Canada COVID-19 Trend Analysis and Model for Predictions

Manpreet Nanreh
mnanreh@ryerson.ca
Ryerson University
Toronto, Ontario

ABSTRACT

COVID-19 has had a great impact all over the planet and it spreading at a rapid rate in most countries such as US. Since the pandemic started, there have been great deal of fatalities. Therefore, it is important to analyze the data and determine the type of trend the spread of virus is having on the people. In this paper, the COVID-19 dataset was analyzed with the goal to perform trend analysis and determine models which would allow to make future predictions. The mathematical SIR model was analyzed in order to understand how the spread of virus depends on factors like disease transmission rate and recovery rate. It was determined that the logistic curve fits the data more than the exponential curve. The application of linear regression showed great results and proved to be a model worth picking if the goal is to make future predictions for COVID-19.

KEYWORDS

covid-19, sir model, curve fit, linear regression

1 INTRODUCTION

As of April 23, 2020, there have been total of 43,285 confirmed cases of COVID-19 and 2,240 fatal cases of COVID-19 in Canada.[3] There have also been tremendous amount of 14,761 recovered cases in Canada as of April 23, 2020. This first case of the pandemic was found in Wuhan City, Hubei Province of China on December 31, 2019 where the patient was informed to have symptoms of pneumonia. On March 11, 2020, the World Health Organization declared COVID-19 as a worldwide pandemic.[6] The symptoms of this viral disease include coughing, fever, difficulty of breathing and pneumonia in both lungs. It is recommended that if these symptoms start showing signs, then the person is recommended to isolate for 14 days as this is the incubation period of the virus.[3] This paper focuses on discussing the trend analysis this pandemic follows in Canada and how machine learning can be used to make future predictions.

2 RELATED WORK

There has been a lot of applications of machine learning on many different diseases and epidemics. The paper by Pouriyeh et al.[4] discusses the application of machine learning techniques such as Decision Tree, Naive Bayes, Multiplayer Perceptron, K-Nearest Neighbour, Single Conjunctive Rule Learned, Radial Basis Function and Support Vector Machine (SVM) on the heart disease from which they found that SVM allowed them to obtain the best results. This in turn gave the motivation to apply machine learning techniques to COVID-19 dataset.

Another paper by Ross and Hamer[7], introduced the mathematical SIR model which can be used to analyze directly transmitted infectious diseases. The SIR model can be represented by a system of quadratic ordinary differential equations as follows:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{1}$$

where $\beta > 0$ refers to the disease transmission rate and $\gamma > 0$ refers to the recovery which can also be represented in terms duration of infection $D = 1/\gamma$. In the equations above, S refers to the susceptible cases, I refers to the infectious cases and R refers to the number of recovered or deceased cases. The solution to this set of ODEs allows to determine the number of susceptible, infectious and recovered under specific theoretical conditions.

The paper by Hu et al.[2] showed the use of LSTM to analyze the trend of time series data. This paper provided techniques such as how to represent trend of the data and the application of moving windows for time series data.

3 DATASET AND FEATURE ENGINEERING

The dataset used in this paper was obtained from Johns Hopkins CSSE GitHub page[1] and is updated daily with information collected around the world. The dataset contains data staring from January 22, 2020 and is ongoing for different countries around the world. This study focuses on the use of time series data for Canada specifically where the dataset provides the number of cases on daily basis for each province. Since the focus is on the entire country, the data was combined where it shows the number of confirmed, recovered and fatal cases daily. As of April 23, 2020, US has the highest number of 869,170 confirmed cases which are more than double the number of confirmed cases in Spain thus far. Spain has the second highest number of 213,024 confirmed cases followed by Italy with the third highest number of 189,973 confirmed cases.

As of April 23, 2020, Germany has reported the highest number of recoveries thus far with 103,300 recovered cases. Spain has had the second highest number of 89,250 recovered cases and US with the third highest number of 80,203 recovered cases. There have been 77,983 recovered cases in China which makes it the fourth highest number of recoveries as it can be seen in **Figure 2**.

From **Figure 4** and **Figure 5**, it was found that the number of confirmed and fatality cases in Canada are rising as the days pass by. But it was also found that the number of recovered cases in Canada are also increasing as shown in **Figure 6**.

In order to prepare the dataset for machine learning techniques, the data was converted to logarithmic scale as this will the linear regression model to learn the data properly. After careful feature extraction and feature engineering, the following features were added to the dataset:

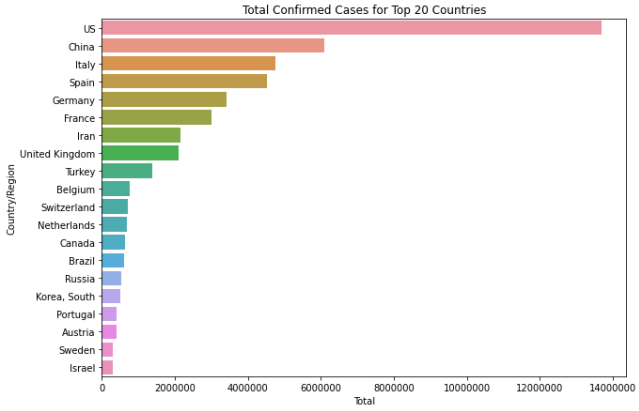

Figure 1: The total number of confirmed cases for top 20 countries.

