Data 583 Project

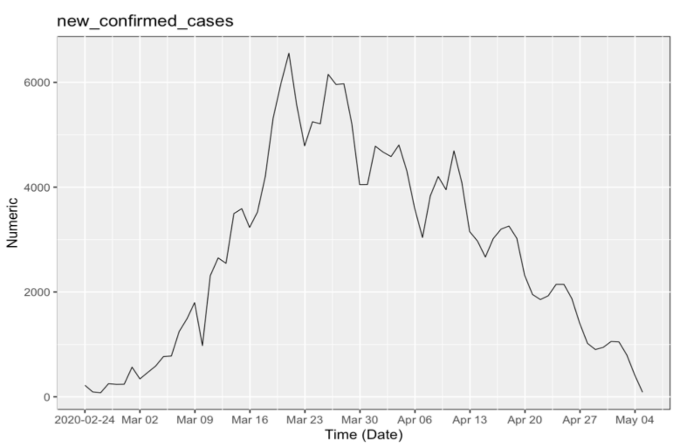
Chris Donoff, Elizaveta Volkova, Sofia Bahmutsky, Jasmine Chen

The time series analysis reported here focused on several countries that differed with respect to how far along they are in responding to the Covid-19 outbreak. Looking at multiple countries provided a more complete account of the virus' trajectory. Italy and South Korea posed as great case studies to use, as they have more data over a wider time window, enabling more accurate models of the Covid-19 infection rate over time. We used autoregressive parameters coming from these richer datasets on Canada and Russia- countries that are still early on in the spread of Covid-19. For all of our datasets, we scaled the daily number of new cases by the daily number of tests, to provide us a more accurate infection rate that accounts for testing.

Italy

The Italy dataset was transformed to make the data stationary, so we can detect potential factors that regression lines would otherwise miss. Plotting the autocorrelations after applying differencing revealed an autocorrelation at lag 7, guiding us to use an arima model with the AR parameter = 7, and differencing = 1. This resulted in the lowest AIC score, and enabled us to predict the future number of confirmed cases. Predictions from this model (20 days into the future) indicate that the number of confirmed cases would return to manageable levels by May 5th (Fig. 1A). With the standard errors associated with these predictions being quite high, this model would greatly benefit from the inclusion of additional variables that influence the variation of Covid-19 infection the greatest. To get a more accurate depiction of the rate of infection, we scaled the number of confirmed cases by the number of new tests performed each day. Interestingly, doing so vastly altered the interpretation of how well Italy is doing at mitigating the spread of infection. Scaling this data produced a trajectory that indicates Italy has already arrived at a manageable infection rate (Fig.1B).

A)



B)

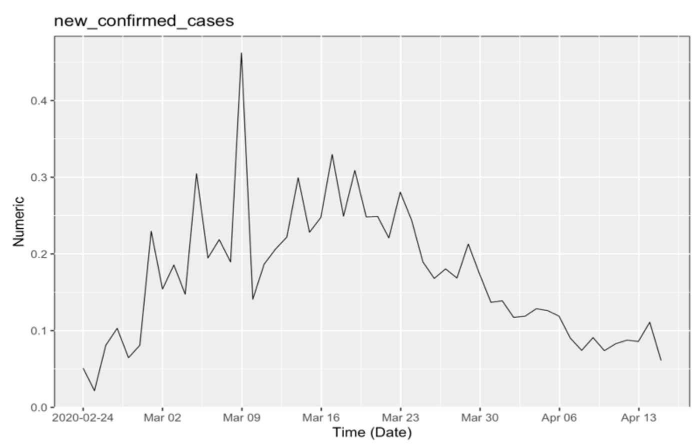


Figure 1. A) Predicted number of new confirmed cases from April 15 - May 5. B) Scaled number of confirmed cases by daily number of tests over time.

South Korea

South Korea responded quickly and decisively to covid-19 by taking an all-government approach and implementing widespread testing. Korea's success in 'flattening the curve' is largely due to its 'aggressive' testing and contact tracing, yet, without resorting to harsh measures such as lockdowns. Using the number of daily confirmed cases, AR = 2 model gives the lowest AIC score. Its 20-day prediction indicates that the number of confirmed cases would remain the level of zero. When scaling the number of daily confirmed cases by the number of daily tests, model with AR = 3, differencing=1, MA=1 gives the lowest AIC score. checkresiduals() function is used to estimate model adequacy. The residual ACF plot for the original data shows autocorrelation (>0.2) at lag=3, while the residual ACF plot for the scaled data shows no significant autocorrelation, thus, the model based on scaled data is more adequate. 20-day prediction based on the scaled model indicates that the infection rate would remain reasonably low, yet still remain around 0.1, despite the country's success in driving numbers down.

