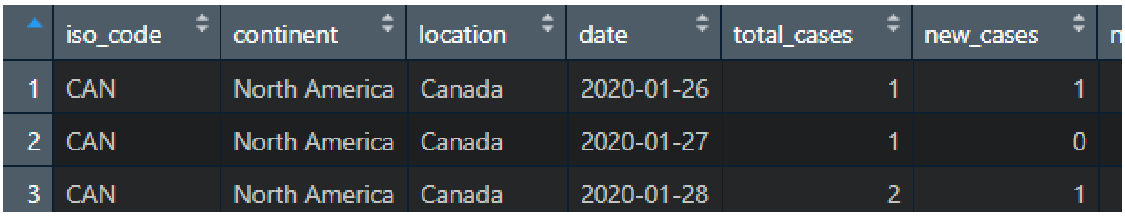
Refer to COVID-19 page on “Our World in Data” website https://ourworldindata.org/coronavirus

Time series models created to forecast the number of daily new cases in Canada from September 14 to September 28, 2021.

Data Exploration and filtering

As the original data set contains Covid 19 information for multiple geographic jurisdictions, it has to be prepared and filtered by Canada relevant information only. By use of the dplyr package, Canada related information were filtered out for later processing (Figure 1).



	iso_code	continent	location	date	total_cases	new_cases
1	CAN	North America	Canada	2020-01-26	1	1
2	CAN	North America	Canada	2020-01-27	1	0
3	CAN	North America	Canada	2020-01-28	2	1

Figure 1. The original data set from the source (https://github.com/owid/covid-19-data/tree/master/public/data) was treated by filtering the 'iso_code' column to retain CAN (Canada) related information only.

The auto.arima was then used to make forecasting using the original time series (i.e., without preprocessing), which suggested a ARIMA(0,1,5) model. Below shows the standard diagnostic charts on the ARIMA(0,1,5) model (Figure 2) and a summary of the suggested model.

