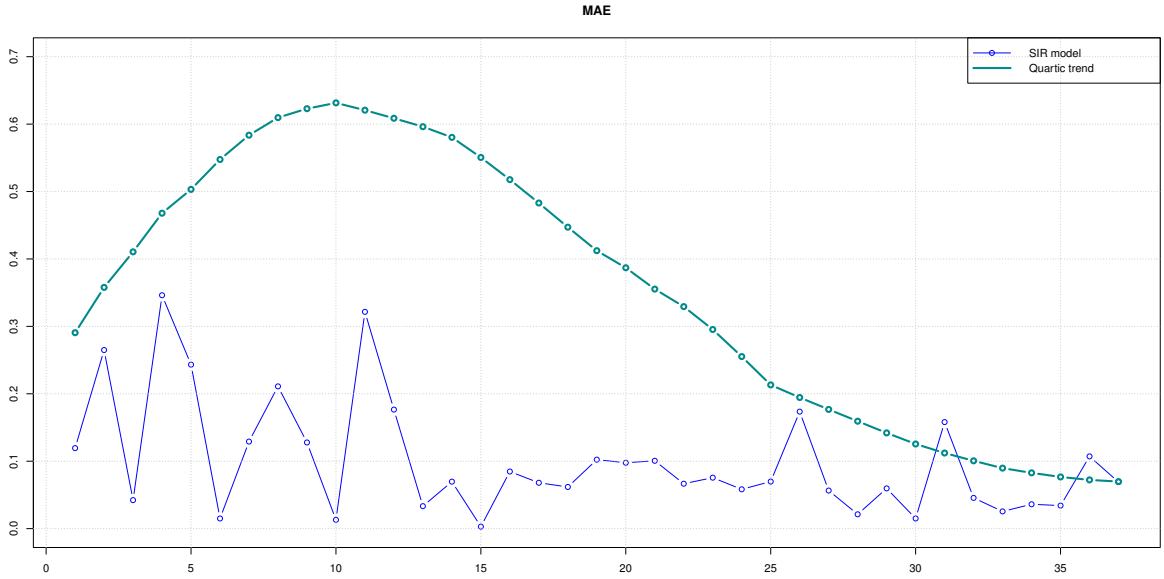
(b) Canada: Quartic trend model

(c) US: SIR model

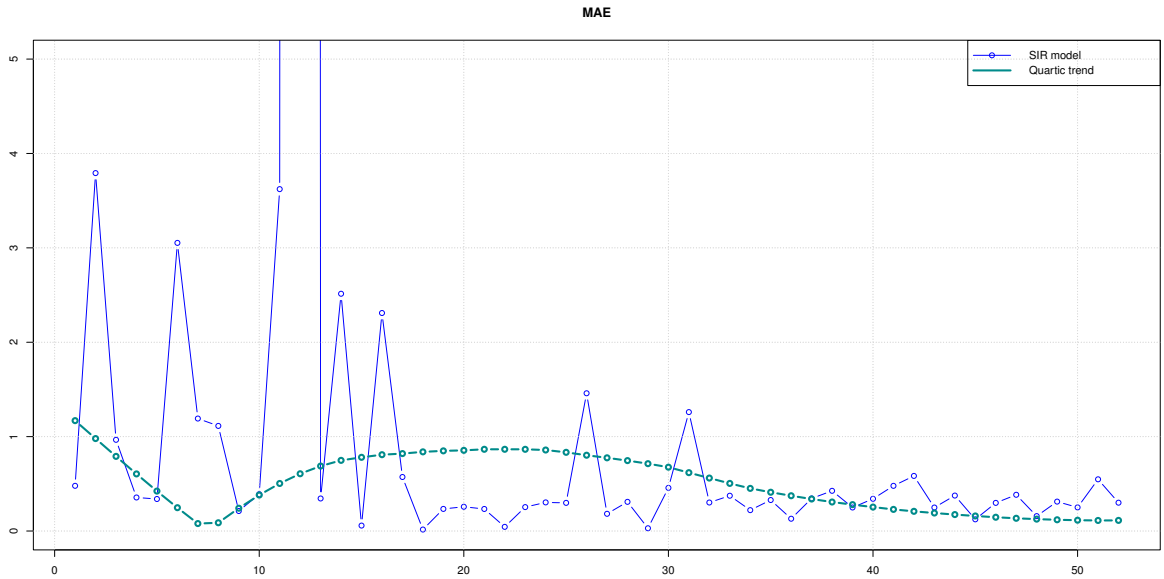
(d) US: Quartic trend model

[a] the brown dots are the actual log-transformed case counts
[b] the dark cyan lines are fitted curves

Figure 4: Plot of mean absolute errors (MAE) of one-to-seven days ahead