CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Analysis and ARIMA-based Forecasts of COVID-19 Daily Confirmed Cases for Selected Canadian Provinces

AUTHOR: Jiawei Xu

SUPERVISOR: Patrick Farrell

DATE: Apr.14th, 2021

Abstract

On March 11, 2020, the World Health Organization (WHO) identified the Novel Coronavirus (Covid-19) outbreak as a global pandemic. The COVID-19 epidemic, which has lasted for more than a year, has had an indelible impact on countries and people around the world. Forecasts of COVID-19 trends are various and complicated. This project aims to introduce time series analysis about dataset and generate appropriate Auto-regressive Integrated Moving Average (ARIMA) models for predicting data of the number of daily confirmed cases for Ontario, Quebec and British Columbia. Also, results of forecasts, the accuracy of models and imperfections are presented in this paper.

Key Words: Canada, COVID-19, Time Series, ARIMA, Forecasts

Table of Contents

1 Introduction

1.1 Motivation and Objectives

The rapid expansion of the scale of the epidemic has led to the introduction of many emergency measures and countermeasures in each country. In March of 2020, Canada has entered a sobering COVID-19 reality and has only just begun to put in place more aggressive public health policies of containment that are focused on slowing disease progression or flattening the curve. An already overburdened healthcare system has insufficient surge capacity and cannot cope unless significant measures are undertaken. Slowing the progress of the virus is critical to slow the rate of inevitable hospitalizations.

In some ways, the epidemic is bad news for a country, not only because it has significantly strained the resources of the national healthcare system and public health system, but also from a macroeconomic point of view, it has affected consumption, investment, import and export trade to varying degrees, thus severely affecting GDP and even bringing about a global economic depression. At a macro level, there is a real tension between slowing progress of COVID-19 and seeking a rapid economic recovery. From a health perspective, a slow pace is an indication of success. From an economic perspective, a slower speed of spread, while blunting the health crisis, deepens business and economic losses.

From the perspective of citizens and individuals, the epidemic has limited employment and work, and increased the financial burden on families. Also, strategies such as social distancing become more difficult over time. Prolonged social isolation can impact mental, physical and spiritual health. We may protect ourselves from this virus, but negatively impact our quality of life and health in many other ways.

All policies and issues require a root cause of information, which is the analysis of the current status data of confirmed cases of the epidemic and the forecasts of data in short-term as well as long-term. A correct and reliable forecasting will greatly influence how policy makers formulate policies, and it will also influence households' expectations for the future. With the third wave of outbreak, Ontario, Quebec and British Columbia face more serious challenges, efficient models

for short-term forecasting are needed to forecast the number of future cases. This is what motivated me to analyze and forecast COVID-19 data for these selected Canadian provinces.

The aim of writing this honours project is not only because I have to complete undergraduate study and get a bachelor's degree of Mathematics, but also that I want to learn more about ARIMA model and its applications in our real lives. Through a semester of self-study and careful research, I have gained an in-depth understanding of statistics model, and I believe that it laid a solid foundation for my upcoming graduate career to study statistics and econometrics.

1.2 Literature Review

The dynamic forecasting of confirmed cases of COVID-19 has been extensively discussed in many publications over the last year (as of March 2021). A typical mathematical epidemiological model is established as a system of differential equations for susceptible infection-removal (SIR) sequences. the SIR model, i.e. the model for immune compromised diseases such as Covid-19, has been presented in many publications. Li-Pang et al. (2021) investigated prediction of the development of COVID-19 in Canada using several models, including SIR model. They focus attention on Canadian data and analyze three provinces, Ontario, Alberta, British Columbia, and Quebec, which have the most severe situations in Canada. To build predictive models and conduct prediction, they employed three models, smooth transition autoregressive (STAR) models, neural network (NN) models, and susceptible-infected-removed (SIR) models, to fit time series data of confirmed cases in the four provinces separately. However, they found that SIR model is not a good one to forecast for reasons of the incomplete assumption, that is the SIR model requires no inbound or outbound infected travellers, it assumes no asymptomatic cases, which is clearly untrue.

Moreover, some experts and researches use ARIMA models to forecast. Tadeusz (2020) used ARIMA (1,2,0) to predict the dynamics of confirmed COVID-19 cases for selected European countries at different stages of the development of the epidemic, i.e., at the first stage of development, when the maximum number of cases per day is reached, and at the stage of the

epidemic's extinction. Also, Benvenuto et. al. (2020) indicated that forecasts can be re-estimated every day because of daily updated COVID-19 databases. Thus, ARIMA models can be considered as an immediate and straightforward system for monitoring the epidemic at national and regional levels.

2 Methodology and Data

ARIMA models have been widely used in researches of statistics and econometrics, for instance, it has been applied in forecasting GDP (gross domestic product), inflation rates, interest rates, prices of stock as well as prediction of epidemics etc. ARIMA model is a typical time series model, which consists of three parts: AR model (autoregressive model) and MA model (moving average model), and the order I of the difference, so ARIMA is called autoregressive integrated moving average model. In this chapter, it will introduce what time series data is, ARIMA model and how to use this model to forecast data. Also, it will do a preliminary analysis of the raw data.

2.1 Time Series Analysis

In mathematics and statistics, a time series is a series of data points indexed (or listed or plotted) in chronological order. Time-series data show temporal variation: variation over period (years, months, weeks, days, hours, even seconds). Most commonly, a time series is one that is collected at continuous, equidistant time points. Therefore, it is a sequence of discrete time data.

According to "*Time Series Analysis and Its Applications with R Examples*", the main goal of time series analysis is to develop mathematical and statistical models that provide a reasonable description of the example data. In order to provide a statistical setup to characterize data that seem to fluctuate in a random manner over time, we assume that a time series can be defined as a collection of random variables indexed according to the order in which they are obtained in time (Shumway and Stoffer, 2017).

Commonly, a time series $(x_1, \ldots, x_e)$ is assumed to be a sequence of real values taken at successive equally spaced points in time, from time $t=1$ to time $t=e$.

Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communication engineering, and mainly in any field of applied science and engineering involving time measurement. Time series are very frequently plotted via temporal line charts, as shown in Figure.1 below:

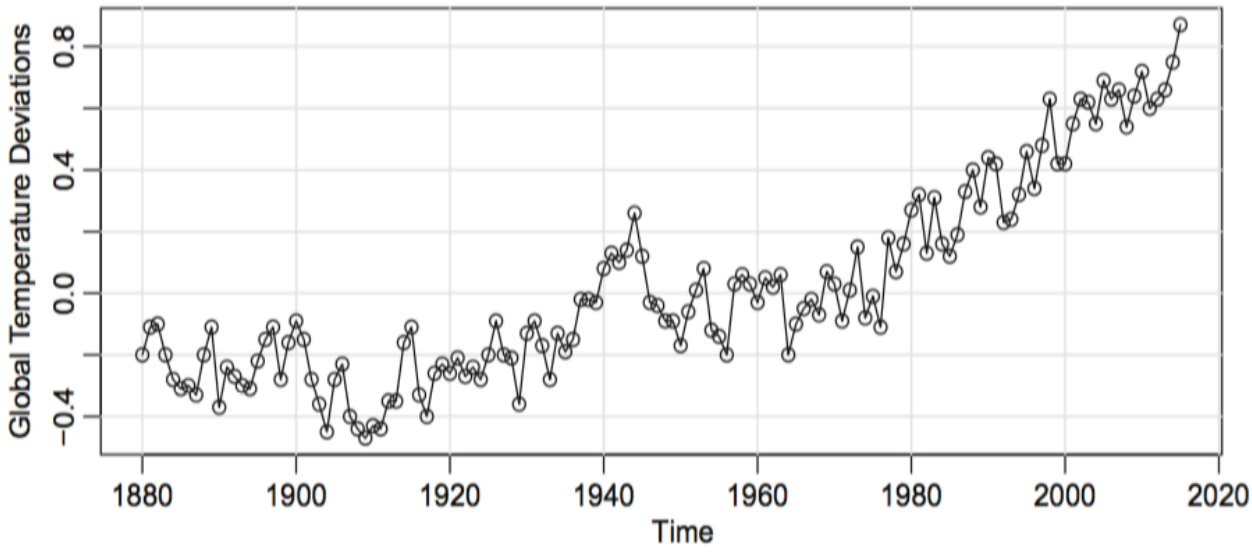


Fig. 1 early average global temperature deviations (1880–2015) in degrees centigrade

Figure.1 shows the global temperature series record. Data of global land-ocean mean temperature indices from 1880 to 2015 (with a base period of 1951-1980) is plotted via a temporal line. We note a clear upward trend in the series during the second half of the 20th century, which has been used as an argument for the global warming hypothesis. We also noticed a flat trend around 1935, and then a sharp rise around 1970.

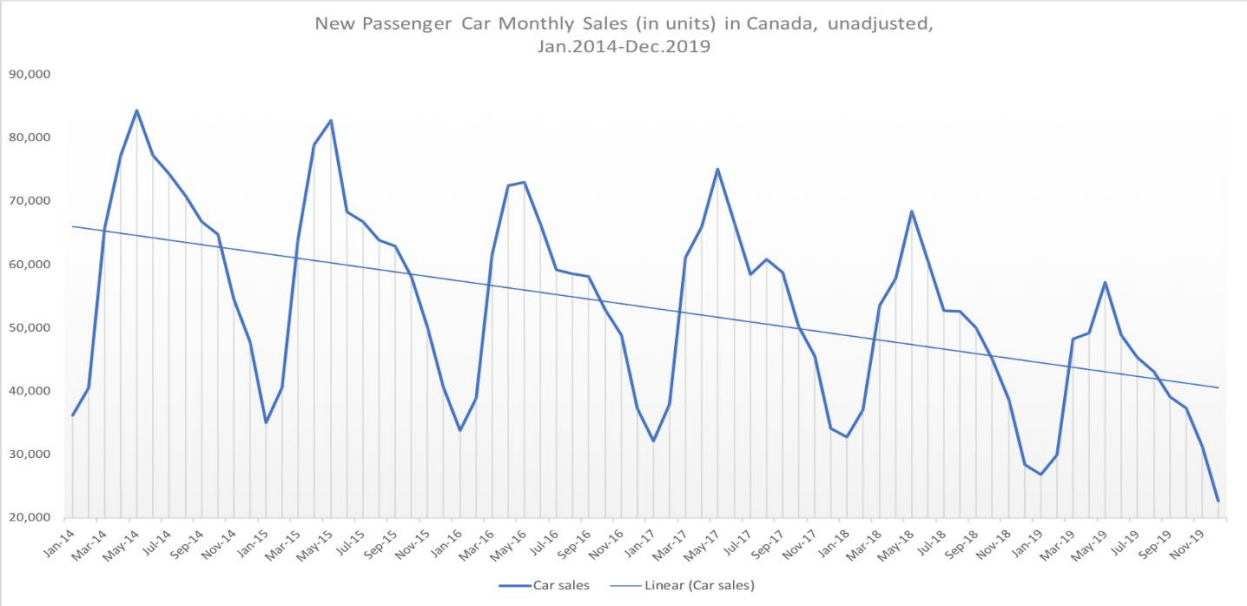


Fig. 2 New Passenger Car sales (in units) in Canada (Jan.2014-Dec.2019)

In Figure.2, Canadian monthly sales of new passenger cars are also plotted via a blue temporal line. Obviously, we can find that new passenger car sales show a downward trend over those five years, it also can be found by the linear trend line in Figure.2. Moreover, every year, the number of new passenger car sales shows an upward trend at the beginning of the year and reaches the largest sales volume of the year in May and June, and then car sales decreases in the second half of the year until the end of the year.

In order to determine the value of the variable at time $\boldsymbol{t}$ (U_t), we have to distinguish some various components, such as trend value (T_t), seasonal variation (S_t), cyclic variation (C_t) and random component (R_t). Book written by Hyndman and Athanasopoulos (2018) gives some explanations and examples of time series patterns, a trend exists when there is a long-term increase or decrease in the data. It could be linear, like trend line in Figure.2, but it does not have to be linear. A seasonal pattern occurs when a time series is affected by seasonal factors, such as the time of the year or the day of the week. Also, seasonality always has a fixed and known frequency. In the Figure.3, the monthly sales of antidiabetic drugs have seasonality which is induced by the change in the cost of the drugs at the end of the calendar year. Moreover, car sales in Figure.2 also have relatively regular seasonal patterns and they are visible to the naked eyes. A cycle occurs when the data exhibit rises and falls that are not of a fixed frequency. The

thing we have to notice is that if the fluctuations are not of a fixed frequency then they are cyclic patterns; if the frequency is unchanged and associated with some aspect of the calendar, then they are seasonal patterns.