A) B)

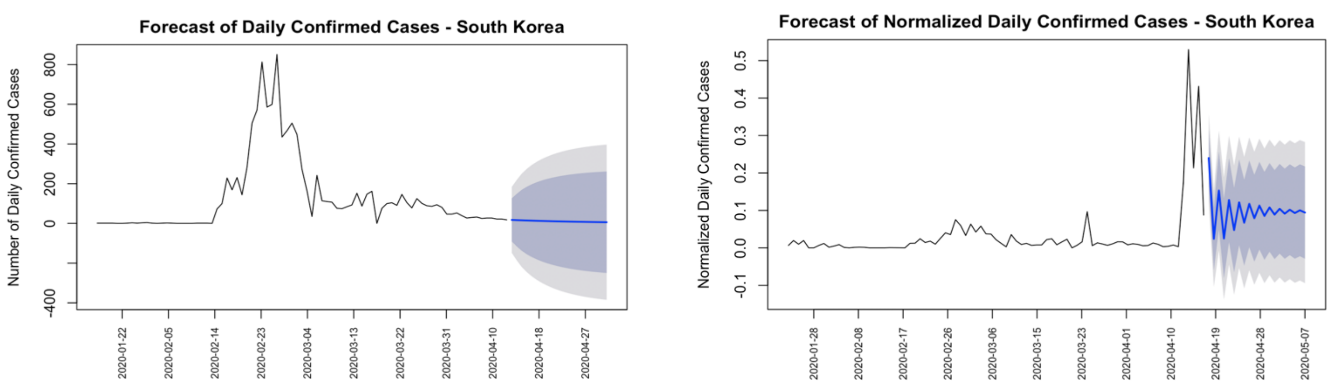


Figure 2. A) Predicted number of new confirmed cases from April 19 - May 7. B) Predicted normalized number of new confirmed cases by daily number of tests April 19 - May 8.

Canada

The first COVID-19 case in Canada was observed in Toronto on January 25, Canada in general was slow to implement social distance protocols, or any serious lockdowns until a few months later. Cases in Canada were quite low for the first 6 weeks, from the end of January to March, but due to the little early initiative, the number of cases quickly grew (reaching a maximum of approximately 2500 cases/day in the last week of march) and has still not reached a flattening of the “peak”. This trend is displayed in Fig. 3A. Since Canada’s population is widespread and not very dense, this may be a factor contributing to the poor preparation and lack of lockdown protocols in the earlier stages of the pandemic.

The shorter duration and lower amount of data in Canada caused issues when using ARIMA, since the predictions would follow the past pattern and a plateau would not be reached. We decided to implement the parameters from Italy and South Korea into the Canada model. After normalizing and differencing the daily cases for Canada, the auto-correlation plot was checked (ACF) to see any notable time lags, however no spikes were observed in the ACF plot. Based on previous knowledge from the ARIMA model of Italy and South Korea, an ARIMA(7, 2, 2) model was determined to be appropriate, based on AIC score, as well as the visual predicted output. The (7, 2, 2) model was used to predict the next 30 days of the proportion of positive tests in Canada over the total number of people tested (Fig. 3B). The predicted trend follows the pattern of the preceding data, it increases from about 0.1 to around 0.2 by the end of the prediction interval, however the uncertainty bands are very wide, so insight and predictions should be considered with caution. Using ARIMA for the Canada data at the given moment is not ideal, since it predicts a continued increasing amount of cases. In reality, based on other countries which have had earlier outbreaks, the amount of cases will eventually plateau and decrease given strict distancing measures, so it is just a matter of time before we see the same trend in Canada.

A) B)