rolling-window forecasts with each in-sample period (window size) containing 100 observations

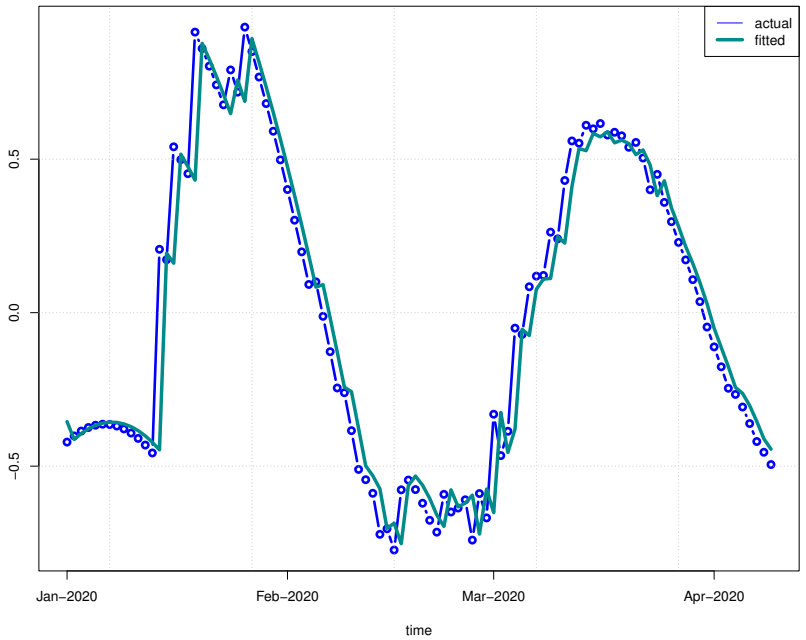


(a) Canada

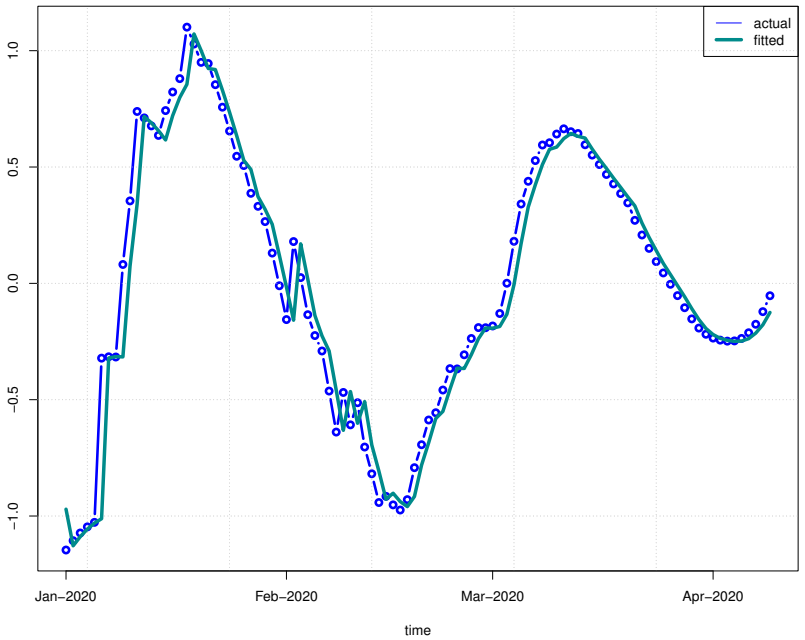


(b) US

Figure 5: Residuals (obtained from the quartic equation fitting) of the log-transformed case counts vs. their fitted autoregressive process of order one