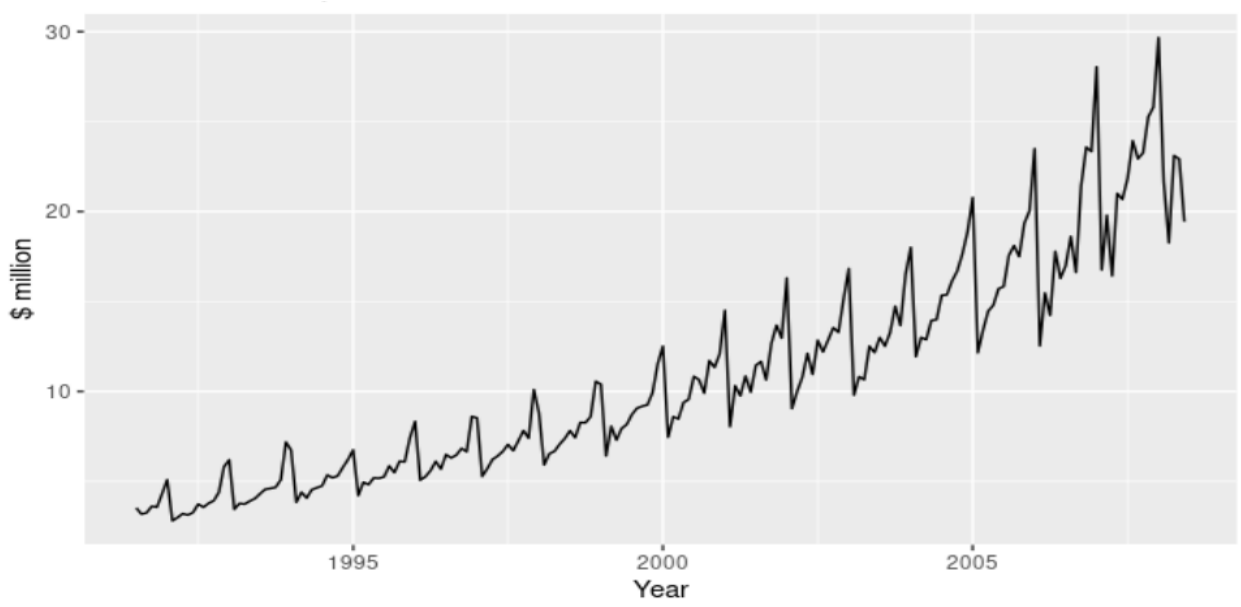


Fig. 3 Monthly sales of antidiabetic drugs in Australia

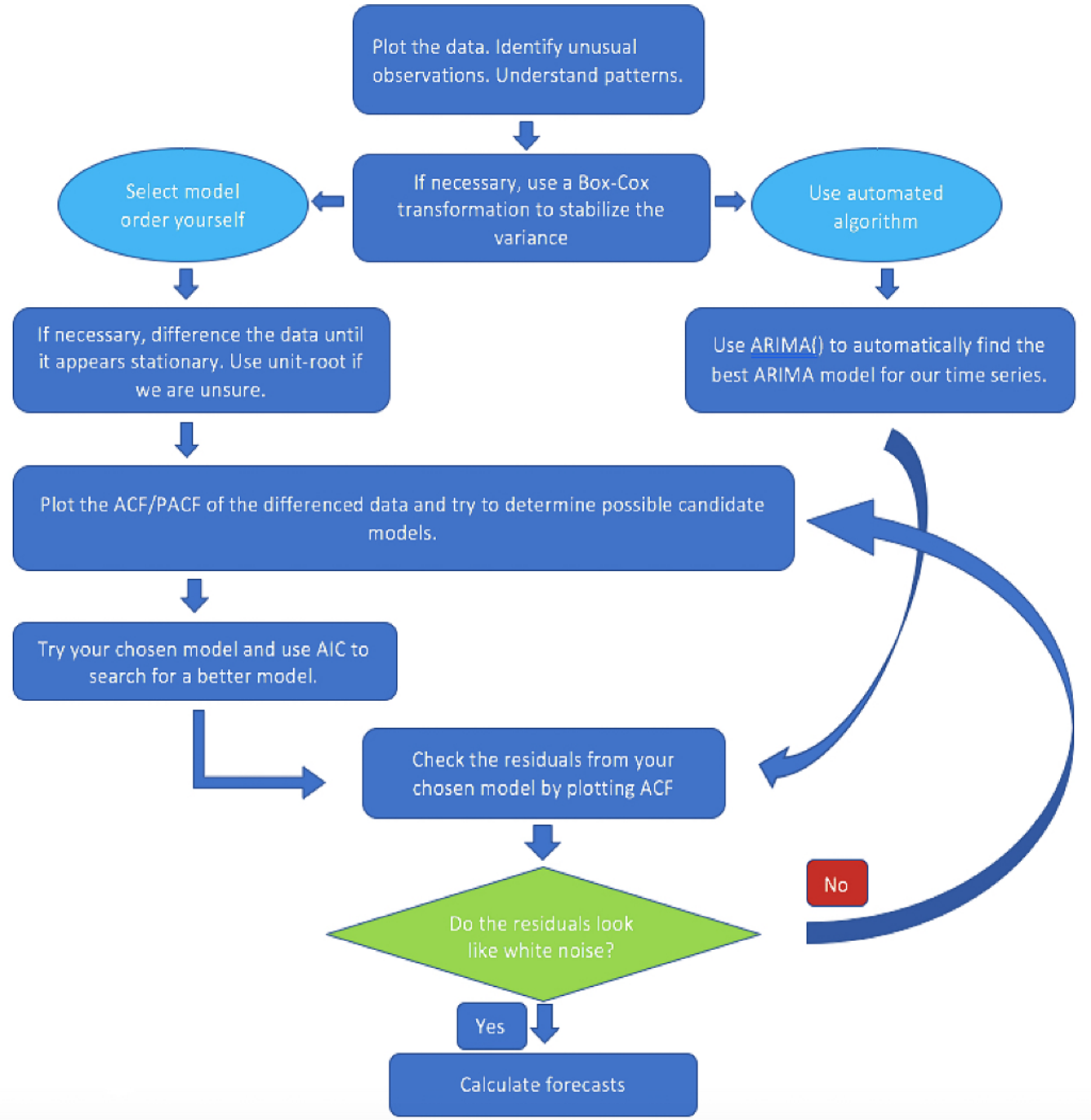
In time series model, if seasonal variations are constant around the trend, we use the additive method, that is, $U_t = T_t + S_t + C_t + R_t$; if seasonal and random fluctuations are a function of the trend, we use the multiplicative method, that is, $U_t = T_t * S_t * C_t * R_t$.

2.2 ARIMA models

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.
Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.
An ARIMA model is characterized by 3 terms: p, d, q where,
p is the order of the AR term
q is the order of the MA term

d is the number of differencing required to make the time series stationary.

The charts below are steps of ARIMA modelling:



In Section 2.2, it will introduce ARIMA (p, d, q) model, stationarity of time series, and auto-correlation function as well as partial auto-correlation function.

2.2.1 Stationarity and White Noise Process

A common assumption in the time series analysis is that the data are stationary.
A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. Stationarity can be defined in precise mathematical terms, but for our purpose we mean a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations (seasonality).

Intuitively, stationarity means that the statistical properties of the process do not change over time. However, several different notions of stationarity have been suggested in econometric literature over the years.

If the time series is not stationary, it can be transformed to stationarity with one of the following techniques:

1) We can difference the data. That is, given the series Z_t, we create the new series $Y_i = Z_i - Z_{i-1}$. The differenced data will contain one less point than the original data. Although you can difference the data more than once, one difference is usually sufficient.
2) If the data contain a trend, we can fit some type of curve to the data and then model the residuals from that fit. Since the purpose of the fit is to simply remove long term trend, a simple fit, such as a straight line, is typically used.

3) For non-constant variance, taking the logarithm or square root of the series may stabilize the variance. For negative data, you can add a suitable constant to make all the data positive before applying the transformation. This constant can then be subtracted from the model to obtain predicted (i.e., the fitted) values and forecasts for future points.

The stability of difference function refers to the convergence of data generated by difference function. For the general p-order difference function, the homogeneous function is

$$y_t - a_1 y_{t-1} - a_2 y_{t-2} - \cdots - a_p y_{t-p} = 0.$$

Suppose that a special solution of this chi-squared equation takes the form:

$$y_t^h = \lambda^t,$$

and substituted into the homogeneous function, then:

$$\lambda^t - a_1\lambda^{t-1} - a_2\lambda^{t-2} - \cdots - a_p\lambda^{t-p} = 0$$

A system of difference equations is stationary, if all of the roots of the characteristic equation fall within the unit circle.

A **strictly stationary** time series is one for which the probabilistic behavior of every collection of values $\{y_{t1}, y_{t2}, ..., y_{tk}\}$ is identical to that of the time shifted set $\{y_{t1+h}, y_{t2+h}, \cdots, y_{tk+h}\}$.

Pr $\{y_{t1} \le c_1, ..., y_{tk} \le c_k\}$ = Pr $\{y_{t1+h} \le c_1, ..., y_{tk+h} \le c_k\}$

That is, for all k = 1,2, ..., all time points $t_1, t_2, ..., t_k$, all numbers $c_1, c_2, ..., c_k$, and all time shifts h = 0, ±1, ±2, ...

Weakly stationarity: also called covariance-smooth or second-order smooth, for a random time series y_t, it is called a weakly stationary random variable if its expectation, variance, and covariance do not vary with time *t*, i.e., for all times, it must satisfy the following conditions:

1) $E(y_t) = \mu$, which is constant;
2) $Var(y_t) = \sigma^2$, which is constant;
3) $\gamma_j = E([y_t - \mu][y_{t-j} - \mu]), j = 0, \pm1, \pm2, L$, where γ_j is the autocovariance function of this stationary time series.

According to Palachy (2019), weakly stationary processes are painted specific pictures with constant mean and variance. Their properties are contrasted nicely with those of their counterparts in Figure 4 below,

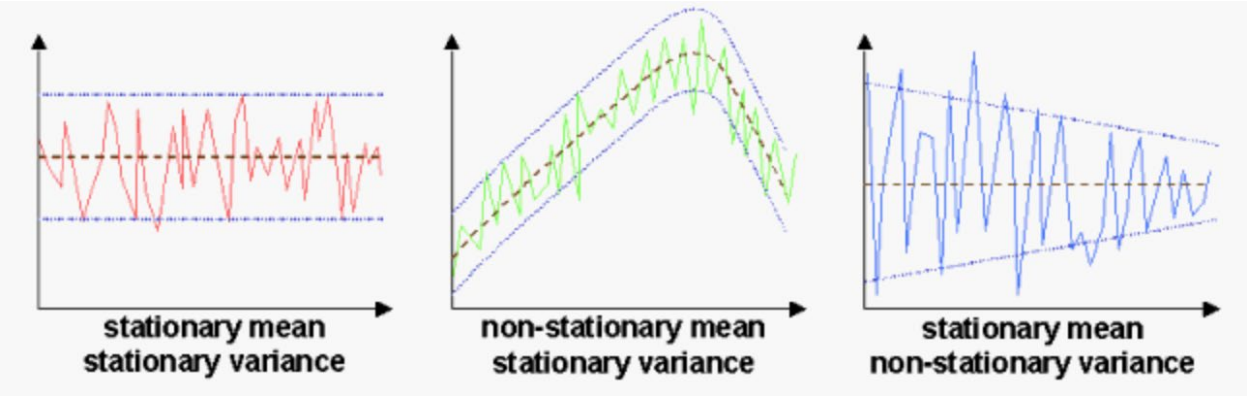


Fig.4 Contrasts of different types of weakly stationarity

White noise process: if a random process is called a white noise process, in which all random sequences are independent of each other, with a mean of 0 and a constant variance. Namely, for all times t and y, if the following conditions are satisfied, it is a white noise process:

1) $E(y_t) = 0$;
2) $Var(y_t) = \gamma_0 = \sigma^2$.

2.2.2 AIC and BIC

Akaike's Information Criterion (AIC)

$$\text{AIC} = \log \hat{\sigma}_k^2 + \frac{n+2k}{n},$$

where $\hat{\sigma}_k^2$ *is* $\frac{SSE(k)}{n}$ *and k is the number of parameters in the model.*

The value of *k* yielding the minimum AIC specifies the best model. The idea is roughly that minimizing $\hat{\sigma}_k^2$ would be a reasonable objective, except that it decreases monotonically as *k* increases. Therefore, we ought to penalize the error variance by a term proportional to the number of parameters.

Bayesian Information Criterion (BIC)

$$\text{BIC} = \log \hat{\sigma}_k^2 + \frac{k \log n}{n},$$

BIC is also called the Schwarz Information Criterion (SIC), notice that the penalty term in BIC is much larger than in AIC, consequently, BIC tends to choose smaller models. Various simulation studies have tended to verify that BIC does well at getting the correct order in large samples, whereas AIC tends to be superior in smaller samples where the relative number of parameters is large.