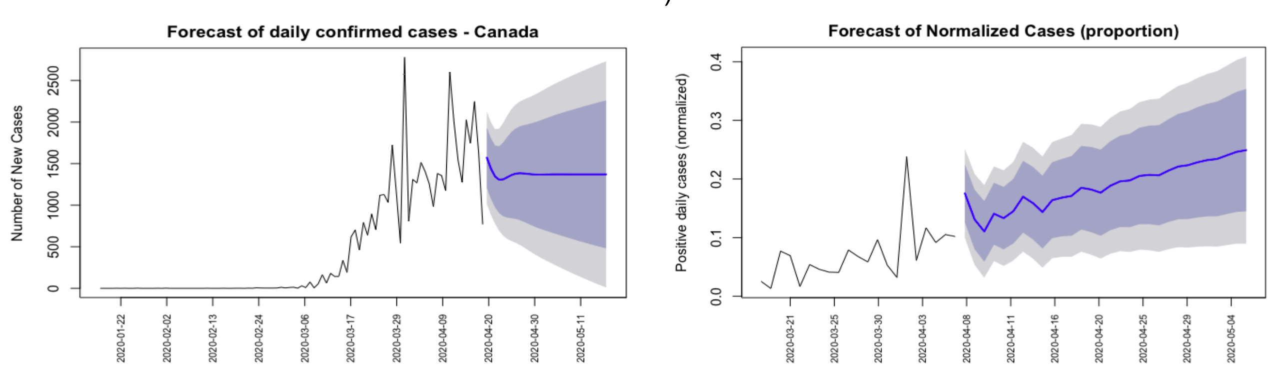


Figure 3A: Left-hand side panel shows the daily new cases confirmed in Canada, along with forecasting using an ARIMA(2, 1, 3) model. **3B:** Right-hand side panel shows the normalized version of the daily cases, followed by forecasting using an ARIMA(7, 2, 2) model. Blue area depicts 80% CI bands, and gray is 95% CI bands.

Russia

The first COVID-19 case in Russia was observed in the Moscow region on February 03. Russia in general was slow to implement social distance protocols, or any serious lockdowns until April 2nd, 2019. Cases in Russia were quite low until April 13th when they started to drastically increase from 174 cases per day to 5966 cases on April 25th with the highest rate of 6060 cases on April 19th. This trend is displayed on figure 4A. Since Russian population is widespread (same as Canadian population) and quarantine protocols (same to Italy) were introduced, this can be an explanation of an expectation of slow increase for the first period of the cases in the country.

To get a more accurate depiction of the rate of infection, we scaled the number of confirmed cases by the number of new tests performed each day (Fig. B) - as Russia is counted to be the second highest rate country by tests after the USA. The arima model for normalized number of daily cases by number of daily tests was performed with the same parameters as Italian data due to similarity of quarantine protocols introduced. Interestingly, doing so notably altered interpretation of the data: scaling the data with a model similar to Italy resulted in a higher rate of cases compared to the rate expected from the daily number of cases.

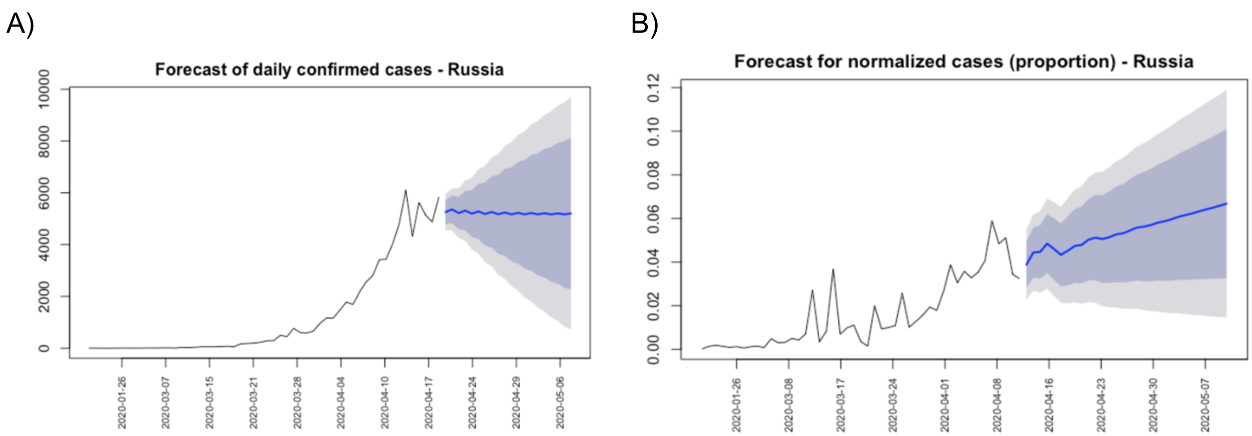


Figure 4: Left-hand figure depicts the daily new cases confirmed in Russia, along with forecasting using an ARIMA (2,1,3) model. Right-side figure depicts the normalized version of the daily, followed by forecasting using an ARIMA(7, 2, 2) model. Blue area depicts 80% CI bands, and gray is 95% CI bands.

To conclude, the ARIMA analysis of COVID-19 cannot be 100% accurate due to various factors, including population spread, the quarantine limitations introduced by authorities, population behaviour, number of tests introduced in each region of the country and also by accuracy rate of such tests. Such analysis can only be used as the expectation reference in case that each country would not introduce/decline any new limitations to their citizens.