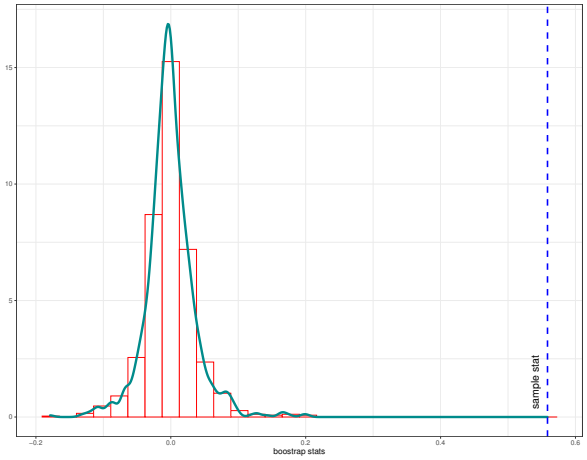


(a) Canada

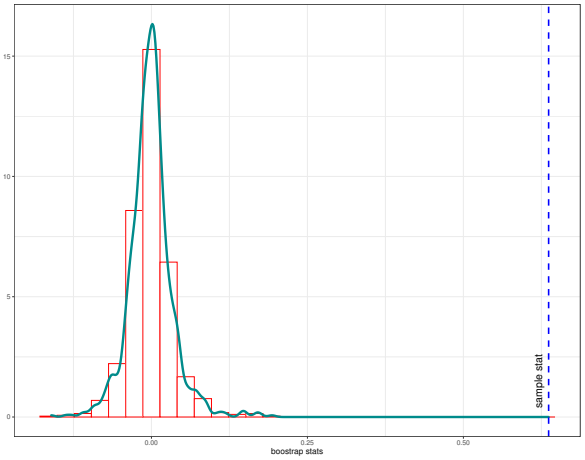


(b) US

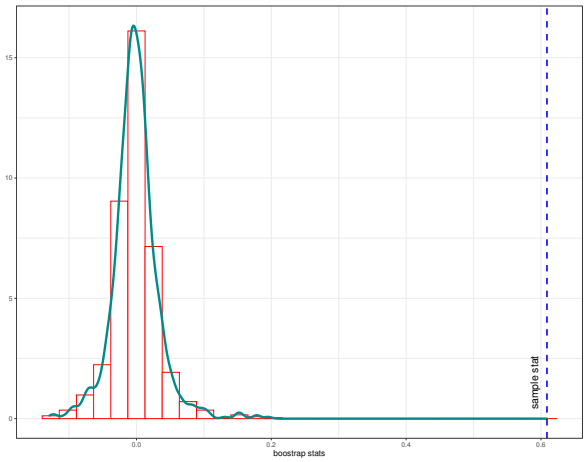
Figure 6: Histograms of the values of the bootstrap distance-based test for independence between the residuals (obtained from the quartic equation fitting) of the log-transformedcase counts and the residuals (also obtained from the quartic equation fitting) of the GT search keywords for Canada



