STA457 Final Project

Predicting number of confirmed cases of COVID-19 in Toronto with ARIMA model

Lingjun Meng

2022/4/12

Contents

Abstract

Given the vaccination rate of second does of Toronto residents aged 18 and above approaching 90% as per 12^{th} April, University of Toronto is planning to shift all the course delivery back to in-person in the following semester. Is this shifting back really safe? Data about confirmed cases of COVID-19 in Toronto provided by Toronto Public Health was collected and cleaned to fit a time series. Time series analysis was conducted, and an ARIMA(1,1,0) was fitted on the log transformed data. The model diagnostics showed that the model performs well, and Spectral Analysis was then done to find the first three predominant frequencies. the predictions made based on the model indicate that the weekly number of confirmed cases of COVID-19 is expected to grow in the following ten weeks.

Keywords: COVID-19, Time Series Analysis, ARIMA Model, Forecasting, Spectral Analysis

Introduction

Background Information

It has been 2 and a half years since the outbreak of the COVID-19. All the students and the teachers at University of Toronto had a difficult time fighting against the pandemic brought by the virus. Now with the vaccination rate of second does of Toronto residents aged 18 and above approaching 90% (as per 12^{th} April)[1], the university is planning to shift all the courses back to in-person.

Online courses indeed have some drawbacks that are not negligible compared to in-person learning, including but not limited to causing social isolation, lack of involvement,

accessibility and academic integrity problems.[2] Thus, the urgent need of transfer the course delivery back to in-person indeed exists. However, we cannot ignore the possible threat that the virus can bring to the society of UofT while moving back to in-person learning.

Purpose of the report

In this report, I will perform a safety assessment of shifting back to in-person learning by conducting a time series analysis of the number of confirmed cases of the COVID-19 in Toronto. More specifically, I will fit a model to explain the trend of the number of infections, and then make predictions of the situation in the near future. Therefore we can have a more clear understanding of how the pandemic will evolve, and hence we can conclude about the level of safety of shifting back to in-person learning.

Data

Originla Data

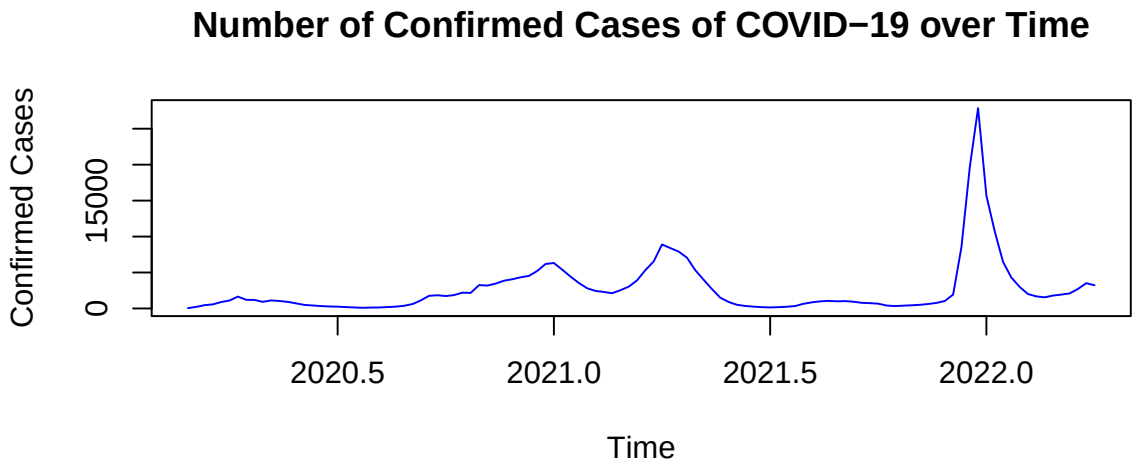
Published and updated by the Toronto Public Health(TPH), the original data[3] can be found at the Toronto open data portal. As a response of the ongoing outbreak brought by the COVID-19, TPH provided data about 18 variables including demographic, geographic, and severity information for all cases reported to and managed by Toronto Public Health, both confirmed and probable.[3] There are total of 309355 rows, each represents a unique observation of the reported case. In addition, the dataset is updated once a week since the pandemic is evolving.