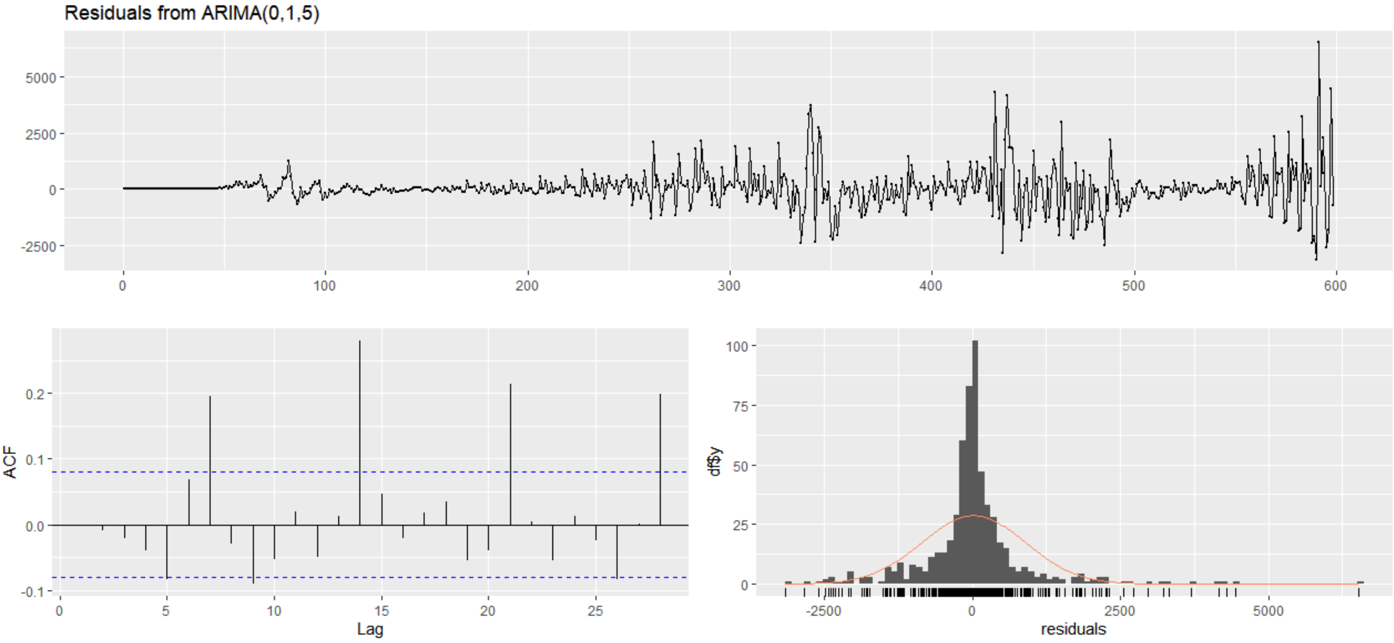


Figure 2. Standard Diagnostic Charts for pre-treated data

As shown in the ACF plot in Figure 2, a seasonal series of alternating of positive and negative lags can be spotted in the data set. The seasonality follows a pattern of strong lags in a set of positive lags followed by sets of negative lags. The pattern seems to repeat every 7 days suggesting a weekly pattern is in place and need to be pre-processed. Otherwise, the data set follows a normal distribution, whilst there appears to be increasing variance over time in the data set. As such, the ARIMA model cannot be applied until the above behaviours are normalized through pre-processing.

```
Series: CANNewcase
ARIMA(0,1,5)

Coefficients:
            ma1        ma2       ma3        ma4       ma5
        -0.9235    -0.0637    0.2963    -0.0014    0.1017
s.e.     0.0405     0.0598    0.0547     0.0738    0.0518

sigma^2 estimated as 771442:    log likelihood=-4892.06
AIC=9796.12      AICc=9796.26      BIC=9822.47
```

Summary of ARIMA(0,1,5)

With reference to the above summary on the suggest model ARIMA(0,1,5), the model has AIC and BIC values of 9796.12 and 9822.47. In the next section, the report will go through the pre-processing steps and improvement procedures to improve the forecasting models and come up with improved ones with lower AIC and BIC values. Below showcase the resultant forecast plot made with the aforementioned model (Figure 3).

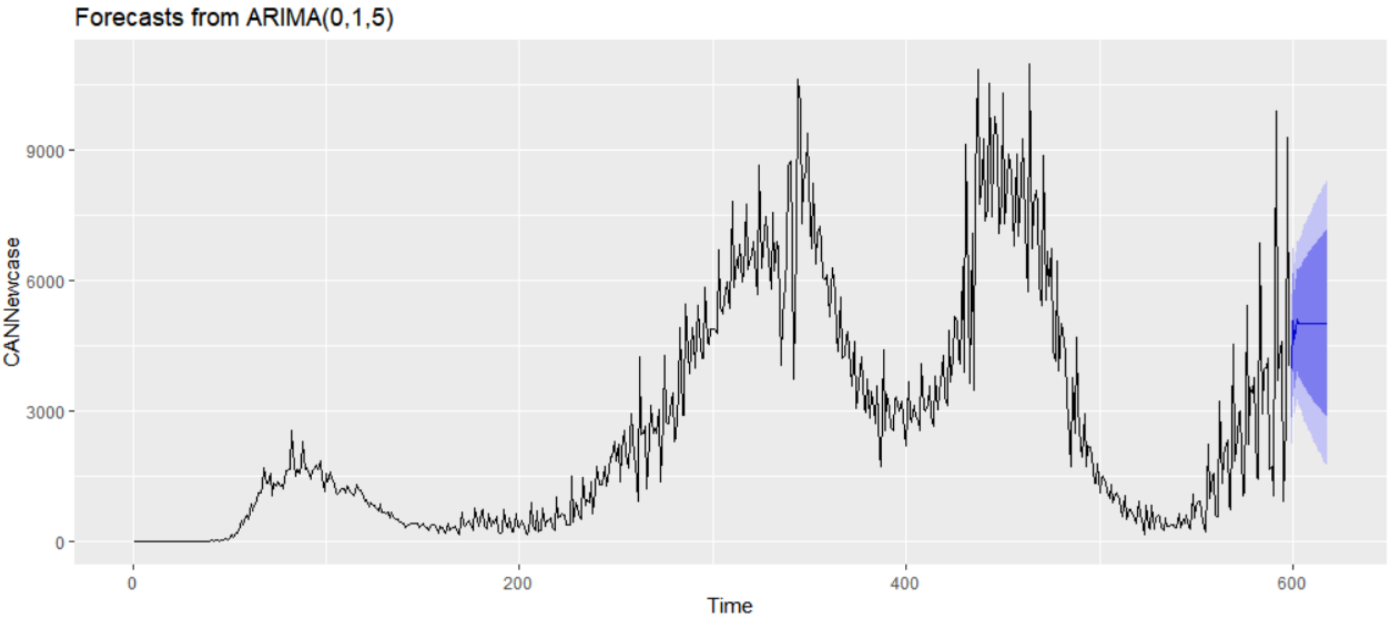


Figure 3. Forecast Plot of Arima (0,1,5)

Pre-Processing

Stabilizing the Variance

It was noted that in the first 100 days, the number of new cases were low, as such the data from the first 100 days were removed from the data set to provide a forecast with stronger emphasis on recency.