2.2.3 ACF and PACF

The **autocorrelation function (ACF)** of a stationary time series will be written as

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}} = \frac{\gamma(h)}{\gamma(0)}.$$

The Cauchy–Schwarz inequality shows again that $-1 \leq \rho(h) \leq 1$ for all h, enabling one to assess the relative importance of a given autocorrelation value by comparing with the extreme values −1 and 1.

In statistics, ACF can help in indicating whether differencing is needed. Because the polynomial $\varphi(z)\,(1 - z)^d$ has a unit root, the sample ACF, $\hat{\rho}(h)$, will not decay to zero fast as h increases. Thus, a slow decay in $\hat{\rho}(h)$ is an indication that differencing may be needed.

In time series analysis, the **partial autocorrelation function (PACF)** gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags. Basically, instead of finding correlations of present with lags like ACF, it finds correlation of the residuals (which remains after removing the effects which are already explained by the earlier lag(s)) with the next lag value. Therefore, if there is any hidden information in the residual which can be modeled by the next lag, we might get a good correlation and we will keep that next lag as a feature while modeling.

The use of this function was introduced as part of the Box–Jenkins approach to time series modelling, whereby plotting the partial autocorrelation functions one could determine the appropriate lags p in an AR (p) model or in an extended ARIMA (p, d, q) model.

When preliminary values of d have been settled, the next step is to look at the sample ACF and PACF of $\nabla^d x_t$ for whatever values of d have been chosen. Note that it cannot be the case that both the ACF and PACF cut off. Because we are dealing with estimates, it will not always be clear whether the sample ACF or PACF is tailing off or cutting off. Also, two models that are seemingly different can actually be very similar. With this in mind, we should not worry about being so precise at this stage of the model fitting. At this point, a few preliminary values of p, d, and q should be at hand, and we can start estimating the parameters.

2.2.4 AR Model and MA model

An **autoregressive model of order p**, abbreviated **AR(p)**, is of the form

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + w_t,$$

where y_t is stationary, $w_t \sim$ white noise $(0, \sigma_w^2)$, and $\varphi_1, \varphi_2, \ldots, \varphi_p$ are constants ($\varphi_p \neq 0$). The mean of y_t is zero. If the mean, μ, of y_t is not zero, replace y_t by $y_t - \mu$,

$$y_t - \mu = \varphi_1 (y_{t-1} - \mu) + \varphi_2 (y_{t-2} - \mu) + \cdots + \varphi_p (y_{t-p} - \mu) + w_t,$$

or write $y_t = \alpha + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + w_t$, where $\alpha = \mu (1 - \varphi_1 - \cdots - \varphi_p)$. And the **autoregressive operator** is defined to be $\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$.

Figure.5 shows a time plot of two AR (1) processes, one with $\varphi = 0.9$ and one with $\varphi = -0.9$; in both cases, $\sigma_w^2 = 1$.

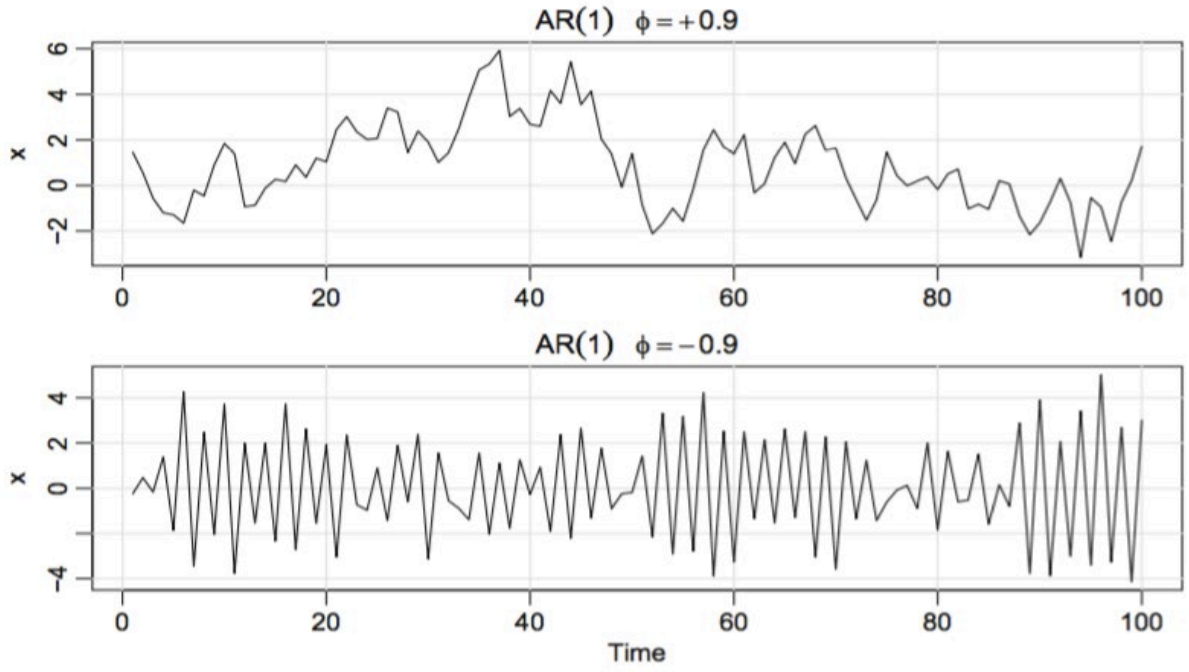


Fig.5 Simulated AR (1) models: $\varphi = 0.9$ (top); $\varphi = -0.9$ $(bottom)$

The **moving average model of order q**, or **MA(q)** model, is defined to be

$$y_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q},$$

where $w_t \sim$ white noise $(0, \sigma_w^2)$, and $\theta_1, \theta_2, \ldots, \theta_q$ ($\theta_q \neq 0$) are parameters.

The system is the same as the infinite moving average defined as the linear process, where ψ0 = 1, ψj = θj, for j = 1, ..., q, and ψj = 0 for other values. We may also write the MA(q) process in the equivalent form

$$y_t = \theta(\mathrm{B})\, w_t.$$

The **moving average operator** is θ(B) = 1 + θ_1B + $\theta_2 B^2$ +⋯+$\theta_q B^q$.

Unlike the autoregressive process, the moving average process is stationary for any values of the parameters $\theta_1, \ldots, \theta_q$.

Figure 6 shows a time plot of two MA (1) processes, one with $\theta = 0.9$ and one with $\theta = -0.9$; in both cases, σ_w^2 =1.

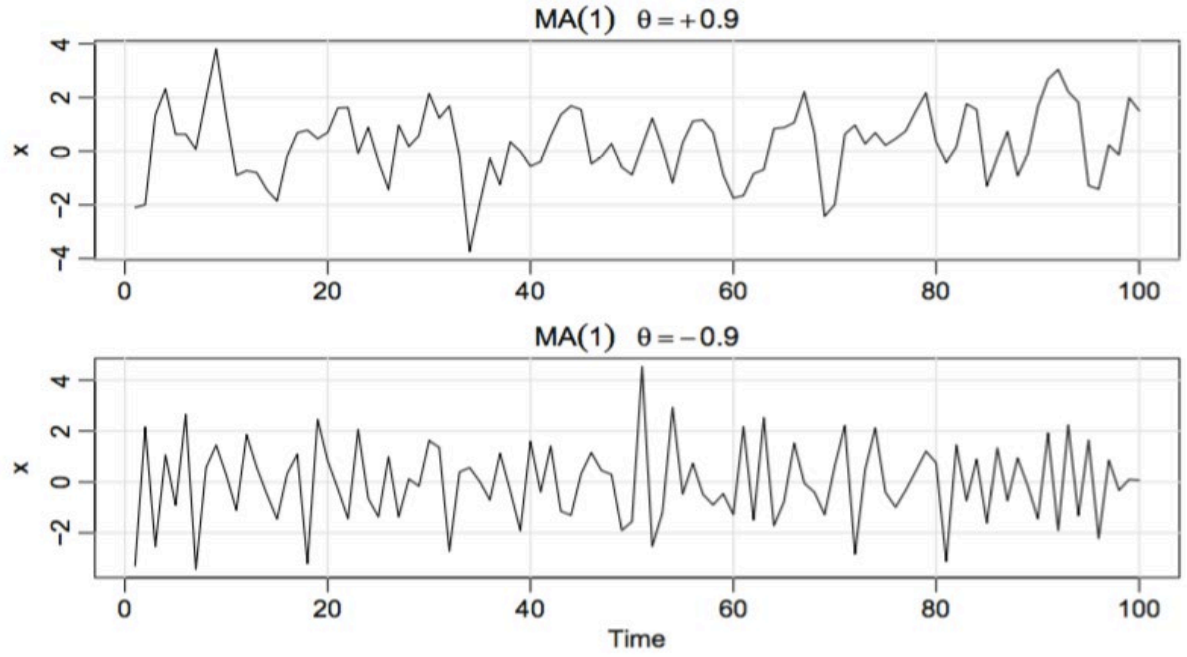


Fig.6 Simulated MA (1) models: $\theta = 0.9$ (top); $\theta = -0.9$ $(bottom)$

The following table shows how to choose order of models by using ACF and PACF plots:

	AR (p) Model	MA (q) Model	ARMA (p, q) Model
ACF	Tails off	Cutting off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

2.2.5 ARIMA Model

If we combine differencing with autoregression and a moving average model, we obtain a **non-seasonal ARIMA model**. ARIMA is an acronym for Auto−regressive Integrated Moving Average (in this context, "integration" is the reverse of differencing). The full model can be written as

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

where y'_t is the differenced series and it may has been differenced by once, twice or more times. The "predictors" on the right-hand side include both lagged values of y_t and lagged errors. We call this an **ARIMA (p, d, q) model**, where

p = order of the autoregressive part;

d = degree of first differencing involved;

q = order of the moving average part.

Also, full model equation can be written in backshift notation as

$$(1 - \phi_1 \mathrm{B} - \cdots - \phi_p B^p)\,(1 - \mathrm{B})^d y_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q)\mathcal{E}_t.$$

The first term refers to AR (p), the second term refers to d differences, and right-hand side refers to MA (q). However, we can also use function **auto.arima()** in R to select appropriate values for p, d, q automatically.

The same stationarity and invertibility conditions that are used for autoregressive and moving average models also apply to an ARIMA model. Moreover, there are some special cases of ARIMA models which are

White noise	ARIMA (0, 0, 0)
Random walk	ARIMA (0, 1, 0) with no constant
Random walk with drift	ARIMA (0, 1, 0) with a constant
Autoregression	ARIMA (p, 0, 0)
Moving Average	ARIMA (0, 0, q)

2.3 Data Source and Initial Analysis

Data used and forecasted in this project is daily reported confirmed cases in Ontario, Quebec and British Columbia. Since the number of Canadian total confirmed COVID-19 cases is too large and accumulative, it needs more information and strict model to forecast. Otherwise, we always get results of increasing trend, because how and when the number of current cases will decline is essential part for analysis. Thus, we may consider that forecasts of the speed of increasing cases are needed, which is necessary for long-term or short-run analysis if we use accumulative data. Using daily confirmed cases is more visual and intuitive to see how the situation of epidemic change in the near future for selected Canadian provinces.

All of data are downloaded from https://www.ctvnews.ca/health/coronavirus/tracking-every-case-of-covid-19-in-canada-1.4852102, https://ici.radio-canada.ca/info/2020/coronavirus-covid-19-pandemie-cas-carte-maladie-symptomes-propagation/index-en.html https://ici.radio-canada.ca/info/2020/09/covid-19-pandemie-cas-deces-propagation-vague-maladie-coronavirus/index-en.html. Some information and background are obtained from each provincial government page about COVID-19.

In the Figure.7, time series plots of the number of daily confirmed cases in Ontario shows a surge trend patterns, not a horizontal pattern. At the end of year of 2020, the number of daily confirmed cases reached the maximum peak which was over 4000 confirmed cases in a day. From February of 2021 until now, we can easily find there has been the third wave of epidemic. From an overall perspective, linear trend line for the date of Ontario is upward sloping all the time.