Cleaned Data

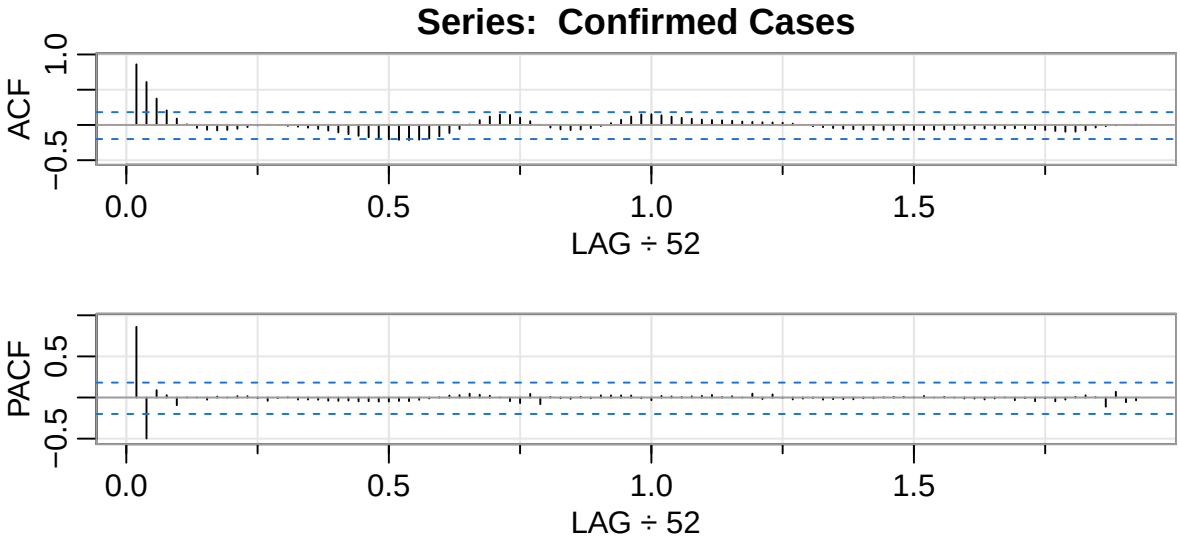
Only cases marked with "Confirmed" were included and the data is summarized to include the number of cases associated with each week specified by the "Episode.time" variable. Lastly, a time series was fitted using the number of confirmed cases of COVID-19 in each week in Toronto, starting from the 9^{th} week in 2020 to the 14^{th} week in 2022, and contains 110 observations.

