Sneha Mikkilineni Durga, Binal Patel, Yuening Li

COVID-19 TIME SERIES ANALYSIS

Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. The main focus of this project is to forecast cumulative confirmed cases for Canada, as a whole and for the following provinces: Alberta, British Columbia, Ontario and Quebec. Since these provinces had an increasing number of cases over the past few months. Different statistical models were implemented on the data to understand the performance of each model.

Dataset

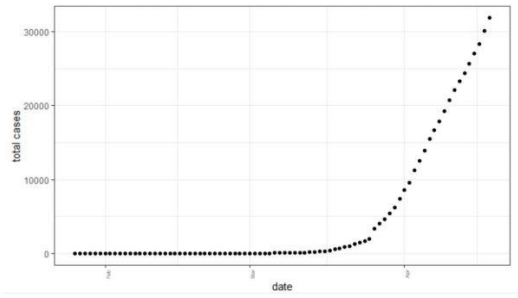
The data is collected from the publicly available information on confirmed and presumptive positive cases during the ongoing COVID-19 outbreak in Canada. The nation's data set is obtained from: (https://github.com/ishaberry/Covid19Canada). It contains the number of new cases, total cases, number of deaths and total deaths of COVID 19 per day in Canada from 1/26/2020, when the first case was detected, to 4/27/2020. Another dataset used for provinces in Canada is 'cases_timeseries_prov.csv'. This dataset has 1261 rows and 4 columns spanning between 25 January 2020 and 23 April 2020. It includes information about the 14 provinces, cases newly recorded per day and cumulative cases.

Analysis of Canada:

When fitting models to data, data splitting was performed. In this case, two subsets of data were created. The first subset includes total cases of COVID-19 in Canada from 2020-01-26 to 2020-04-18 used to fit some models. The second subset includes total cases of COVID-19 in Canada from 2020-4-19 to 2020-4-27 used to test the models.

Inspect Data for model fitting:

Total cases of COVID 19 in Canada from 2020-01-26 to 2020-04-18



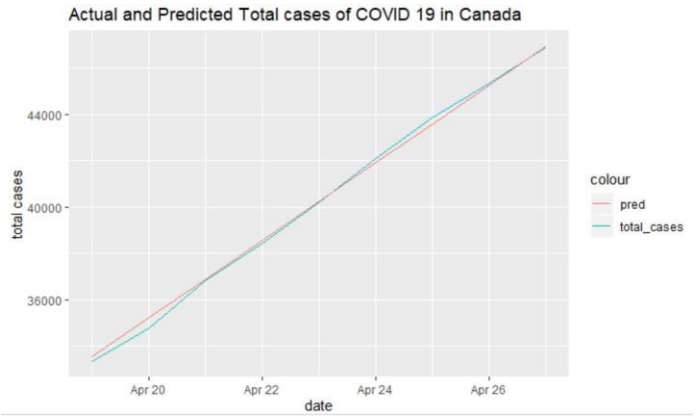
Nonstationary Arima Model (Auto regressive time series)

Apply auto.arima() function to determine the best combination of parameters for the time series model. The model obtained is $Zn=-0.4607\epsilon n-1+\epsilon nZn=-0.4607\epsilon n-1+\epsilon n$ ARIMA(0,2,1) with standard error 0.0902, which means order 1 Moving Average and degree of differencing 2.

Using the fitted time series model, total cases of COVID-19 were predicted for the following 9 days and compared the number of predicted cases with the actual total cases from 2020-4-19 to 2020-4-27 by subtracting the actual values from the predictions and dividing by the estimated standard errors.

```
## [1]  1.11531199  1.31405534  0.12811850  0.21119574  0.05815873 -0.17418153
## [7] -0.22154960 -0.05817533  0.01852049
```