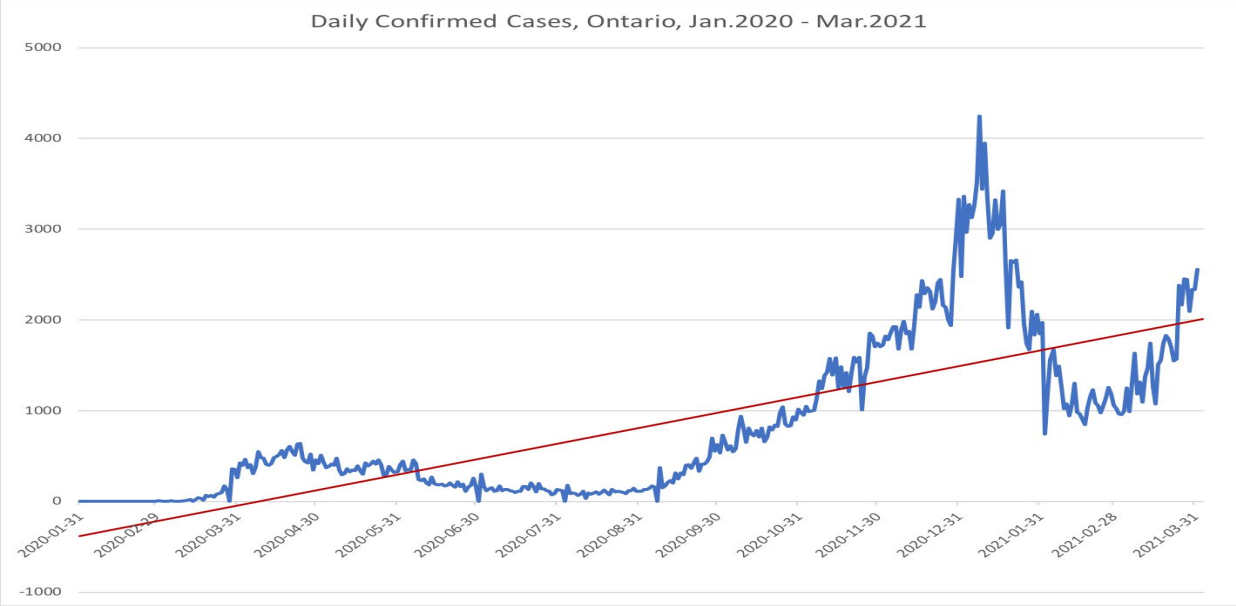


Fig.7 Time Series plot of daily confirmed cases for Ontario (Jan.2020 – Mar.2021)

In the first half of 2020, the epidemic was somewhat more severe in Quebec than in Ontario, especially back at the end of April, when the number of confirmed cases per day in Quebec exceeded 2,000 for the first time, as depicted in Figure.8. It's not the only case that Quebec also reached the maximum of daily confirmed cases at the end of year of 2020. Moreover, the plot showed a rebounding trend from the middle of March of 2021.

For British Columbia, there are some obvious differences compared with Ontario and Quebec. Time series plot of the number of daily confirmed cases in British Columbia has visual seasonal patterns whose cycle is about a week, as shown in the Figure.9.

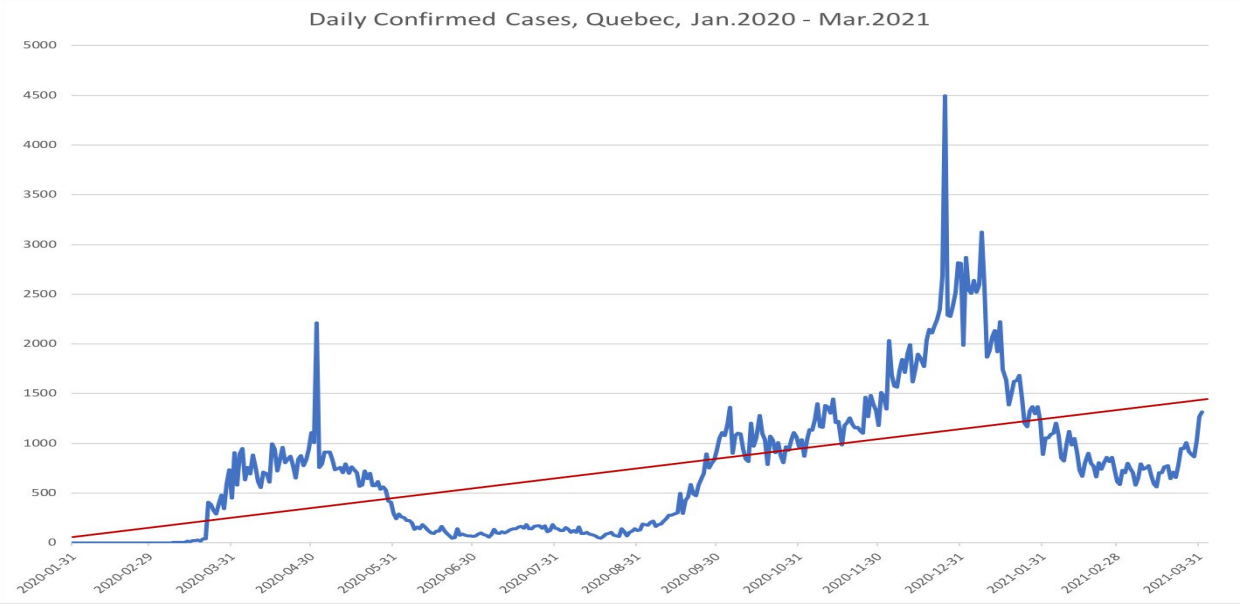


Fig.8 Time Series plot of daily confirmed cases for Quebec (Jan.2020 – Mar.2021)

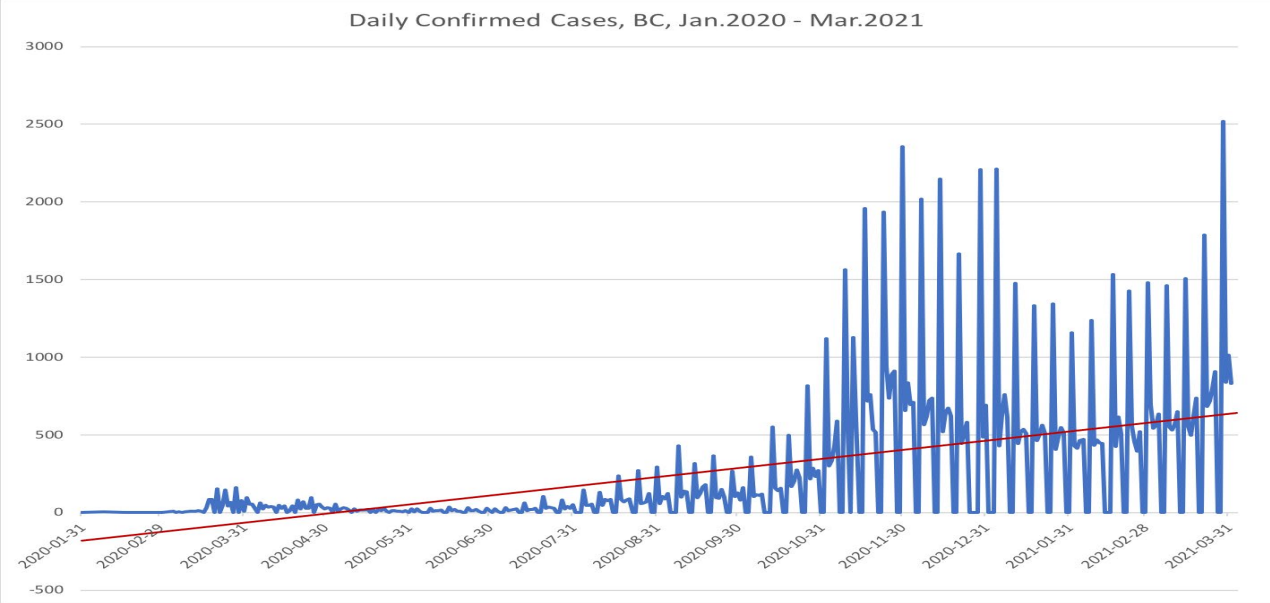


Fig.9 Time Series plot of daily confirmed cases for British Columbia (Jan.2020 – Mar.2021)

3 Applications: ARIMA-based Forecasts

Firstly, we use **ts()** function to transform these data as time series analysis. For each province, we view its time series plot. In this part, we will not show time series plots again in this section, they are the same as plots in the methodology section. We could see that time series is not stationary. Thus, difference equation will be applied to make them stationary. Secondly, we use **diff()**

function and **plot.ts()** function to see how the residuals look like. By observing Figure.10 below, it is not hard to find that residuals are more stable, which also shows variances fluctuates above and below the line of zero and become much more stable after the first order of difference. Thus, we can get that order of first differencing, **d**, could be 1.

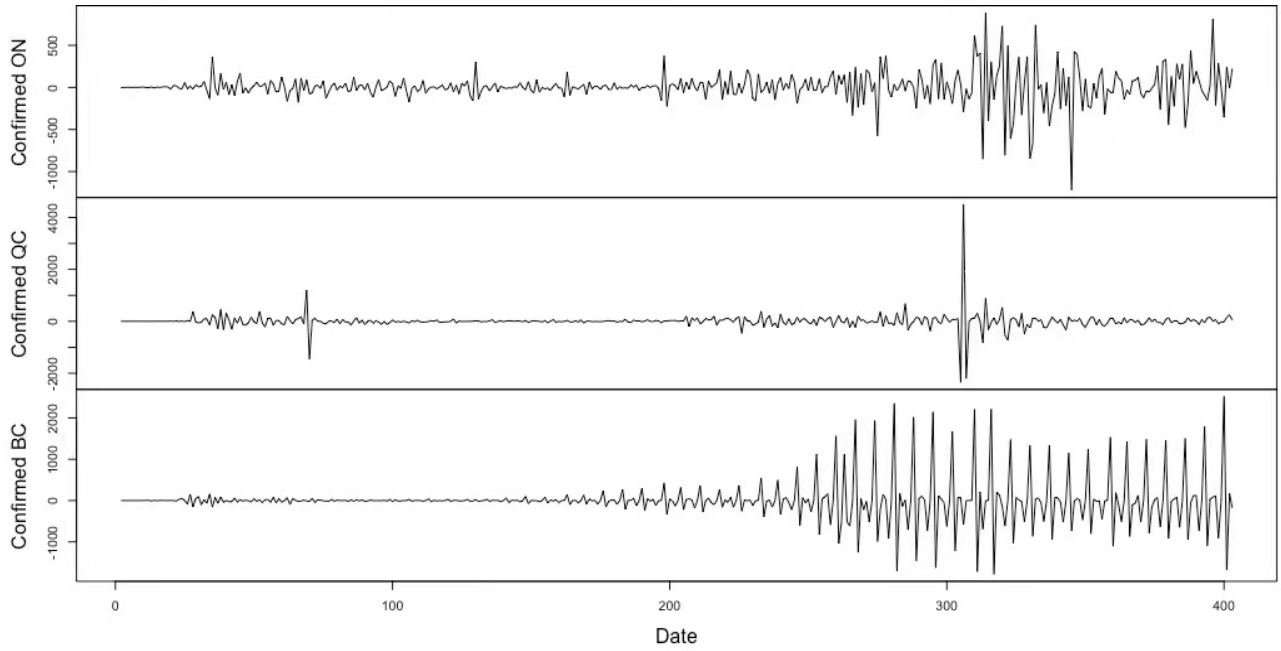


Fig.10 Residuals of the 1st differenced daily confirmed cases, ON, QC, BC (in R)

By observing autocorrelation function plots of the differenced data (Figure.11-Figure.13) below, we can determine orders of the autoregressive part, that is, $p_{ON} = 3,\ p_{QC} = 1,\ p_{BC} = 2$.