Statistical Method

Exploring the original time series

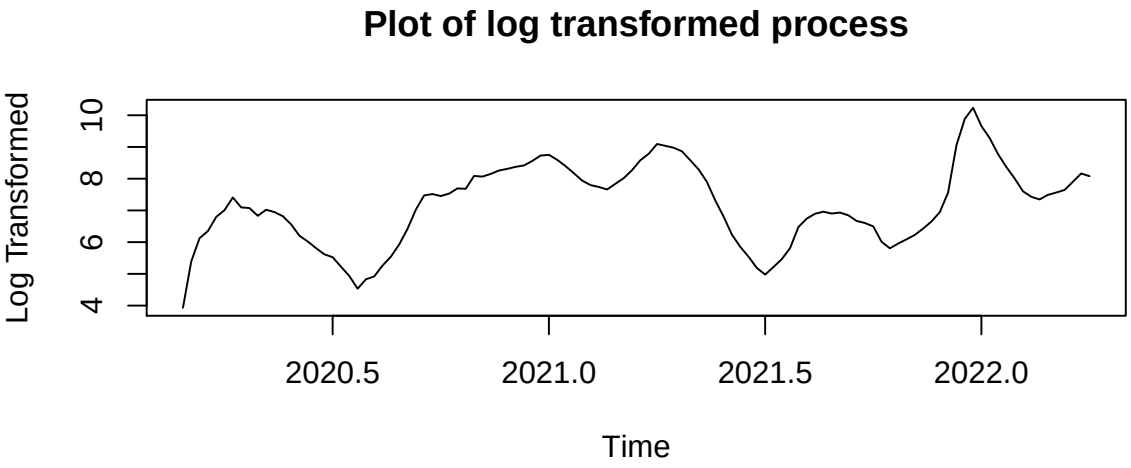


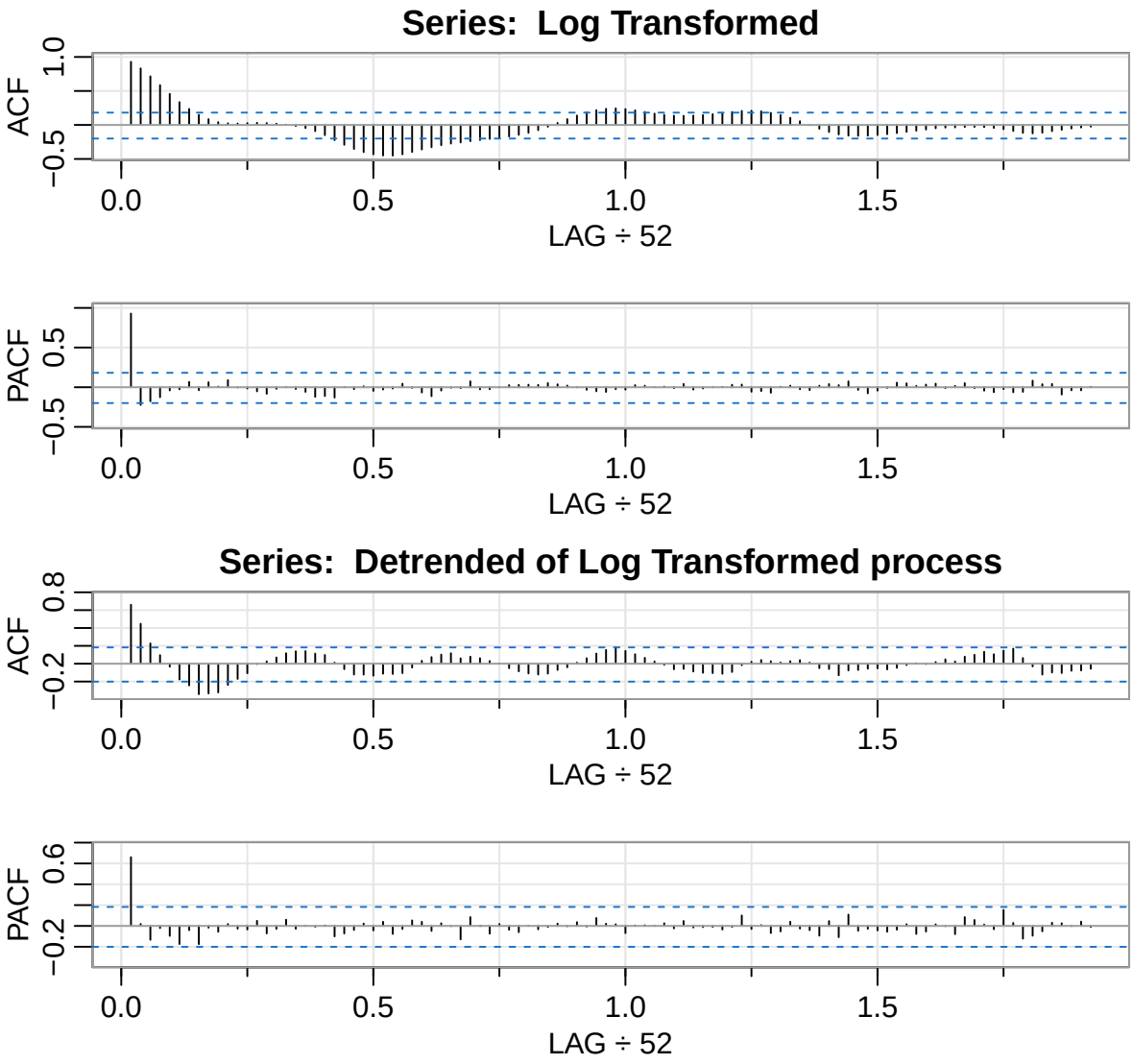
We can see from the plot that the process is not stationary, where a process is considered stationary when it has a constant mean and a variance that is independent with time. Here the variance seems to be increasing as time increases. We will start analyzing the aotocorrelation function($\rho(h)$) of the sample without any transformations. Please see the following graph for the sample ACF and PACF for the process.