As mentioned in the previous section, there was an increased in variance over time in the data set. To stabilize the variance, Box-Cox transformation and Log transformation were attempted (Figure 4).

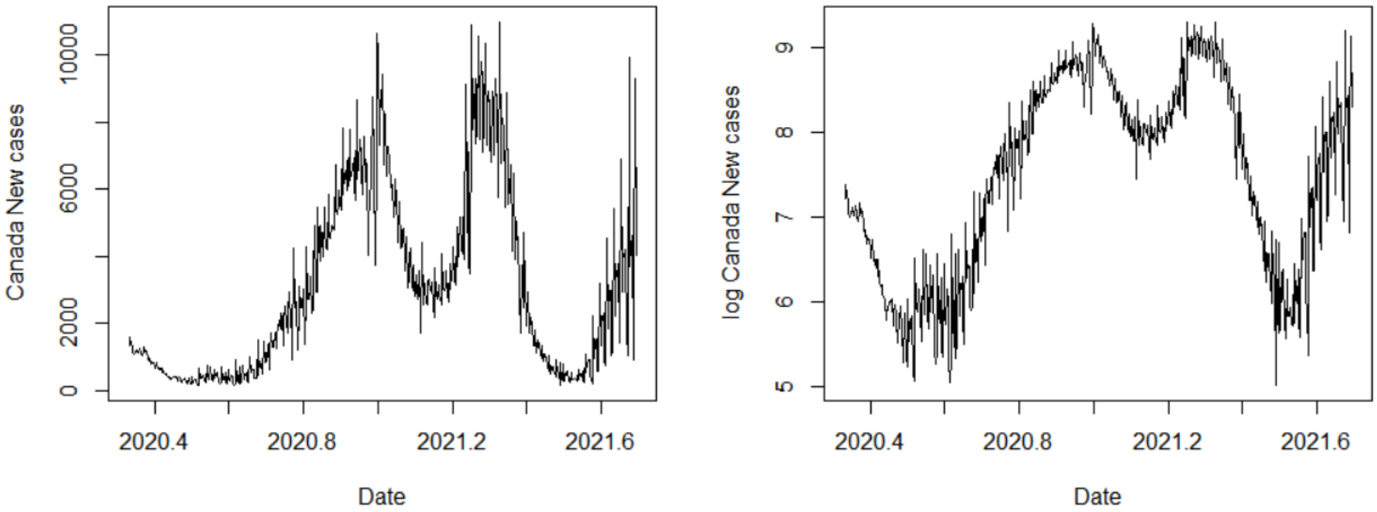


Figure 4. Box-Cox transformed data (Left) versus Log transformed data (Right) plotted against time

As shown in Figure 4, the increasing variance were effectively transformed by both methods, which resembles similar trends over time. For simplicity, the log transformed data will be used for further analysis.

Removing Seasonality

As discussed previous and illustrated in Figure 2, there was a repeated pattern of strong positive lag lags followed by sets of negative lags every 7 days, suggesting a weekly repetitive pattern in the data set. Difference transformation was applied using the diff() function in R with 7 as parameter. Figure 5. Illustrated the transformed data set after removing variance and seasonality and re-examined ACF result verifies that the seasonality pattern has been removed.