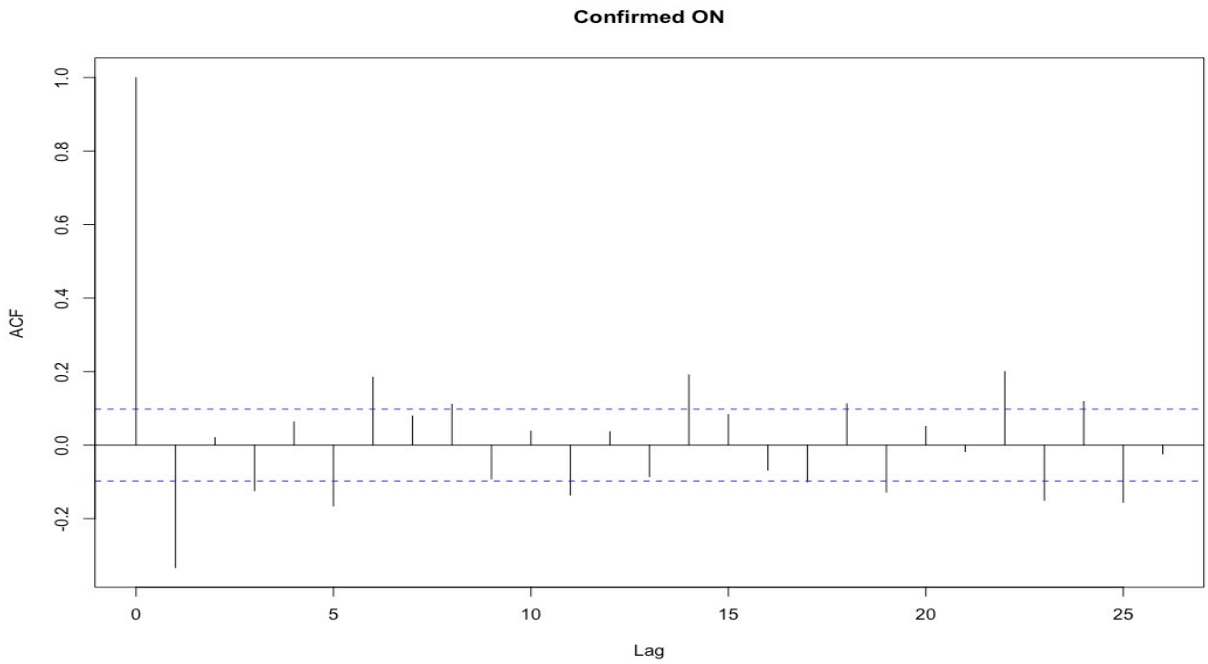


Fig.11 ACF plot of the 1st differenced daily confirmed cases, ON

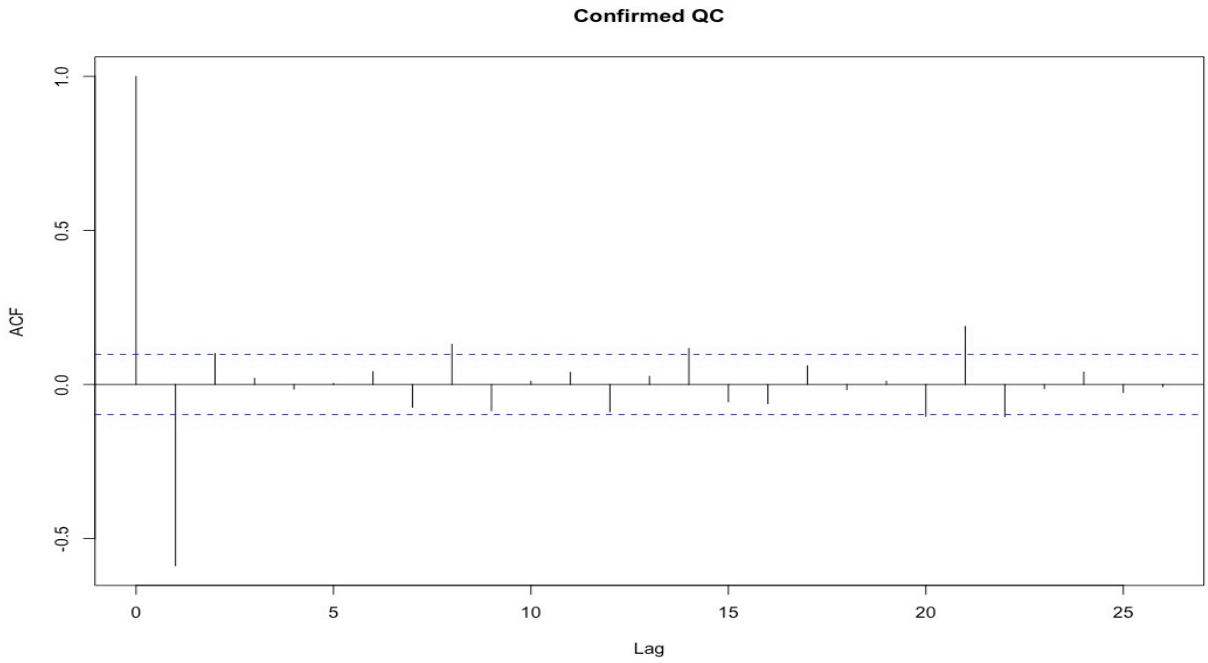


Fig.12 ACF plot of the 1st differenced daily confirmed cases, QC

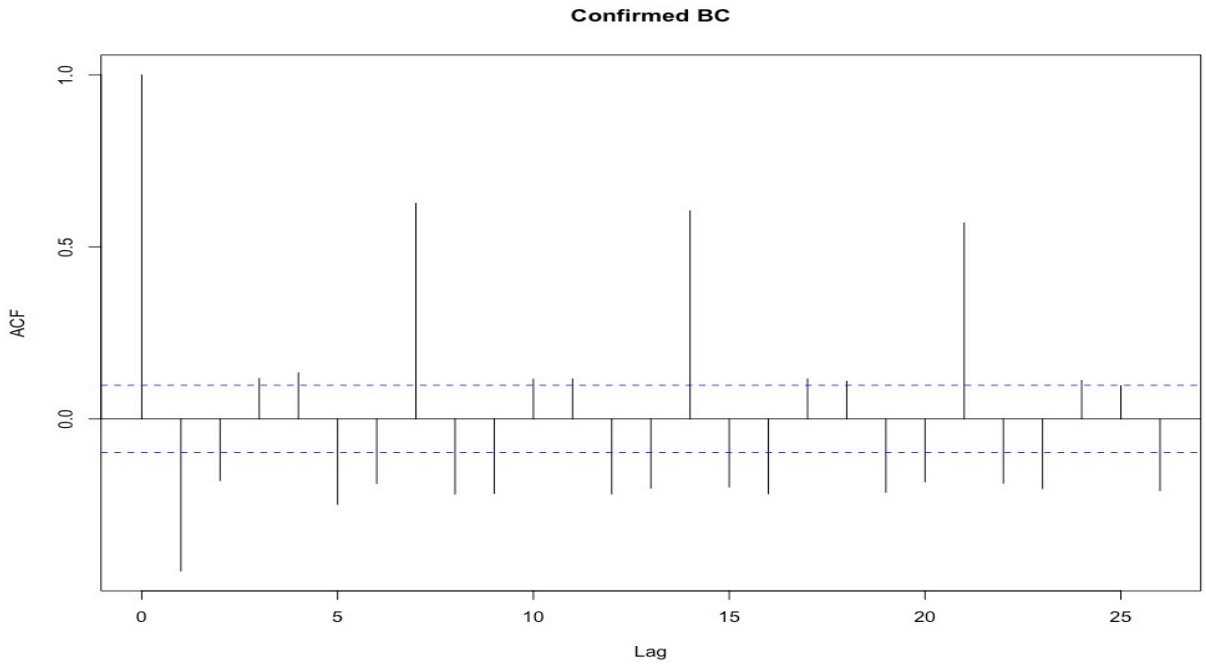


Fig.13 ACF plot of the 1st differenced daily confirmed cases, BC

By observing the partial autocorrelation plots of the differenced data (Figure.14-Figure.16) below, we can determine orders of the moving average part, that is, $q_{ON} = 3, q_{QC} = 3, q_{BC} = 3$.

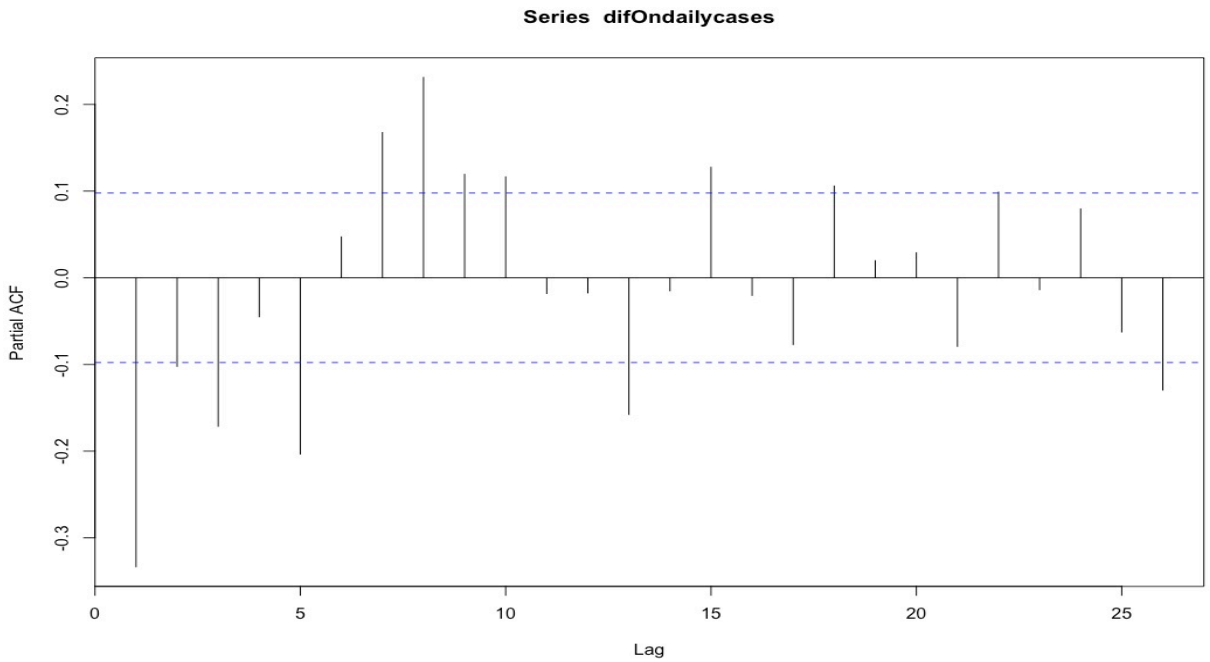


Fig.14 PACF plot of the 1st differenced daily confirmed cases, ON

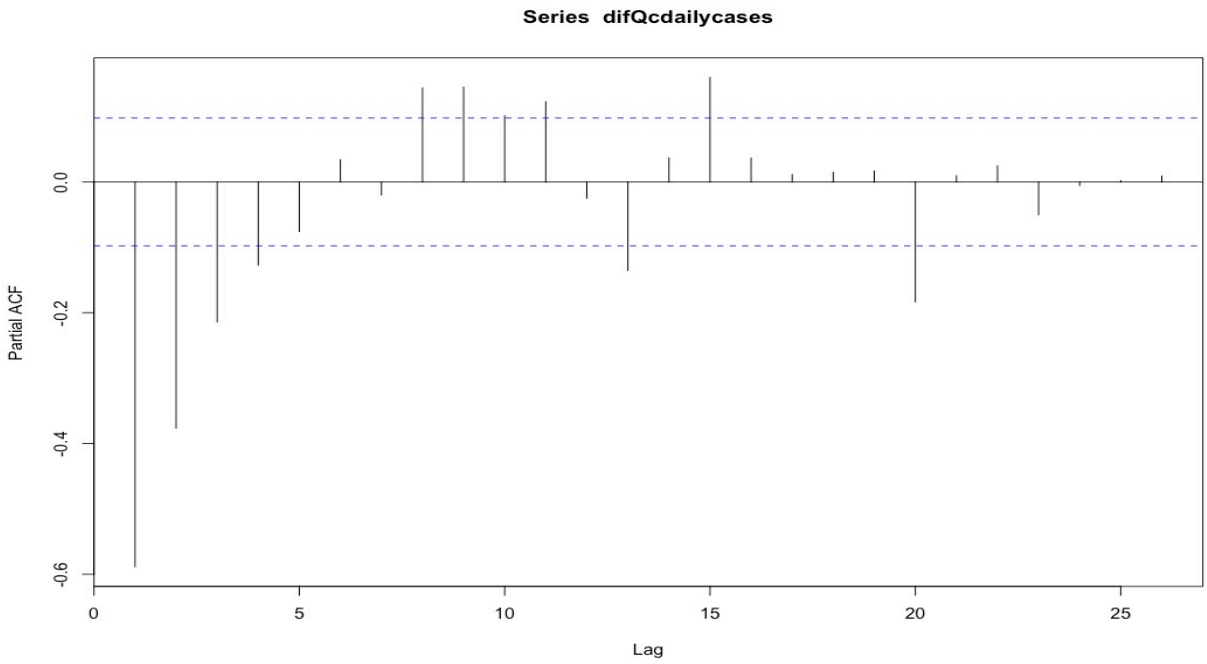


Fig.15 PACF plot of the 1st differenced daily confirmed cases, QC

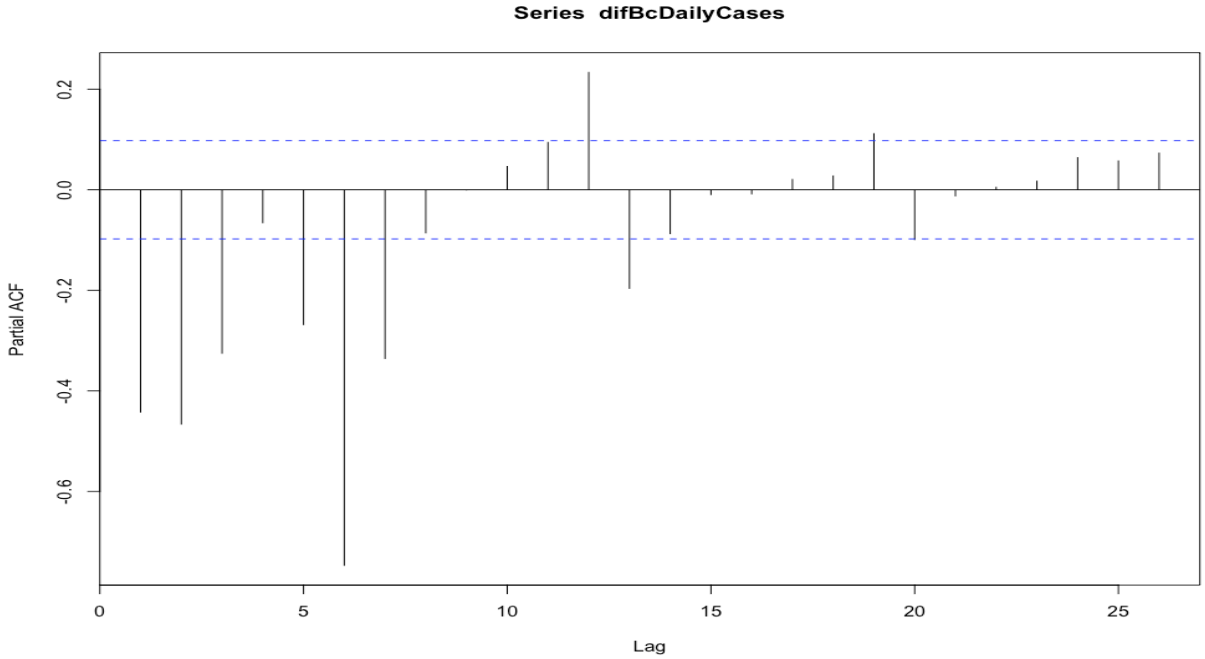


Fig.16 PACF plot of the 1st differenced daily confirmed cases, BC

In this paper, **auto.arima()** function is used. We can use "xxx=auto. arima (sample)" to generate the most appropriate and the best model for data of the number of COVID-19 daily confirmed cases for Ontario, Quebec and British Columbia. For Ontario, its ARIMA model should be