The $\rho(h)$ quickly decrease to 0, this indicates no differencing is required(d=0). Also, since the PACF cuts off at lag 2 while ACF tails off, the graph generated indicates we should try fitting an AR(2) model. Therefore, the first proposed model is ARIMA(2,0,0).

Exploring the transformed process





We can see that the log transformed process now has a statble variance. It is illustrated that the sample ACF $\rho(h)$ is not decaying quickly to 0, which is a indicator that differencing is required. Now we difference the process and check the new sample ACF. It can be observed that the new sample PACF cuts off at lag 1 while the new sample ACF tails off. Thus, the graph suggests the AR(1) model after first differencing(with d = 1) Therefore, the second model proposed is a ARIMA(1,1,0) fitted with the log transformed data.

Results

Summary of estimated parameters of proposed models

First Model fitted with Original Process

Table 1: Summary table for the estimated parameters for the first model

	Estimate	SE	t.value	p.value
ar1	1.2854	0.0815	15.7712	0
ar2	-0.4959	0.0814	-6.0912	0

According to the summary table of the parameters, the model can be represented as the following:

$x_t + 1.285_{(0.082)}x_{t-1} - 0.196_{(0.081)}x_{t-2} = w_t$

Second Model fitted with Log Transformed Process

Table 2: Summary table for the estimated parameters for the second model

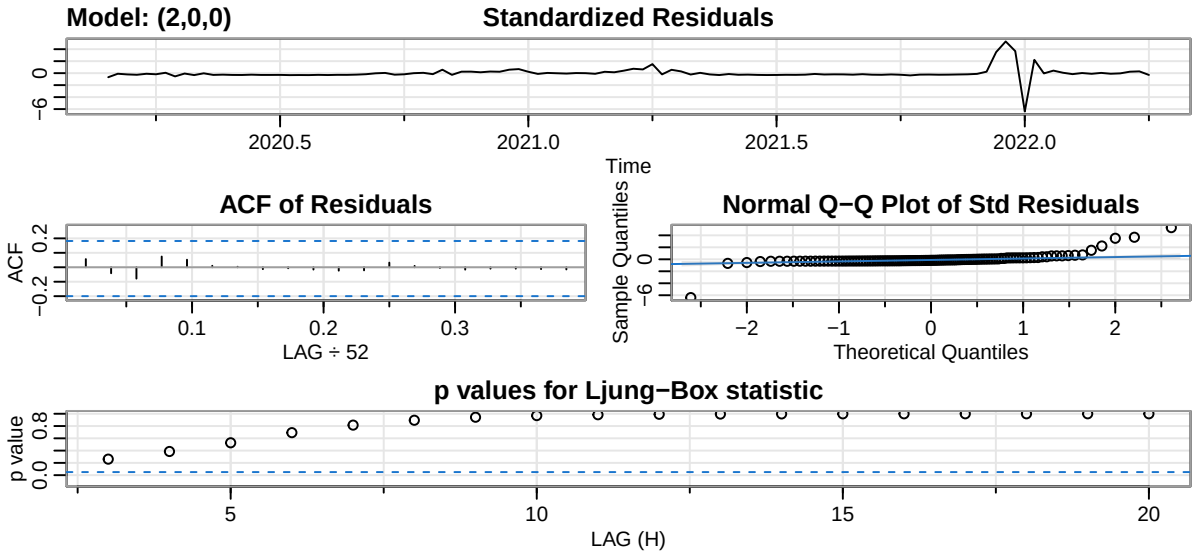
Info	ar1
Estimate	0.7669
SE	0.0706
t.value	10.8552
p.value	0.0000

According to the summary table of the parameters, the model can be represented as the following:

$x_t + 0.767_{(0.071)}x_{t-1} = w_t$

Note that the x_t, x_{t-1} and w_t is different from that in the previous model since this model is fitted with log transformed data. In addition, we can see that the all the parameters estimated for both models have a p-value of 0. Thus, with $\alpha = 0.001$, all the coefficients are significant since $P = 0 < 0.001$.