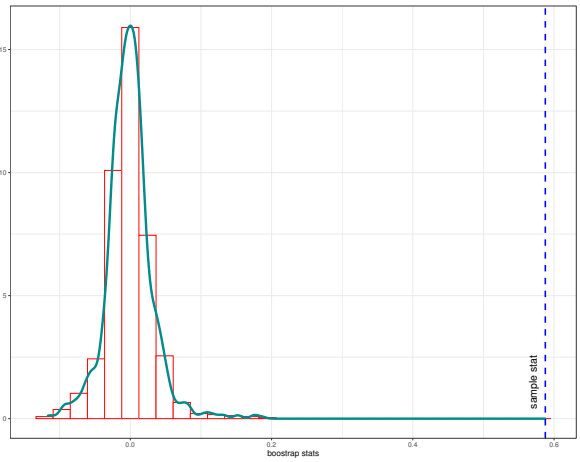
(a) bandwidth = 5



(b) bandwidth = 10

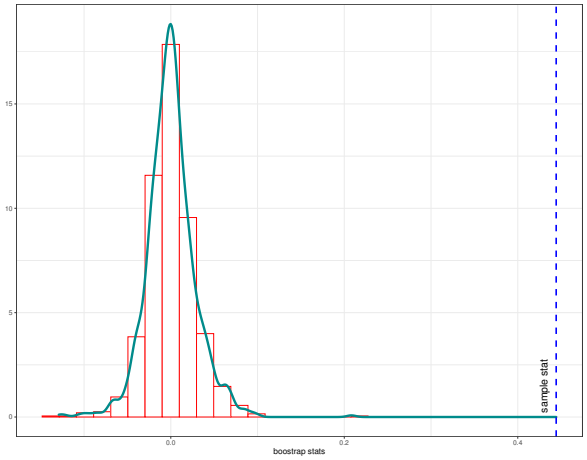


(c) bandwidth = 15

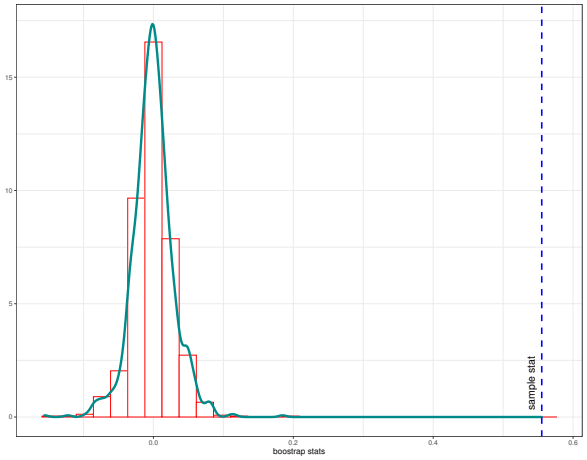


(d) bandwidth = 20

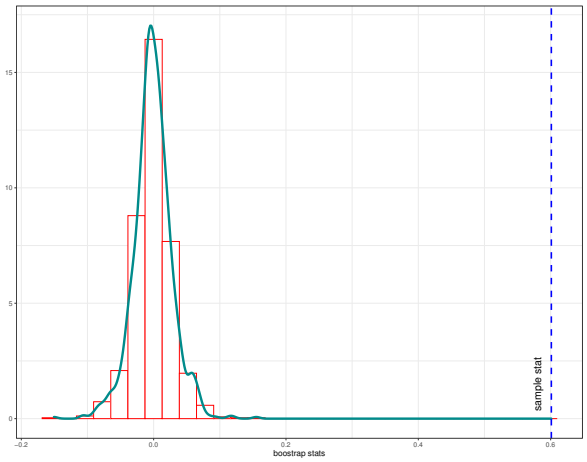
Figure 7: Histograms of the values of the bootstrap distance-based test for independence between the residuals (obtained from the quartic equation fitting) of the log-transformedcase counts and the residuals (also obtained from the quartic equation fitting) of the GT search keywords for the US