ARIMA (3,1,3); For Quebec, it is ARIMA (1,1,3); For British Columbia, it is ARIMA (2,1,3), these results are consistent with findings from auto-correlation function plots and partial auto-correlation function plots. Then, **forecast()** function is used to generate forecasting based on different ARIMA models, that is, "xxx<- forecast (model name, h, significance level)". Figure.17 to Figure.19 are the results for forecasts of the number of daily confirmed cases in 40 days at the significance level of 5%.

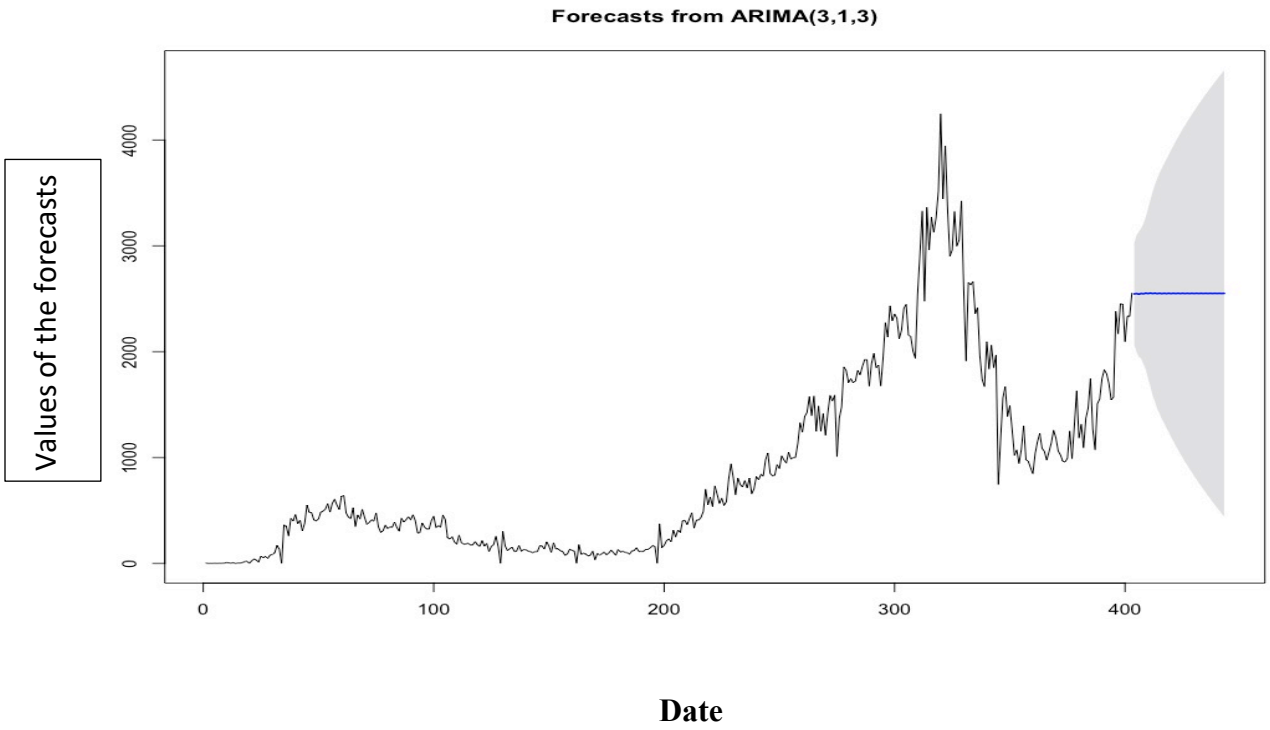


Fig.17 Forecasts of daily confirmed cases from ARIMA (3,1,3), ON

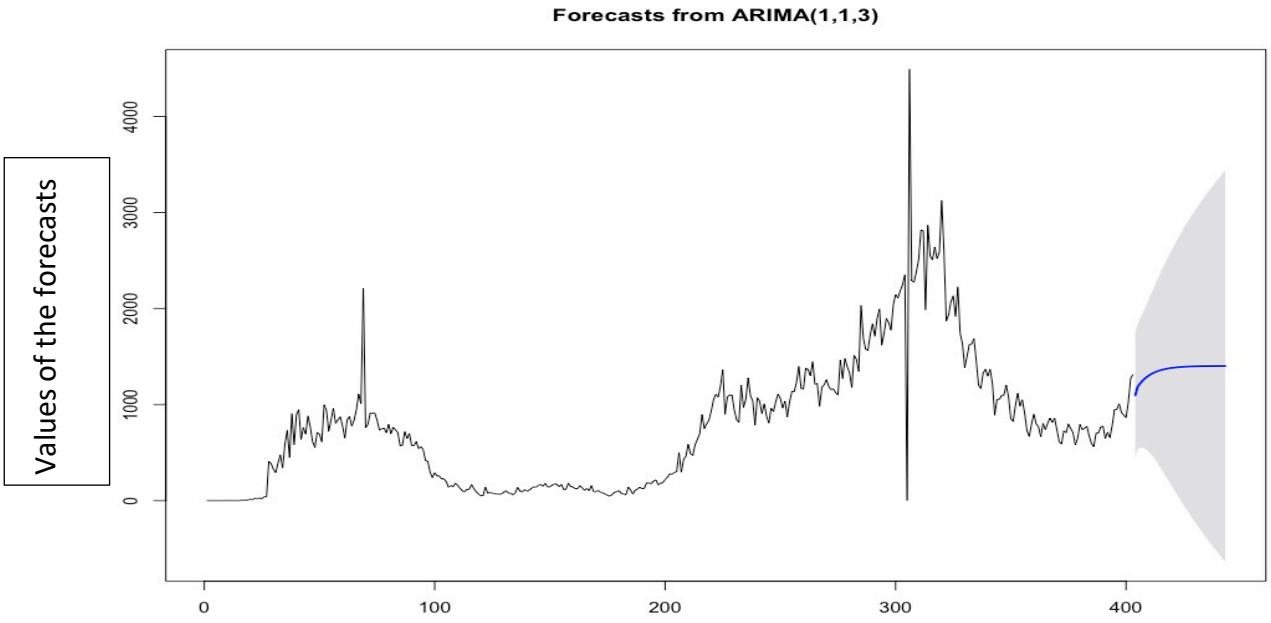


Fig.18 Forecasts of daily confirmed cases from ARIMA (1,1,3), QC

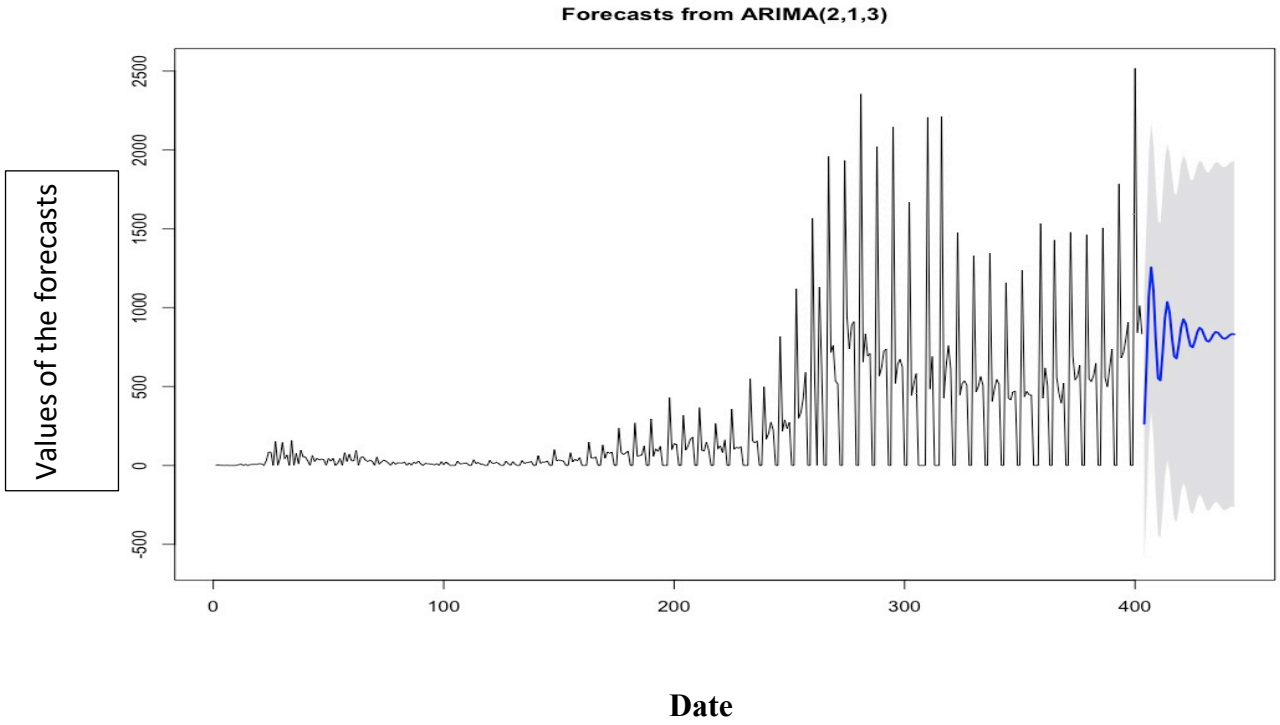


Fig.19 Forecasts of daily confirmed cases from ARIMA (2,1,3), BC

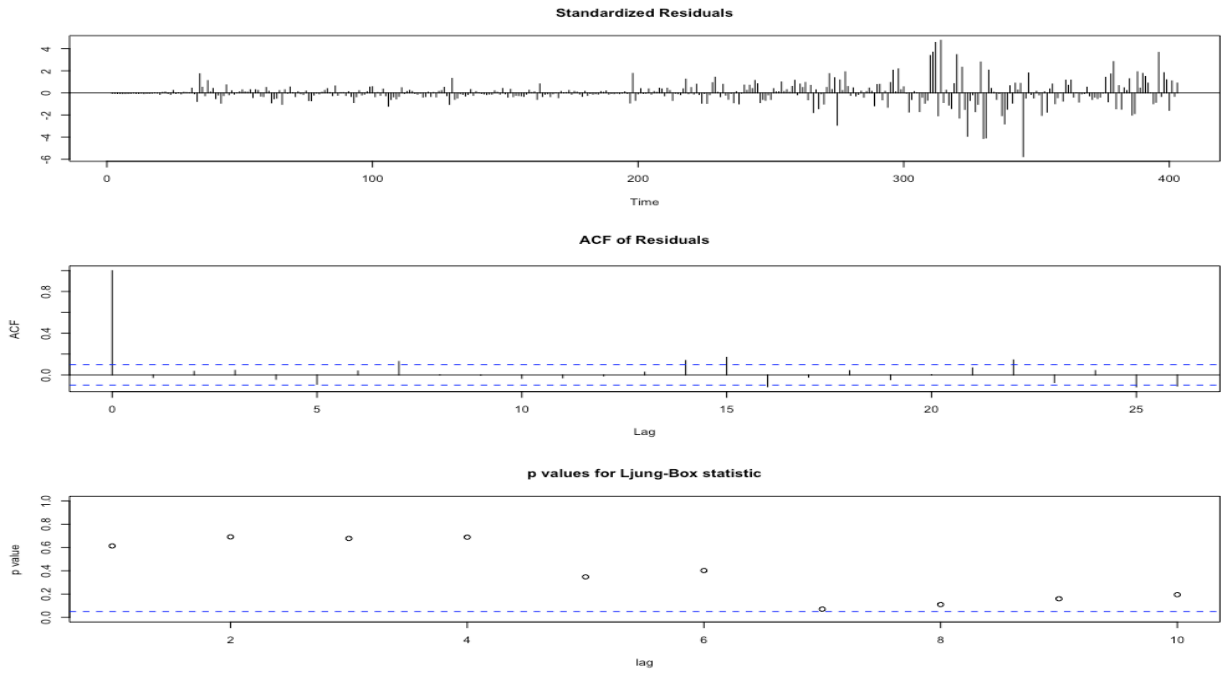


Fig.20 The diagnostics test of ARIMA (3,1,3) for Ontario

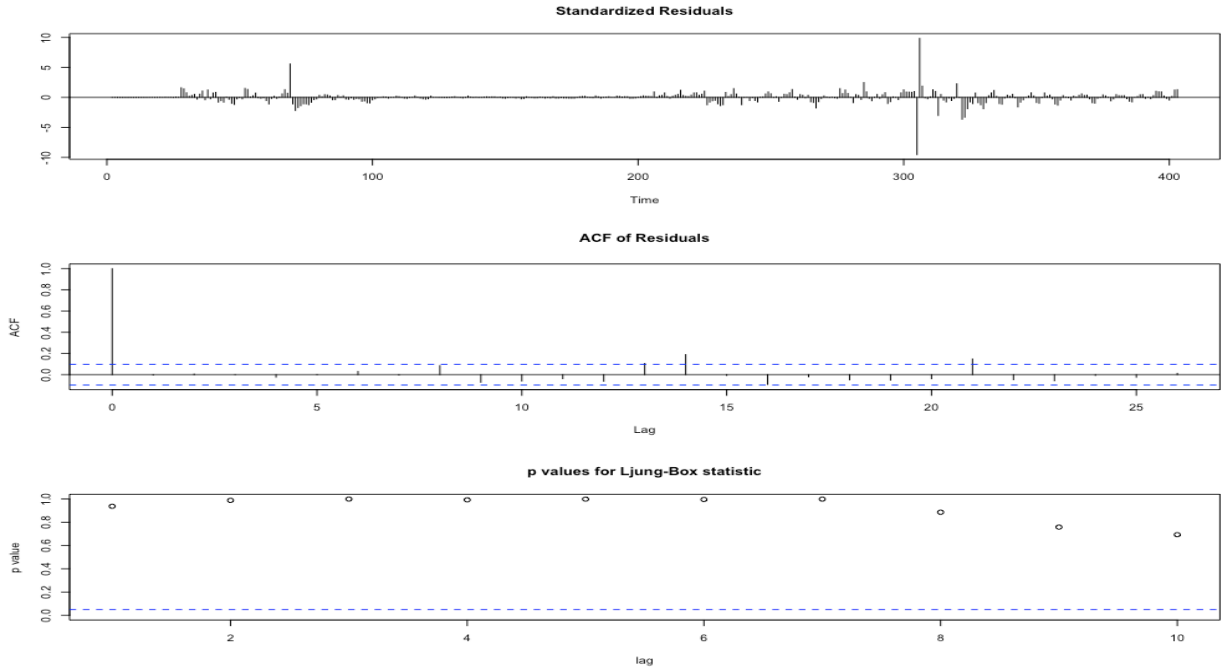


Fig.20 The diagnostics test of ARIMA (1,1,3) for Quebec

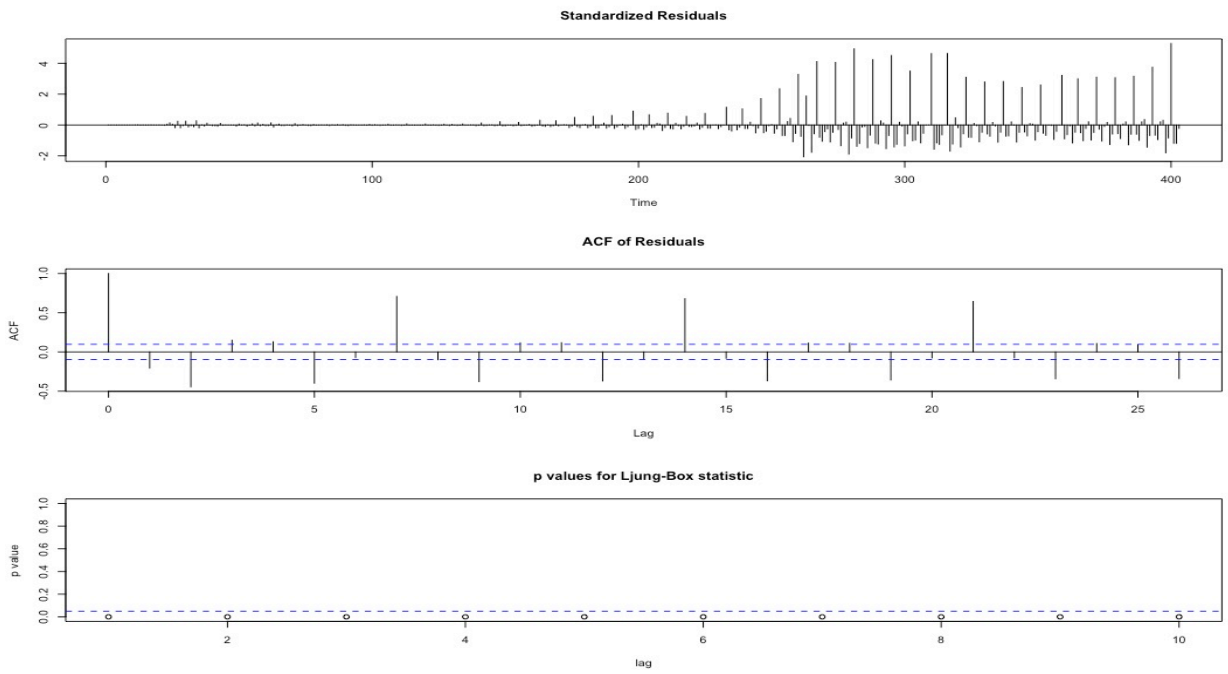


Fig.22 The diagnostics test of ARIMA (2,1,3) for British Columbia

The diagnostics test of ARIMA models illustrated in Figure.19 to Figure.22, including standardized residuals, autocorrelation function of residuals and p-values for Ljung-Box test. As we know, the Ljung-Box chi-squared statistic and the autocorrelation function of the residuals are used to determine whether the model meets the assumptions that the residuals are independent. If the assumption is not met, the model may not fit the data and you should use caution when you interpret the results. In these diagnostics tests generated in R, we have p-values of Ljung-Box test instead of chi-squared statistic. Through observing the third plot of each diagnostics test, points of p-values for Ontario and Quebec are above the line of significance level, 5%, so p-values are greater than the significance level, α. Thus, null hypothesises cannot be rejected at $\alpha = 5\%$ and ARIMA models meet the assumptions that the residuals are independent. However, for British Columbia, all of p-values of Ljung-Box test are below the line of significance level, 5%, so p-values are less than the significance level, α. Thus, null hypothesis is rejected at $\alpha = 5\%$ and ARIMA model may not meet the assumptions that the residuals are independent.

Also, the ACF of residuals, for Ontario and Quebec, are within the critical regions, which indicated that there is no correlation in the residual series. Thus, it is consistent with the assumption that the variances are constant. Thus, we could conclude that ARIMA (3,1,3) for Ontario and ARIMA (1,1,3) are both good fits. However, British Columbia still has many ACF of residuals are not within the critical region. WHY? It will be discussed in the last chapter.

4 Discussion and Conclusion

4.1 Conclusion and Comparation

In this project, we attempted to forecast the daily COVID-19 confirmed cases for selected Canadian provinces, including Ontario, Quebec and British Columbia. Autoregressive Integrated Moving Average (ARIMA) models are applied to analyze and investigate the trends of daily confirmed cases. In the last chapter, we use **auto.arima()** techniques to determine the most appropriate model for each forecast of different provinces. Also, we can easily find that in 40 days, Ontario will have fluctuations about 2,300 confirmed cases for each single day, based on ARIMA (3,1,3) at confidence level of 99.5%, and Quebec will have a slight upward-sloping trend, the number of daily confirmed cases will increase in the short term based on ARIMA (1,1,3), but it will not be as bad as the Christmas'. However, ARIMA (2,1,3) for British Columbia is not ideal or the most appropriate model by checking and testing the accuracy of the model.