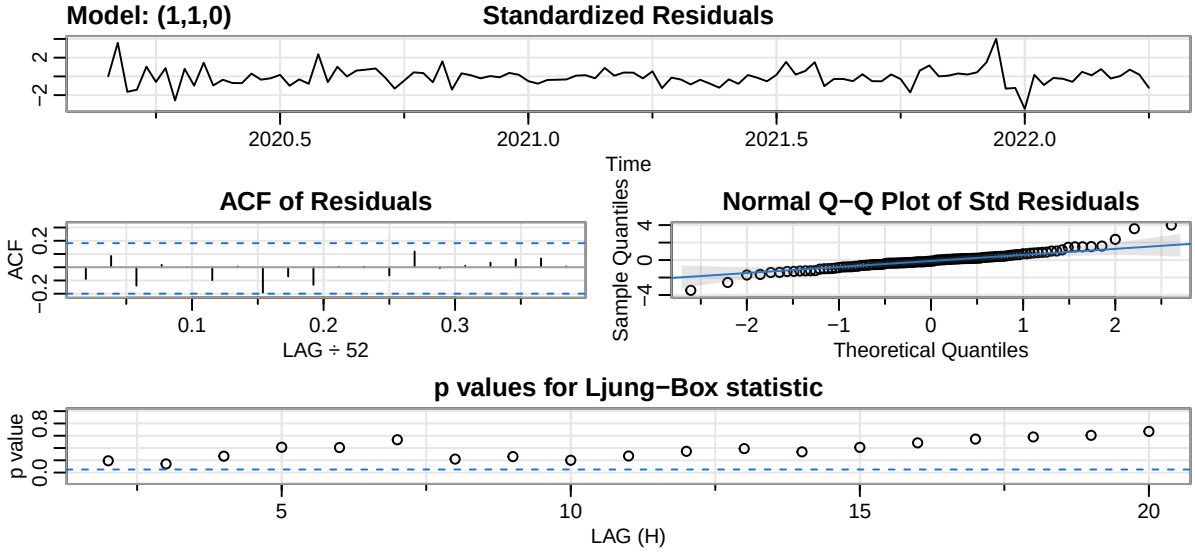
Model diagostics of the proposed model



Inspection of the Standardized Residuals plot showed obvious patterns, most of the residuals stays within 2 standard deviations while there are outliers stay at around 6 standard deviation at about the end of year 2021. The sample ACF of the residuals showed no significant spikes, then we can conclude that the model randomness assumption is not violated.

However, we can see from the Normal Q-Q Plot of the residuals that the normality assumption is violated the dots there are significant departure from the straight line at the top right corner. Lastly, we can see from the Ljung-Box statistics that the residuals are independently and identically distributed as all the dots are above the line.

To conclude, The model fitted performs well except the violated normality assumption inspected from both the Standardized Residual plot and the Normal Q-Q plot. If we use this model to make predictions, the model may perform differently for different time period as the residuals are not normally distributed.



The model diagnostics for the second model fitted with the Log Transformed data performs much better than the previous model. We can see from the Standardized Residual plot and the Normal Q-Q plot that the normality assumption of the second model is not violated as there are no observable trend in these two plots. In addition, all the residuals are staying within 3 standard deviations and there are only few dots that are not on the straight line in the Normal Q-Q plot.

Although there are 1 significant spike in the sample ACF of the residuals, we can still conclude that the randomness assumption holds. And we can conclude the residuals are independently and identically distributed as all the dots are over the line in the Ljung-Box statistics.

Model Selection

Akaike Information Criterion

The AIC of a model is calculated as $AIC = 2k - 2\ln(\hat{L})$ where k is number of estimated parameters in the model while $\hat{L}$ refers to the maximum value of the likelihood function for the fitted model.[4] The smaller the AIC is, the better the model performs.

Information Criterion	Original model	Log Transformed model
AIC	17.82	0.11
AICc	17.82	0.11
BIC	17.92	0.19

We can see that the model fitted with the log transformed data have all the Information Criterion being much smaller than the other model. Then we will select the model fitted with the log transformed data as our final model.