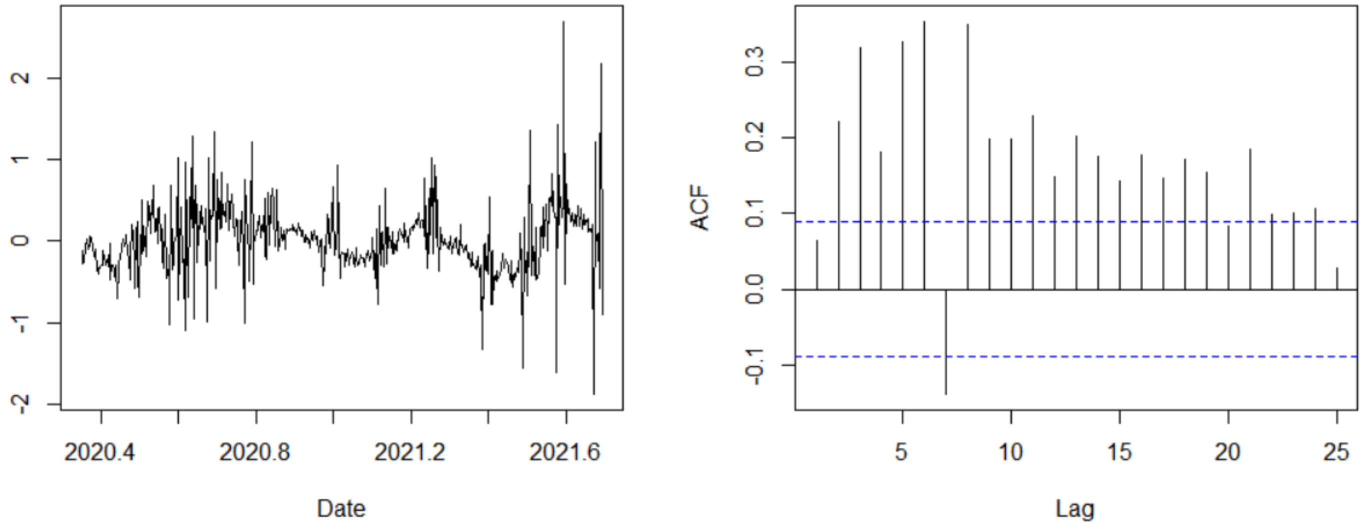


Figure 5. Data after log transformation and removing seasonality (left) and resultant ACF plot (right)

Stationarity check

Augmented Dickey-Fuller Test was used to check for stationarity. Using the adf.test() function in R, the treated data was checked and confirmed that the series is stationery. Below is an extract of the result from the Augmented Dickey-Fuller Test.

```
        Augmented Dickey-Fuller Test

data:    logCANNewcase.processed.deSeasonality
Dickey-Fuller = -4.2018, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(logCANNewcase.processed.deSeasonality, alternative = "stationary") :
    p-value smaller than printed p-value
>
```

Summary of Augmented Dickey-Fuller Test on the treated data

Automatic ARIMA Modeling

Leveraging on the auto.arima function, the treated data was applied, and it was recommended that an ARIMA(0,0,5) model should be adopted. Below is the summary on the ARIMA(0,0,5) model.

```
Series: logCANNewcase.processed.deSeasonality
ARIMA(0,0,5) with zero mean

Coefficients:
             ma1       ma2       ma3       ma4       ma5
         -0.1256    0.5182    0.1208    0.1739    0.5640
s.e.      0.0335    0.0387    0.0421    0.0348    0.0354

sigma^2 estimated as 0.129:    log likelihood=-192.98
AIC=397.97      AICc=398.14      BIC=423.15
```

Summary of ARIMA(0,0,5) model

The quality of the fit was checked again using the checkresiduals() function in R. As noted there were no significant autocorrelation and increasing/decreasing trend in variance. The data set also follows a normal distribution (Figure 6)