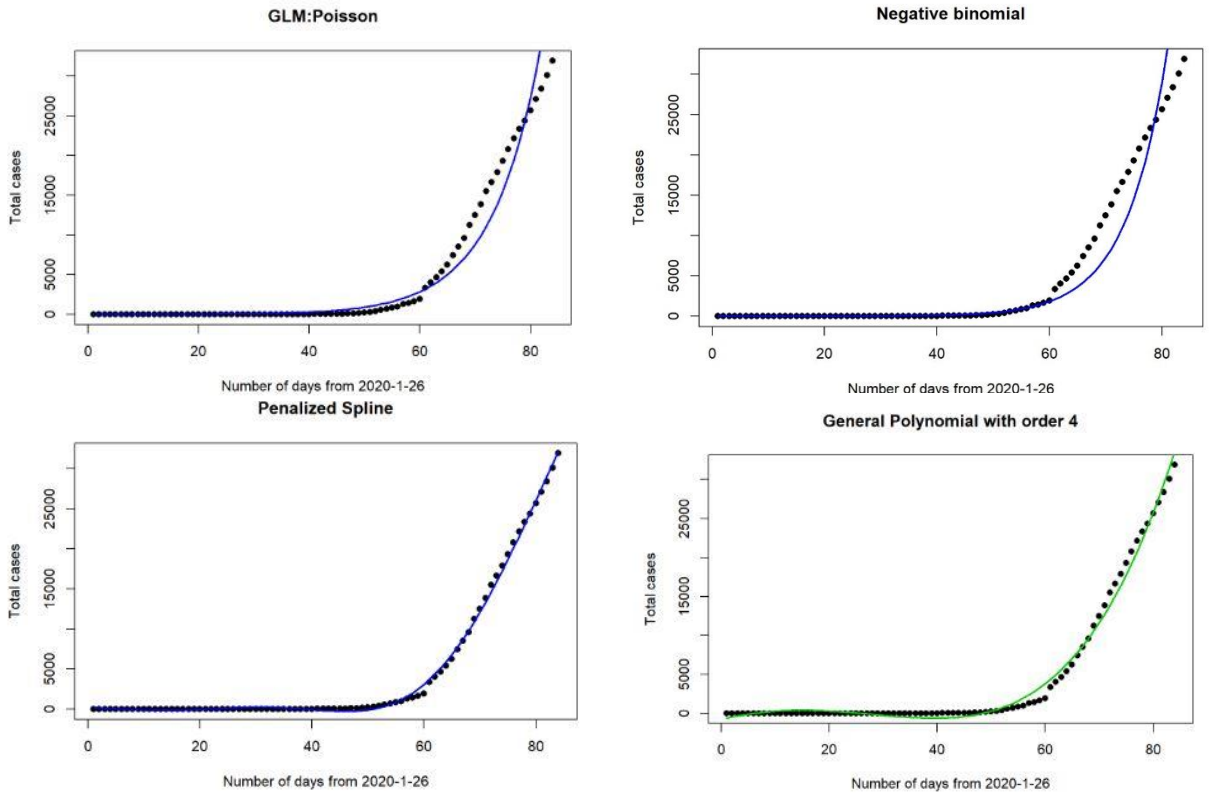
All result obtained are between -2 and 2, suggesting the model is quite accurate. The plot on the right is the actual (in blue) and predicted total cases (in red) of COVID-19 in Canada from 2020-4-19 to 2020-4-27. Using the ARIMA models the predicted number of cases are close to the actual number of cases.



Other Models

Other models were also used on the data set, including "General Polynomial Model", "Penalized Spline Model", "Generalized Linear Model (Quasipoisson regression)", and "Negative Binomial Model".

Figure 1. shows different models that were fitted to the data set. Black dots are the actual value of total cases of COVID-19, color lines are predicted values from the models. For count responses, like total number of COVID-19 cases, Poisson or Quasipoisson GLM were reasonable guesses for modelling. Both Quisilkelihood and Negative Binomial model, are models for counts which allows for overdispersion. The Polynomial and Penalized Spline, on the other hand, though seem to provide better line fitting, it does not account for the dependency of times.



Different models fitted to the data set

Using the predict() function with the total cases from the second subset as newdata object, the Mean Square Error (MSE) for each model were calculated to decide which of the above models is best for the dataset. We obtained: "ARIMA MSE": 44615.71; "polynomial with order 4 MSE": 1644079206; "Penalized Splines Model MSE":1632304310; "GLM quasipoisson MSE": 1635227843; and "GLM Negative binomial Model MSE":1635227843.

Analysis of provinces:

The first step to any analysis is Data Wrangling where the missing values and scale of the data was handled. Further to this step, data exploration was performed to understand the statistics and the observed provinces with the most number of cases during this period. To perform time series analysis, the dataset was converted into time series objects before inputting it into the statistical models. The 'zoo' package with 'xts' function was implemented to perform this operation. 'xts'. They are nothing but a matrix of observations combined with an index of corresponding dates and times. Figure 2. shows the time series plot of all the provinces.

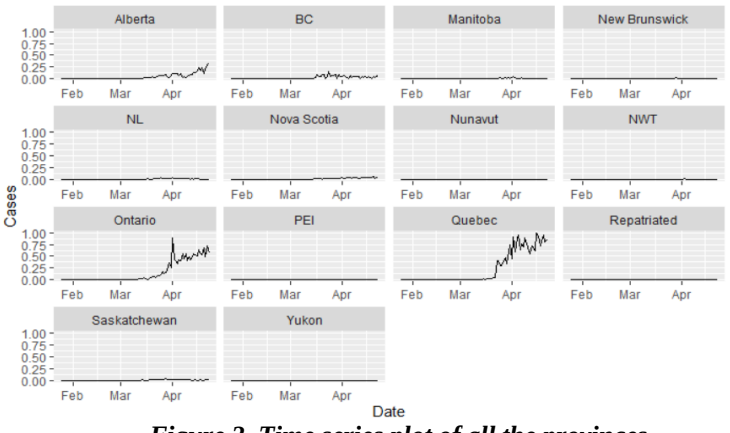


Figure 2. Time series plot of all the provinces

It is important to check if the obtained time series is stationary or non-stationary. For a time series to be stationary, it should hold the following conditions true:

1. The mean value of the time series should be constant.
2. Variance is constant and does not increase over time.
3. Seasonality is minimal.