Forcasting for the next 10 steps

The weekly number of confirmed cases of COVID-19 are expected to grow every week for the following ten weeks based on the prediction. It is also apparent that the further the prediction goes, the larger the confidence interval. That is, the predictions for the time that is further away from now tend to be less accurate. Please refer to Appendix for the table of predicted values as well as their 95% confidence interval.

Spectral Analysis of the data

Please refer to the following table for the first three predominant periods as well as their confidence intervals.

Series	Dominant.Frequency	Period	Lower	Upper
Confirmed Cases	3.0333	0.3297	0.1100566	6.427741
Confirmed Cases	0.8667	1.1538	0.3851479	22.494168
Confirmed Cases	2.6000	0.3846	0.1283826	7.498056

The significance of the Periods can not be estimated as all the periods are in the confidence interval of other periods

All the results generated above were programmed using `R version 4.0.5`

Discussion

Findings

According the predictions made by the ARIMA model fitted with the log transformed data, the weekly number of new confirmed cases of COVID-19 in Toronto is expected to grow slightly for the following ten weeks. Every individual in the society of UofT and even all of the residences aged 18 and above in Toronto should pay extra attention when involved in in-person activities. We are still in the half-way of the war against the pandemic brought by COVID-19, every individual should take their responsibilities to fight against the virus. People can easily make an effort by keep social distancing and wear a face musk, and the victory is in the near future.

Limitations

The data used to fit the model has only 110 observations, predictions made based on the model fitted could have a relatively lower accuracy. The data and the model fitted with

the data is time sensitive. Since the pandemic is evolving everyday, the predictions made based on the model fitted will only be accurate in the near future. It is also the fact that accuracy for the prediction starting for the third week starts to decrease dramatically. It would be better if the model can be updated weekly and only predict for next 2 to 3 weeks. Restricted to the small number of observations in the dataset, seasonality problem is failed to be accounted for in the model. The final model indeed have seasonality and this could lead to a decreased accuracy of the predicted value.

Next steps

As discussed in the limitations above, the key limitation of the model is that we have too few number of observations in the data. The original data set is updated weekly by the Toronto Public Health, researchers interested in similar topic can update the model fitted with the newest data. As the number of observations increase, the seasonality can also be accounted for in the model. Researchers interested in similar topics can also come up with other mathematical models to explain and predict the variation of the number of confirmed cases of COIVD-19 in Toronto, then check the accuracy of the predictions made based on the fitted model with the newest data updated.

Reference

1. Toronto Public Health, COVID 19: Vaccine Data. (2022). Retrieved 12 April 2022, from https://www.toronto.ca/home/covid-19/covid-19-pandemic-data/covid-19-vaccine-data/
2. Tamm, S. (2022). 10 Biggest Disadvantages of E-Learning - E-Student. Retrieved 12 April 2022, from https://e-student.org/disadvantages-of-e-learning/
3. Toronto Public Health, Covid-19 cases in Toronto, Open Data Dataset. (2022). Retrieved 12 April 2022, from https://open.toronto.ca/dataset/covid-19-cases-in-toronto/
4. Akaike information criterion - Wikipedia. (2022). Retrieved 12 April 2022, from https://en.wikipedia.org/wiki/Akaike_information_criterion
5. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
6. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
7. Yihui Xie, 2021, knitr: A General-Purpose Package for Dynamic Report Generation in R, R package version 1.34.
8. Brian Ripley et al, 2021, MASS: Support Functions and Datasets for Venables and Ripley's MASS, R package version 7.3-5.
9. David Stoffer, 2021, astsa: Applied Statistical Time Series Analysis (more than just data), R package version 1.14

Appendix

Table 3: Predicted Value for the next ten weeks with 95 Confidence Interval

Weeks	Lower Limit	Predicted Value	Upper Limit
1	1891.82	3075.51	4999.83
2	1124.69	3016.44	8090.11
3	667.48	3024.30	13703.01
4	402.89	3083.78	23603.52
5	249.42	3185.38	40680.38
6	158.86	3323.13	69515.13
7	104.13	3493.32	117192.41
8	70.17	3693.72	194430.61
9	48.53	3923.13	317138.00
10	34.38	4181.11	508532.68