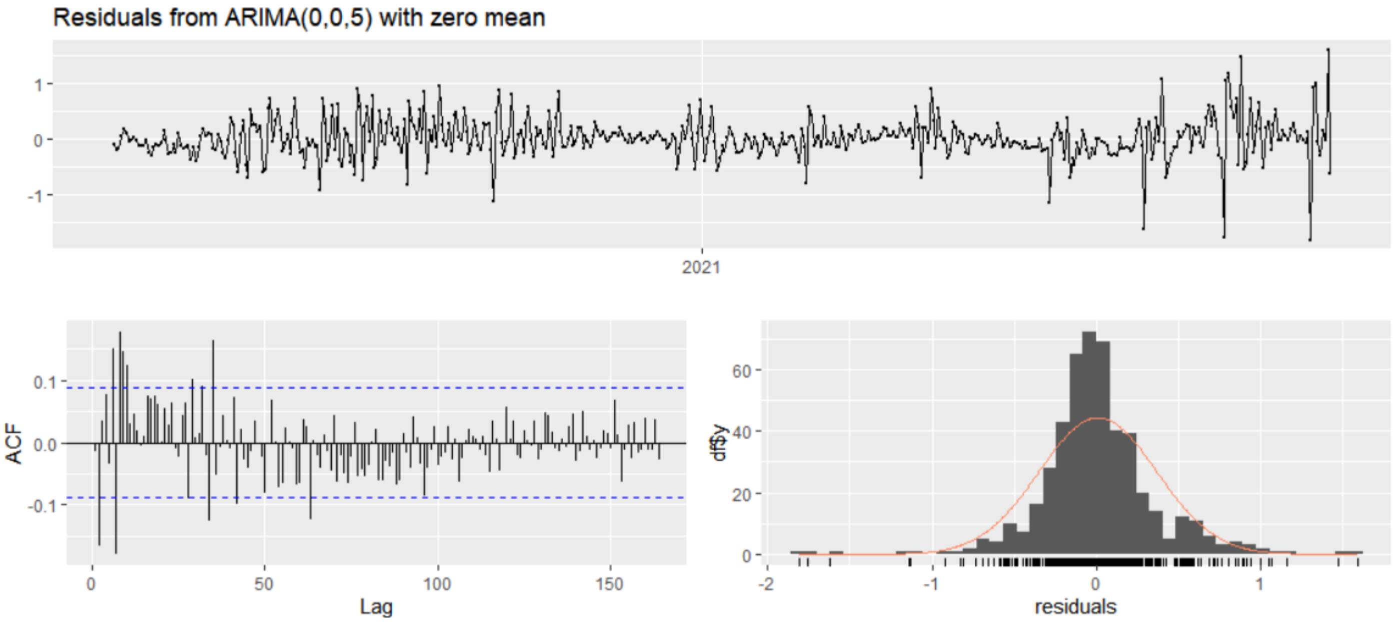


Figure 6. Standard Diagnostic Charts for the treated data

The auto selected model was then used to make forecast for the next 14 days (Sept 14 – 28), and the resulted forecast underwent exponential transformation before yielding the final forecast, which was summarised in Figure 7.