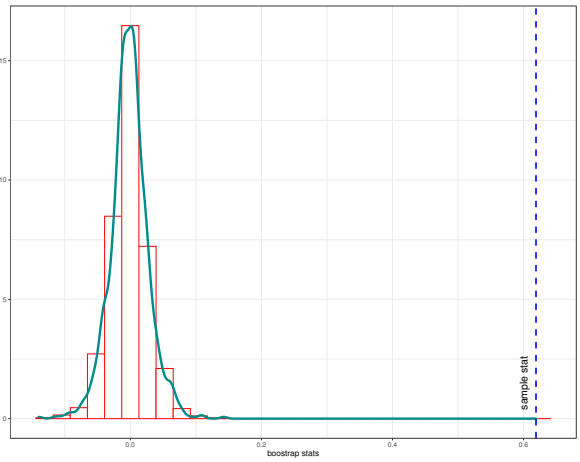
(a) bandwidth = 5



(b) bandwidth = 10

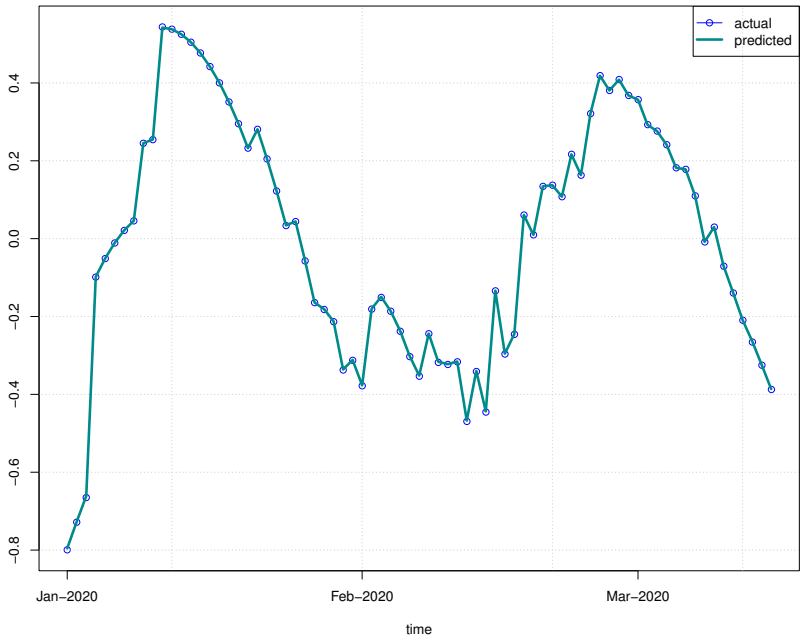


(c) bandwidth = 15

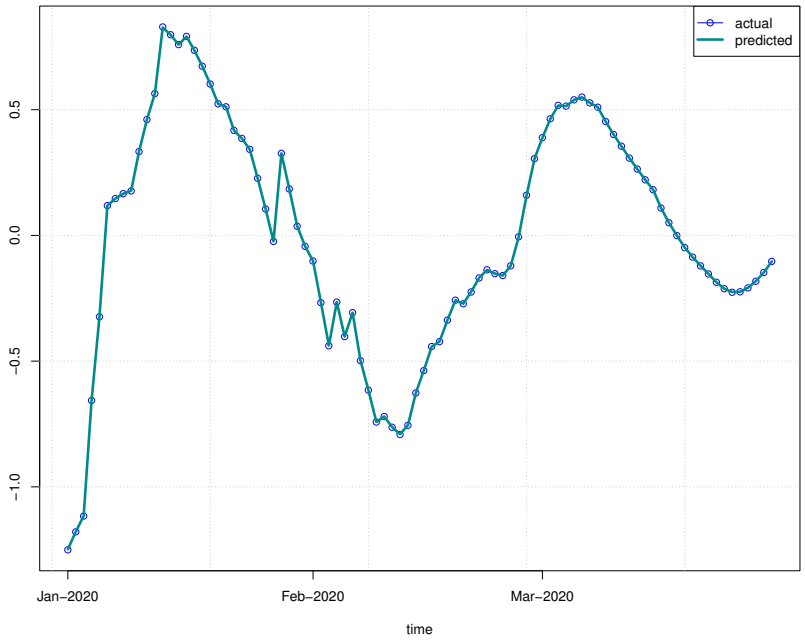


(d) bandwidth = 20

Figure 8: 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