4.2 Limitation and Imperfection

However, there are some limitations and imperfections in this research. First of all, time series of British Columbia seems like seasonal patterns, so using seasonal autoregressive integrated moving average (SARIMA) model could be more appropriate than using the ARIMA model, then it can get more accurate results for the number of daily confirmed cases in the near future. For Ontario and Quebec, they do not have perfect and small number for AIC and BIC, although their residuals are independent. Moreover, Canada had not entered the stage of a full-blown

outbreak before April of 2020, so we may use recent data, just including between the first wave and the third wave. Separating data from different periods after an outbreak for further study will reduce the impact of iterations between periods and lead to better short-term predictions.

4.3 Further Research

It seems that no single model can predict the data with perfect accuracy, and we may need a combination of two or several models to be used. According to research by Chakraborty and Ghosh (2020), for the COVID-19 datasets, we can propose a hybridization of stationary ARIMA and nonstationary WBF model to reduce the individual biases of the component models. The COVID-19 cases datasets for different provinces are various and complex in nature, like their epidemic did not start at the same time and the numbers of imported cases are different for these provinces. Thus, the ARIMA model may fail to produce random errors or even stationary residual series for each province. Moreover, the behavior of the residual series generated by ARIMA model is mostly oscillatory and periodic; thus, we could choose the wavelet function to model the remaining series. Several hybrid models based on ARIMA and neural networks are available in the field of time series forecasting. Thus, in the future research, we could use this better model to generate predictions of COVID-19 data. Moreover, as we mentioned in the first chapter, this epidemic has brought economic recession to the whole economy of all the world. In the further research, we could generate forecasts for inflation and gross domestic product for those countries with large economies, with combination of business cycles in the short run and long run under this epidemic.

5 Reference

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief*, *29*, 105340. https://doi.org/10.1016/j.dib.2020.105340

Chen, L. P., Zhang, Q., Yi, G. Y., & He, W. (2021). Model-based forecasting for Canadian COVID-19 data. *PLOS ONE*, *16*(1), e0244536. https://doi.org/10.1371/journal.pone.0244536

El Allaoui, A., Melliani, S., & Chadli, L. S. (2020). A Mathematical Modeling and Epidemic Prediction of COVID-19. *SSRN Electronic Journal*, 5. https://doi.org/10.2139/ssrn.3580162

Kufel, T. (2020). ARIMA-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries. *Equilibrium*, *15*(2), 181–204. https://doi.org/10.24136/eq.2020.009

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). OTexts.

Interpret the key results for ARIMA - Minitab. (n.d.). (C) Minitab, LLC. All Rights Reserved. 2019. Retrieved March 12, 2021, from https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/time-series/how-to/arima/interpret-the-results/key-results/

Palachy, S. (2019, September 22). *Stationarity in time series analysis - Towards Data Science*. Medium. https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322#:%7E:text=In%20t%20he%20most%20intuitive,not%20itself%20change%20over%20time.

Palachy, S. (2019b, September 22). *Stationarity in time series analysis - Towards Data Science*. Medium. https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322#:%7E:text=In%20t%20he%20most%20intuitive,not%20itself%20change%20over%20time.

Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics)* (4th ed. 2017 ed.). Springer. https://doi.org/10.1007/978-3-319-52452-8_2

6 Appendix (R codes)

```
> library(forecast)
> library(fUnitRoots)
> library(readr)
> OnQcBcDaily <- read_csv("~/Desktop/OnQcBcDaily.csv")

── Column specification ──────────────────────────────────────────────────────────────────────────────────────────────────────────────
cols(
  Date = col_date(format = ""),
  `Confirmed ON` = col_double(),
  `Confirmed QC` = col_double(),
  `Confirmed BC` = col_double()
)

> View(OnQcBcDaily)

# Time series (time series transformation using the ts() function)
# Time series object is a type of object designed for time series analysis.
# which includes two dimensions, a numeric value describing the metric and a dimension of time.
# A time series object is similar to a general numeric vector, except that a time description is added.
# In the R language you can use ts(data vector, frequency= indicates the time interval that separates the time, start=c(year, month represented by the first data))

> dailycases <- ts(OnQcBcDaily)
> plot.ts(dailycases, xlab="Date", ylab="The Number of Daily Confirmed Cases")
```

 # Ontario codes

```
> OnDaily <- read_csv ("~/Desktop/OnQcBcDaily.csv",
+     col_types = cols_only (`Confirmed ON` = col_guess()))
> View (OnDaily)

> OnDailyCases<-ts(OnDaily)
> plot.ts (OnDailyCases, xlab="Date", ylab="The Number of Daily Confirmed Cases, ON")

# (Partial) Auto correlation function plots of daily cases for Ontario
> acf (OnDailyCases)
> pacf (OnDailyCases)

#Unit Root Test for Ontario
> unitrootTest (OnDailyCases)

Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 1
  STATISTIC:
    DF: -0.2192
  P VALUE:
    t: 0.6069
    n: 0.6305

> difOnDailyCases<-diff(OnDailyCases)
> acf(difOnDailyCases)
> pacf(difOnDailyCases)
```

```
> unitrootTest(difOnDailyCases)

Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 1
  STATISTIC:
    DF: -17.9921
  P VALUE:
    t: < 2.2e-16
    n: 0.002844

# White noise test
> Box.test (difOnDailyCases, type="Ljung-Box")

        Box-Ljung test

data:  difOnDailyCases
X-squared = 45.031, df = 1, p-value = 1.94e-11
>par(opar)

#The auto.arima() function inside R is to select the model by picking the minimum values of
AIC and BIC.
> pdq_On = auto.arima(OnDailyCases)
> pdq_On
Series: OnDailyCases
ARIMA(3,1,3)

Coefficients:
```

```
        ar1     ar2      ar3      ma1      ma2     ma3
      0.3191  0.6231  -0.5865  -0.7553  -0.6031  0.8098
s.e.  0.0696  0.0546   0.0540   0.0605   0.0808  0.0512

sigma^2 estimated as 29660:  log likelihood=-2637.94
AIC=5289.89   AICc=5290.17   BIC=5317.86

# Methods for estimation of model parameters, "ML" maximum likelihood estimation, "CSS"
conditional least squares estimation, "CSS-ML"
> arima_On <- arima (OnDailyCases, order = c(3,1,3))
> arima_On

> forecast_On<-forecast (arima_On, h=40, level=c(99.5))
> forecast_On
    Point Forecast   Lo 99.5  Hi 99.5
404       2545.940 2066.1336 3025.747
405       2549.527 1998.6951 3100.358
406       2543.509 1952.7824 3134.236
407       2550.310 1935.0838 3165.537
408       2546.627 1884.1309 3209.124
409       2553.220 1835.1511 3271.288
410       2549.040 1746.9732 3351.106
411       2553.973 1675.1862 3432.761
412       2549.077 1587.0583 3511.095
413       2553.040 1527.3643 3578.716
414       2548.360 1460.2045 3636.515
415       2552.208 1416.6510 3687.765
416       2548.195 1363.3893 3733.001
417       2552.057 1326.4388 3777.676
418       2548.532 1277.4448 3819.620
419       2552.168 1241.0979 3863.238
```

420 2548.866 1193.1272 3904.605
421 2552.145 1156.9731 3947.318
422 2549.002 1111.0026 3987.001
423 2551.979 1076.4711 4027.487
424 2549.047 1033.5162 4064.578
425 2551.810 1001.0399 4102.580
426 2549.118 960.8209 4137.416
427 2551.701 929.8798 4173.522
428 2549.227 891.6615 4206.794
429 2551.626 861.7280 4241.524
430 2549.335 825.0707 4273.600
431 2551.550 795.9685 4307.131
432 2549.422 760.7524 4338.092
433 2551.467 732.4830 4370.450
434 2549.495 698.6538 4400.335
435 2551.387 671.1917 4431.582
436 2549.563 638.6371 4460.489
437 2551.317 611.8989 4490.735
438 2549.630 580.4792 4518.781
439 2551.254 554.3786 4548.130
440 2549.693 523.9786 4575.407
441 2551.196 498.4573 4603.935
442 2549.750 468.9942 4630.506
443 2551.141 444.0146 4658.268

> plot(arima_On)

> plot(forecast_On)

> accuracy(model_On)

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	9.020076	170.7176	101.1508	-Inf	Inf	0.8738263	-0.02649525

```
> par(mar = c(5,5,3,1))
> qqnorm(model_On$residuals)
> qqline(model_On$residuals)
> Box.test(model_On$residuals, type = "Ljung-Box")

        Box-Ljung test

data:  model_On$residuals
X-squared = 0.28502, df = 1, p-value = 0.5934

# From the drawn QQ plots and the results of the LB test, the residuals are consistent with the normality assumption and are not correlated.
# then the model is considered to fit the data sufficiently and can be used for the next step of prediction.

# First-order differencing of the original series, with smoothness and white noise tests
# First-order differencing
# Syntax: (default) diff(x, lag = 1, diff= 1, ...)
```

#Quebec codes

```
> QcDaily <- read_csv ("Desktop/OnQcBcDaily.csv",
+     col_types = cols_only (`Confirmed QC` = col_guess()))
> View(QcDaily)

> QcDailyCases<-ts(QcDaily)
> plot.ts(QcDailyCases, xlab="Date", ylab="The Number of Daily Confirmed Cases, QC")
> acf(QcDailyCases)
> pacf(QcDailyCases)
> unitrootTest(QcDailyCases)
Title:
```

Augmented Dickey-Fuller Test

Test Results:
PARAMETER:
Lag Order: 1
STATISTIC:
DF: -1.3871
P VALUE:
t: 0.1538
n: 0.4099

> difQcDailyCases<-diff(QcDailyCases)
> acf(difQcDailyCases)
> pacf(difQcDailyCases)
> unitrootTest(difQcDailyCases)

Title:
Augmented Dickey-Fuller Test

Test Results:
PARAMETER:
Lag Order: 1
STATISTIC:
DF: -26.409
P VALUE:
t: < 2.2e-16
n: 0.0002776

Description:
Tue Apr 20 04:13:14 2021 by user:

```
> Box.test(difQcDailyCases,type="Ljung-Box")

        Box-Ljung test

data:  difQcDailyCases
X-squared = 140.33, df = 1, p-value < 2.2e-16

> pdq_Qc=auto.arima(QcDailyCases)
> pdq_Qc
Series: QcDailyCases
ARIMA(1,1,3)

Coefficients:
         ar1      ma1     ma2      ma3
      0.8518  -1.8197  1.0384  -0.1404
s.e.  0.0662   0.0848  0.1235   0.0627

sigma^2 estimated as 53825:  log likelihood=-2758.7
AIC=5527.39   AICc=5527.54   BIC=5547.38
>
> arima_Qc <- arima(QcDailyCases, order = c(1,1,3))
> arima_Qc

Call:
arima(x = QcDailyCases, order = c(1, 1, 3))

Coefficients:
         ar1      ma1     ma2      ma3
      0.8518  -1.8197  1.0384  -0.1404
s.e.  0.0662   0.0848  0.1235   0.0627
```

sigma^2 estimated as 53290: log likelihood = -2758.7, aic = 5527.39

> forecast_Qc<-forecast (arima_Qc, h=40, level=c(99.5))

> forecast_Qc

	Point Forecast	Lo 99.5	Hi 99.5
404	1096.173	448.182555	1744.164
405	1184.020	535.695976	1832.345
406	1216.329	548.690423	1883.967
407	1243.848	550.653635	1937.043
408	1267.289	543.116208	1991.462
409	1287.255	527.684380	2046.826
410	1304.262	505.879170	2102.646
411	1318.749	479.050664	2158.447
412	1331.088	448.347444	2213.828
413	1341.598	414.719507	2268.476
414	1350.550	378.937898	2322.163
415	1358.176	341.620408	2374.731
416	1364.671	303.257534	2426.085
417	1370.204	264.235982	2476.171
418	1374.916	224.858792	2524.973
419	1378.930	185.362020	2572.498
420	1382.349	145.928315	2618.770
421	1385.262	106.697851	2663.825
422	1387.742	67.777072	2707.707
423	1389.855	29.245678	2750.465
424	1391.655	-8.837819	2792.148
425	1393.188	-46.431659	2832.808
426	1394.494	-83.507306	2872.495
427	1395.606	-120.046623	2911.259
428	1396.553	-156.039631	2949.146
429	1397.360	-191.482709	2986.203

```
430      1398.048 -226.377160 3022.473
431      1398.633 -260.728063 3057.995
432      1399.132 -294.543360 3092.807
433      1399.557 -327.833126 3126.947
434      1399.919 -360.608985 3160.446
435      1400.227 -392.883647 3193.337
436      1400.489 -424.670548 3225.649
437      1400.713 -455.983552 3257.409
438      1400.903 -486.836733 3288.643
439      1401.066 -517.244192 3319.375
440      1401.204 -547.219925 3349.627
441      1401.321 -576.777717 3379.421
442      1401.422 -605.931064 3408.775
443      1401.507 -634.693117 3437.707
> plot(forecast_Qc)
> tsdiag(arima_Qc)
```

#British Columbia

```
> BcDaily <- read_csv("~/Desktop/OnQcBcDaily.csv",
+     col_types = cols_only(`Confirmed BC` = col_guess()))
> View(BcDaily)

> BcDailyCases <- ts(BcDaily)
> plot.ts(BcDailyCases, xlab = "Date", ylab = "The Number of Daily Confirmed Cases,BC")
> acf(BcDailyCases)
> pacf(BcDailyCases)
> unitrootTest(BcDailyCases)

Title:
 Augmented Dickey-Fuller Test
```

Test Results:
PARAMETER:
Lag Order: 1
STATISTIC:
DF: -8.2798
P VALUE:
t: 1.957e-14
n: 0.04567

> difBcDailyCases<-diff(BcDailyCases)
> acf(difBcDailyCases)
> pacf(difBcDailyCases)
> unitrootTest(difBcDailyCases)

Title:
Augmented Dickey-Fuller Test

Test Results:
PARAMETER:
Lag Order: 1
STATISTIC:
DF: -28.0733
P VALUE:
t: < 2.2e-16
n: 0.0001761

Description:
Tue Apr 20 02:31:42 2021 by user:

> Box.test(difBcDailyCases, type="Ljung-Box")

```
        Box-Ljung test

data:  difBcDailyCases
X-squared = 79.202, df = 1, p-value < 2.2e-16

> pdq_Bc=auto.arima(BcDailyCases)
> pdq_Bc
Series: BcDailyCases
ARIMA (2,1,3) with drift

Coefficients:
         ar1      ar2      ma1     ma2      ma3   drift
      1.1393  -0.8201  -2.3677  2.2084  -0.7887  2.0767
s.e.  0.0374   0.0375   0.0313  0.0613   0.0390  1.1952

sigma^2 estimated as 96181:  log likelihood=-2876.19
AIC=5766.37   AICc=5766.66   BIC=5794.35

> arima_Bc <- arima(BcDailyCases, order = c(2,1,3))
> arima_Bc

Call:
arima(x = BcDailyCases, order = c(2, 1, 3))

Coefficients:
         ar1      ar2      ma1     ma2      ma3
      1.1390  -0.8197  -2.3607  2.1978  -0.7822
s.e.  0.0376   0.0377   0.0314  0.0616   0.0392

sigma^2 estimated as 95441:  log likelihood = -2877.63,  aic = 5767.26
```

```
> forecast_Bc<-forecast (arima_Bc, h=40, level=c(99.5))
> forecast_Bc
    Point Forecast    Lo 99.5  Hi 99.5
404       264.0254 -603.16784 1131.219
405       642.5114 -245.74739 1530.770
406      1074.0875  162.71186 1985.463
407      1255.3936  343.61665 2167.170
408      1108.1281  177.96658 2038.290
409       791.7762 -179.44188 1762.994
410       552.1733 -442.56194 1546.909
411       538.5887 -458.17185 1535.349
412       719.5222 -278.82630 1717.871
413       936.7376  -63.96062 1937.436
414      1035.8277   35.00615 2036.649
415       970.6347  -37.56009 1978.829
416       815.1554 -207.83845 1838.149
417       691.5072 -341.75957 1724.774
418       678.1230 -357.58382 1713.830
419       764.2351 -271.51464 1799.985
420       873.2862 -162.46350 1909.036
421       926.9060 -109.61770 1963.430
422       898.5870 -142.11359 1939.288
423       822.3793 -225.32006 1870.079
424       758.7936 -294.55828 1812.146
425       748.8395 -307.01526 1804.694
426       789.6240 -267.02864 1846.277
427       844.2364 -212.99039 1901.463
428       873.0071 -185.58045 1931.595
429       861.0098 -200.61081 1922.630
430       823.7613 -242.10013 1889.623
431       791.1704 -278.49128 1860.832
```

```
 432      784.5831 -287.52446 1856.691
 433      803.7955 -269.78529 1877.376
 434      831.0779 -243.77819 1905.934
 435      846.4032 -230.19164 1922.998
 436      841.4947 -237.66715 1920.657
 437      823.3416 -258.94826 1905.631
 438      806.6891 -278.57612 1891.954
 439      802.6027 -285.02854 1890.234
 440      811.5986 -277.88296 1901.080
 441      825.1945 -266.00179 1916.391
 442      833.3058 -259.83980 1926.451
 443      831.3998 -264.10336 1926.903
> plot(forecast_Bc)

> tsdiag(arima_Bc)
```