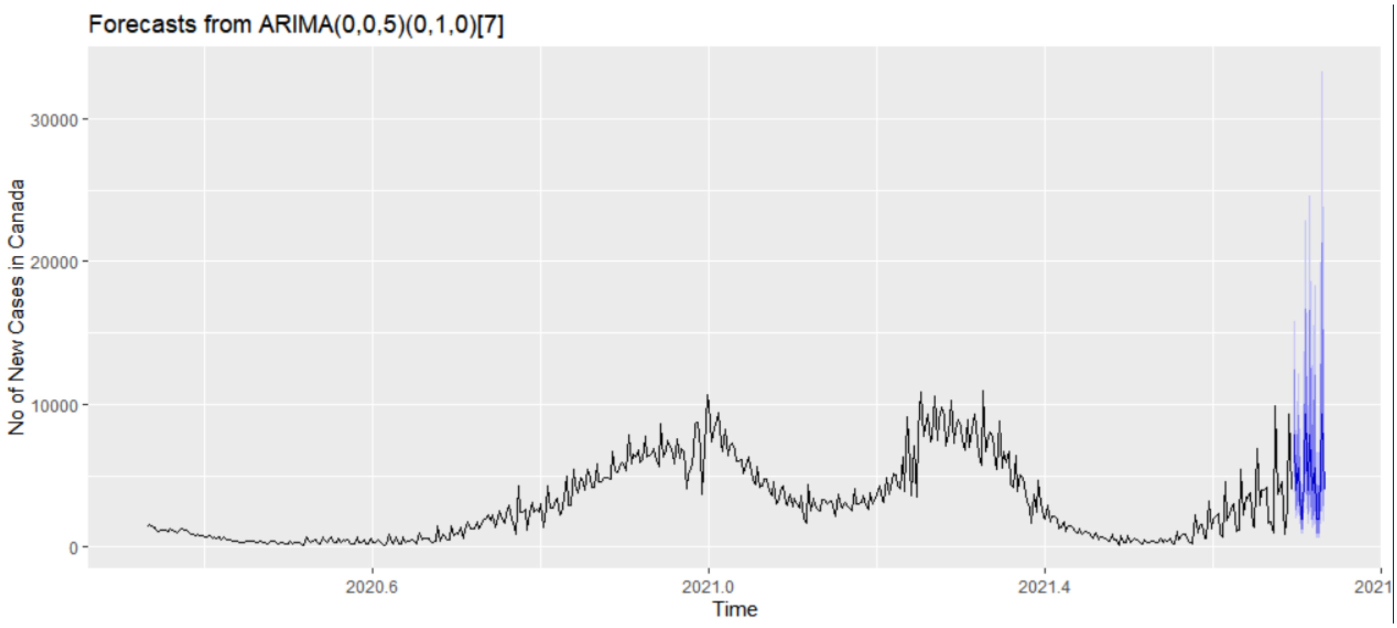


Figure 7. 14 day forecast of New Covid Cases

Improving the Auto-selected models

In attempt to improve the ARIMA model, varying criterion were used to generate other models manually to yield a lower AIC than the existing model – the extracted results are summarised below.

```
> AIC(Arima(log(CANNewcase.processed), order=c(0,0,5), seasonal=list(order=c(0,1,0),period=7)))
[1] 397.9696
> AIC(Arima(log(CANNewcase.processed), order=c(0,0,7), seasonal=list(order=c(0,1,0),period=7)))
[1] 285.8336
```

```
> AIC(Arima(log(CANNewcase.processed), order=c(0,0,9), seasonal=list(order=c(0,1,0),period=7)))
[1] 244.8222
> AIC(Arima(log(CANNewcase.processed), order=c(0,0,11), seasonal=list(order=c(0,1,0),period=7)))
[1] 235.1486
> AIC(Arima(log(CANNewcase.processed), order=c(0,0,13), seasonal=list(order=c(0,1,0),period=7)))
[1] 238.7721
> AIC(Arima(log(CANNewcase.processed), order=c(0,0,15), seasonal=list(order=c(0,1,0),period=7)))
[1] 232.0103
> AIC(Arima(log(CANNewcase.processed), order=c(0,0,17), seasonal=list(order=c(0,1,0),period=7)))
[1] 232.7905
> AIC(Arima(log(CANNewcase.processed), order=c(0,0,19), seasonal=list(order=c(0,1,0),period=7)))
[1] 233.7181
> AIC(Arima(log(CANNewcase.processed), order=c(0,0,21), seasonal=list(order=c(0,1,0),period=7)))
[1] 233.9424
>
```

Summary of criterion change and resultant AIC values

As illustrated, the model score a lower AIC score via using the parameters (0,0,15) [highlighted]. Based on the manually yield criteria, a revised forecast was performed with a ARIMA(0,0,15) model and the resulted was illustrated in Figure 8.

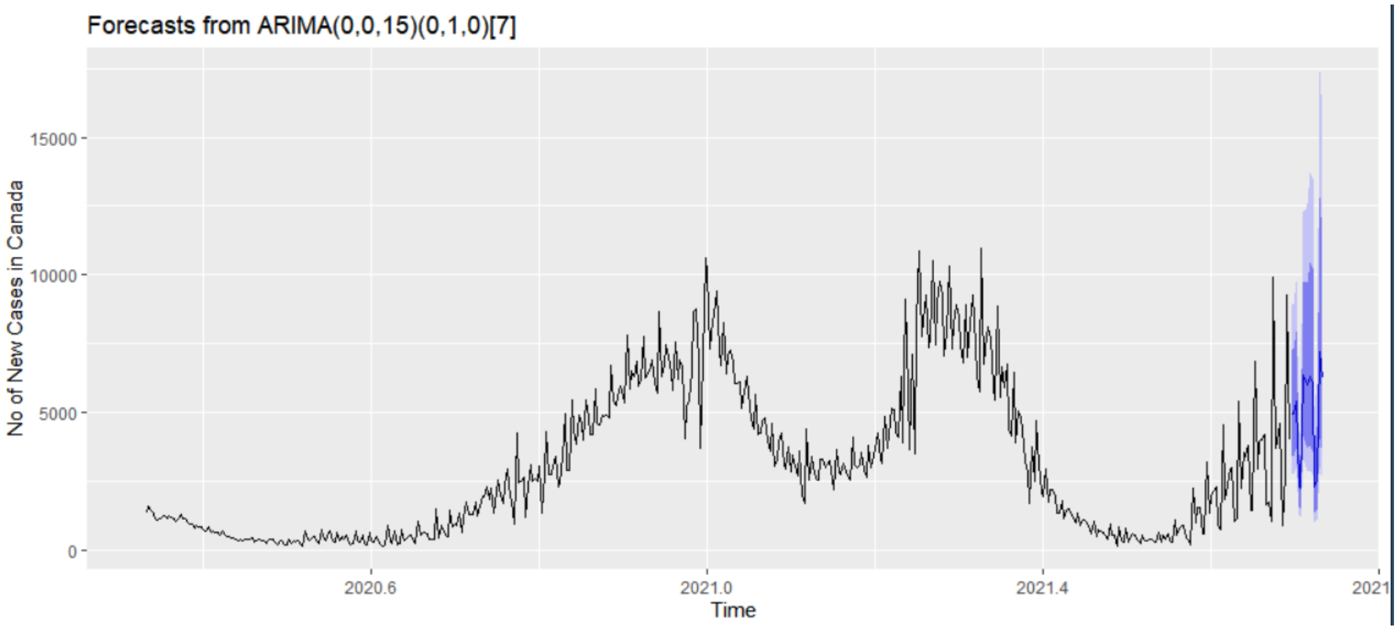
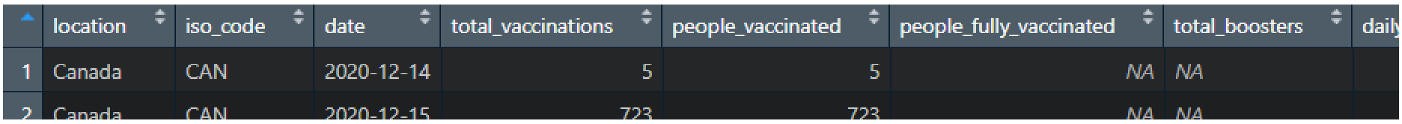