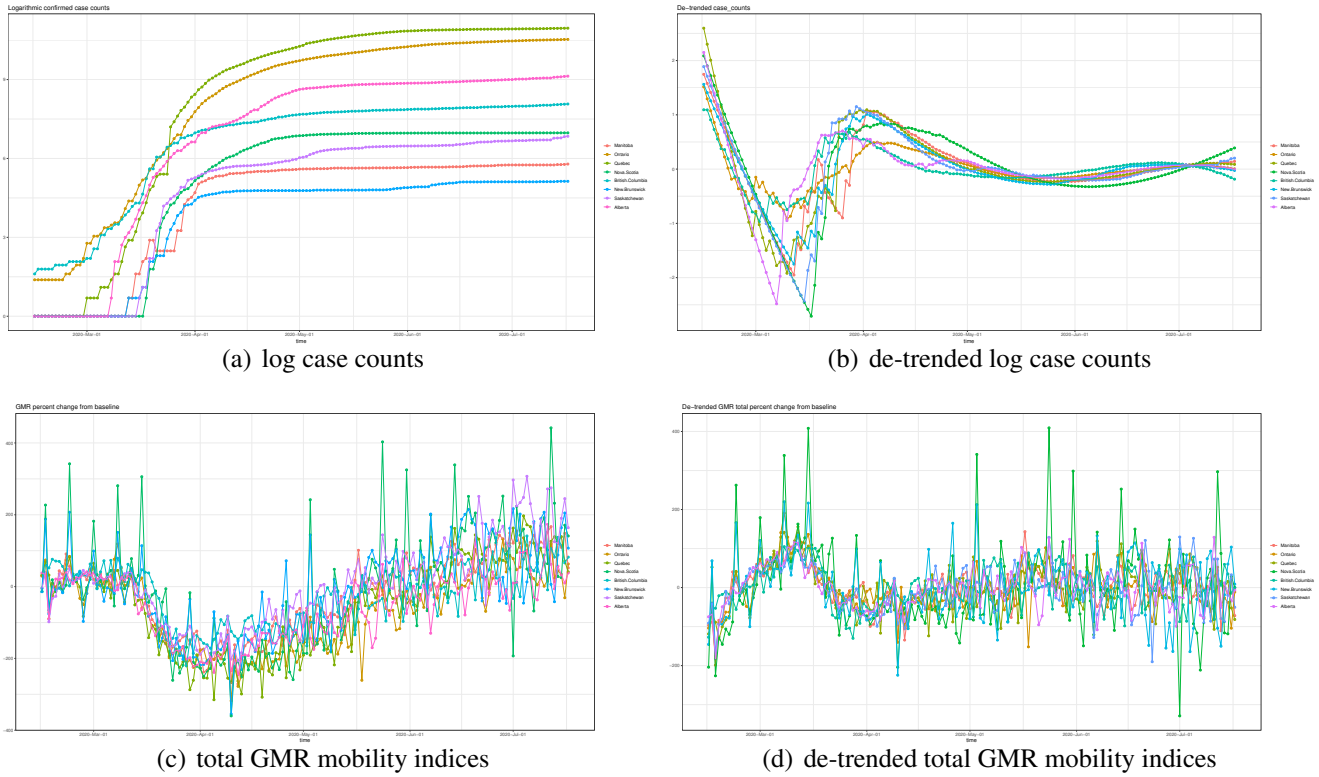


(a) Canada



(b) US

Figure 9: Plots of logarithmic case counts, de-trended logarithmic case counts, total Google Mobility Reports (GMR) mobility trends, and de-trended total GMR mobility trends across eight Canadian provinces