Visual and statistical tests were used to check for stationarity.

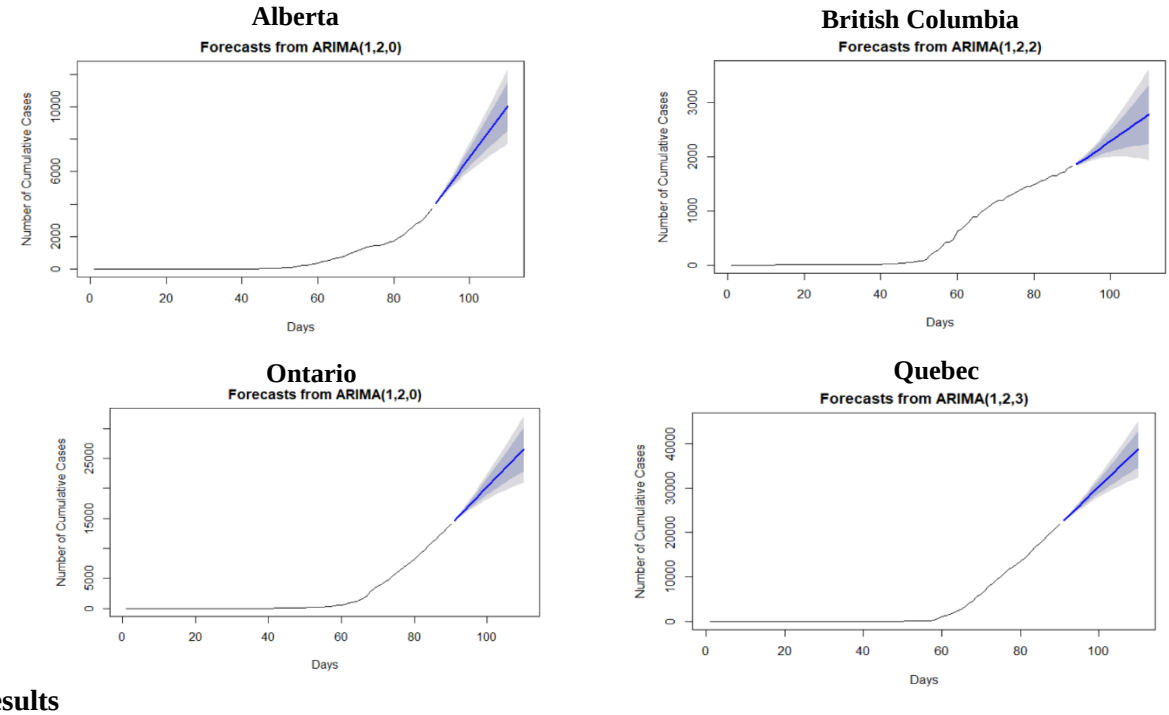
1. Autocorrelation plot:

This plot helps us to find correlation of time series with lags of itself, visually. The autocorrelation shows gradual decay which evidently shows the non-stationarity of the time series.

2. Augmented Dickey-Fuller(ADF) test:

This test is performed to test for stationarity statistically. The null hypothesis states that the time series is non-stationary and the alternative hypothesis states that it is stationary. All the non-stationary time series were transformed into a stationary time series by performing differencing twice until it turns stationary. The reason to perform differencing was to eliminate the trend in the time series data.

auto.arima() function was used to determine the best combination of parameters for the time series model. Using ARIMA model the following forecast plots were obtained for the next 20 days.



Results

Based on our observation, Arima model performs much better than other models above (in Figure 1). The total cases of COVID-19 virus infections for Canada are still growing exponentially. All the four provinces show an increase in the number of cumulative cases for COVID-19 for the next 20 days. The forecasts are shown on the blue line, 80% of the prediction interval is the dark gray area and 95% of the prediction interval is the light gray shaded region that is, chances these predictions will lie in the respectful shaded areas. Since, the p, d and q parameters vary for each of the provinces, there is variation in the prediction. Choosing the appropriate parameters for these models play a crucial role in their predictions. In conclusion, the overall number of cases in Canada are increasing every day.