Figure 8. Figure 7. 14 day forecast of New Covid Cases with revised model

Dynamic Regression

The vaccination time series data provided from the source materials was used to serve as a predictor (X), which was incorporated into the covid time series model through dynamic regression. Again, only Canada related information are used for further analysis (Figure 9.). In the following regression, the new daily vaccination count was used to forecast the number of new covid cases. It should be noted that the log transformed data was used to perform the forecast.



	location	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	total_boosters	daily
1	Canada	CAN	2020-12-14	5	5	NA	NA	
2	Canada	CAN	2020-12-15	723	723	NA	NA	

Figure 9. (Extract) Canada related Covid vaccination statistics

Lagged predictors were introduced by increments in time periods of 90 days(i.e. T, T+90 days, T+180 days and T+270 days) as the vaccination effect on a community has not been systematically verified as of today. Therefore, quarterly cut-off times are used to cover larger historical periods.

```
> # Compute Akaike Information Criteria
> AIC(fit1)
[1] 664.6429
> AIC(fit2)
[1] 658.8672
> AIC(fit3)
[1] 651.0322
> AIC(fit4)
[1] 195.6021
```

```
> # Compute Bayesian Information Criteria
> BIC(fit1)
[1] 675.4382
> BIC(fit2)
[1] 673.246
> BIC(fit3)
[1] 668.9309
> BIC(fit4)
[1] 220.9659
```

Summary of AIC results for T+0,T+90,T+180,T+270 lagged models, arranged from fit 1 to 4 respectively

As illustrated, model fit 4 performs best in terms of AIC and BIC, and suggests using an ARIMA(1,0,1) model.

```
Series: log(CANNewcase.Tminus274[5:274])
Regression with ARIMA(1,0,1) errors

Coefficients:
              ar1       ma1    intercept   DailyVaccinations0   DailyVaccinations1   DailyVaccinations2
DailyVaccinations3
          0.9863    -0.707        7.5716                        0                        0e+00
0e+00                               0
s.e.    0.0073      0.038        0.2431                        0                        1e-04
1e-04                               0

sigma^2 estimated as 0.1008:    log likelihood=-89.8
AIC=195.6       AICc=196.15      BIC=224.39
```

Summary of the ARIMA(0,0,0) model

2 forecasts were simulated by 1) assuming the daily vaccination rate to be the average of the last 14 days (and 2) assuming the daily vaccination rate to be the median of the original distribution. The results of which are shown in Figure 10 and 11.