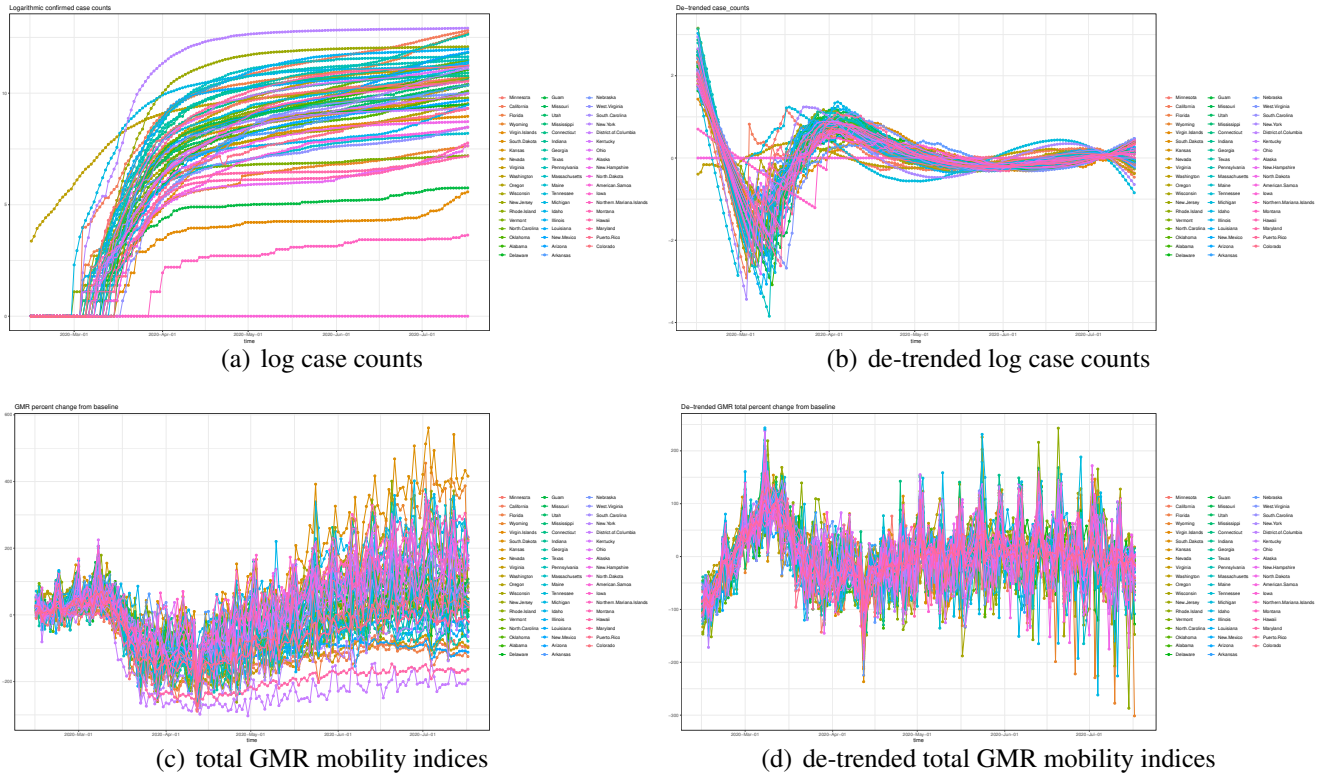
(a) log case counts

(b) de-trended log case counts

(c) total GMR mobility indices

(d) de-trended total GMR mobility indices

Figure 10: Logarithmic case counts, de-trended logarithmic case counts, total Google Mobility Reports (GMR) mobility trends, and de-trended total GMR mobility trends across 56 U.S. states and territories



(a) log case counts

(b) de-trended log case counts

(c) total GMR mobility indices

(d) de-trended total GMR mobility indices