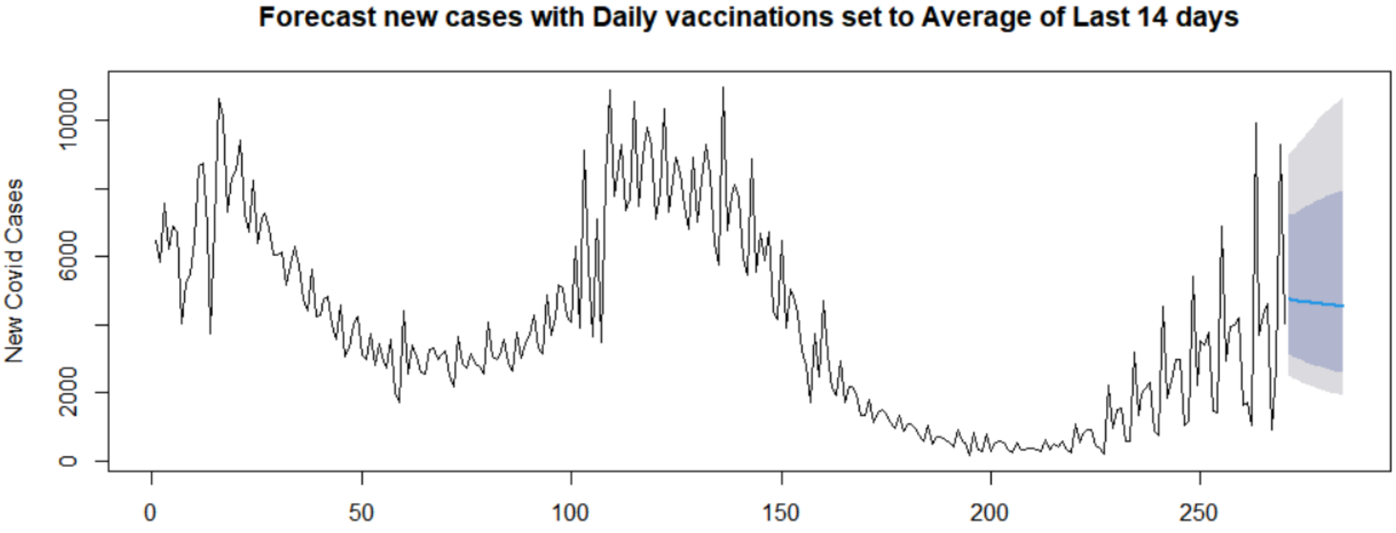


Figure 10. 14 days forecast of new cases with future new vaccination count set to average of last 14 days

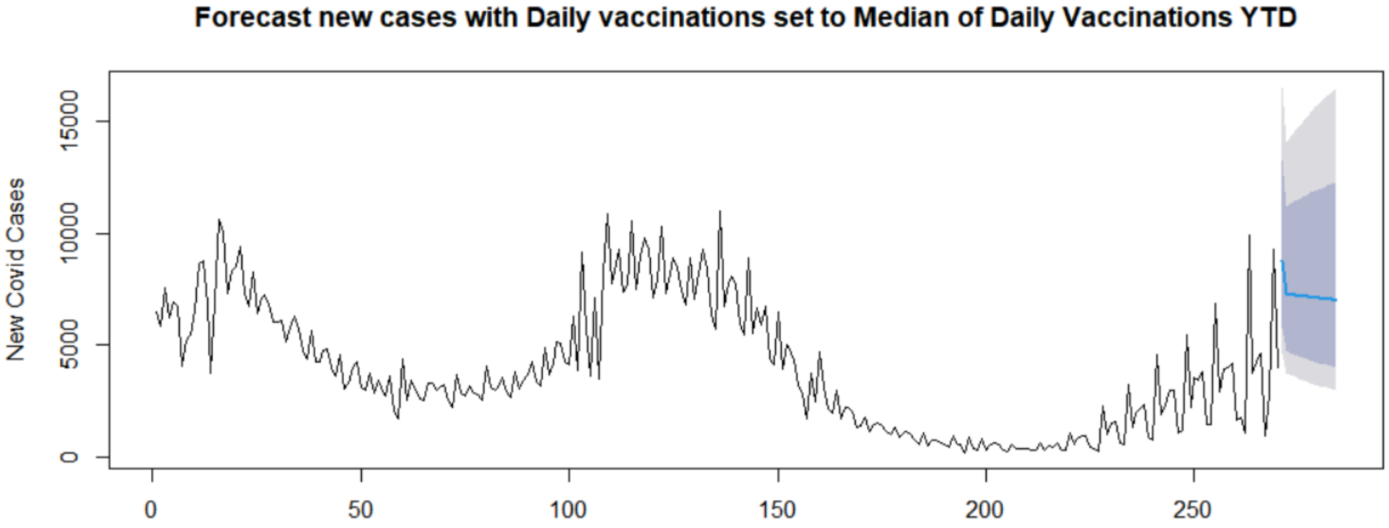


Figure 11. 14 days forecast of new cases with future new vaccination count set to average of last 14 days

As illustrated, the forecasts seem to suggest that there would be a downward trend in new cases in the next 14 days. However, they also seem to suggest that the higher the daily vaccination count, the higher the upper bound of the new cases would be. The forecasts can be further improved if forecasted vaccination figures can be used as the basis for forecasting. However, due to limitation in data, such forecasted data is not readily available.