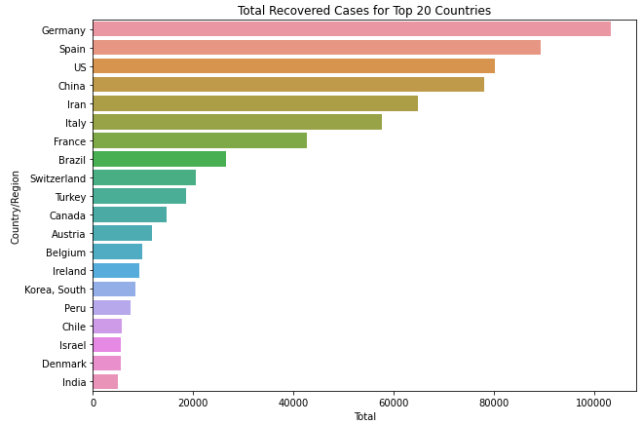

Figure 2: The total number of recovered cases for top 20 countries.

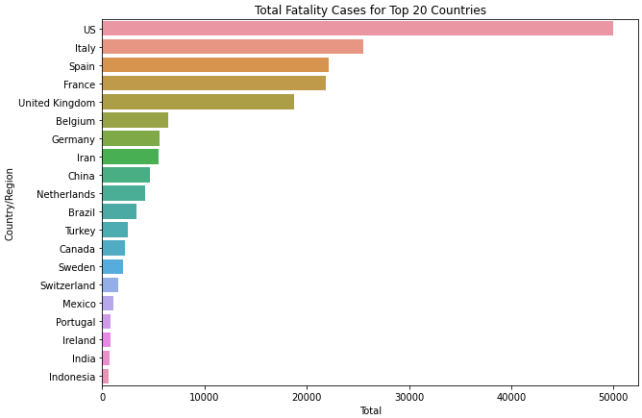

Figure 3: The total number of fatal cases for top 20 countries.

- year, indicating the year the report was submitted
- month, indicating the month the report was submitted

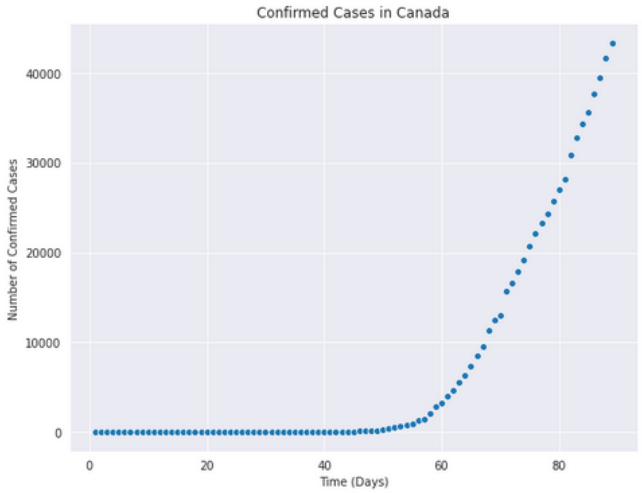

Figure 4: The number of confirmed cases in Canada over time.

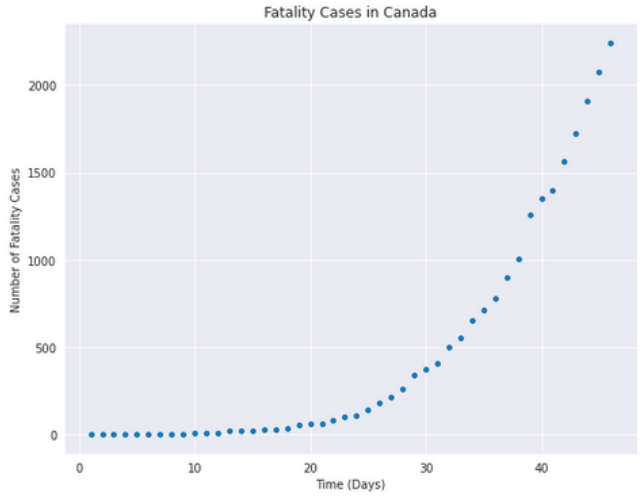

Figure 5: The number of fatality cases in Canada over time.

- day, indicating the day of the month the report was submitted
- day_count, indicating the count of the current day since recordings started
- is_weekday, indicating whether the day was a weekday
- log_cases, referring to logarithmic conversion of case count
- log_lag_i, referring to the logarithmic conversion of the lag of i days
- log_trend_i, referring to the logarithmic conversion of the trend observed in comparison to i days back

The lag was computed using the following equation: $log\_lag\_i = log(x(t) - x(t-i))$ where $x(t)$ refers to the number of cases reported on the current day being analyzed and $x(t-i)$ refers to the number of cases reported $t-i$ days in the past. The trend for the current day being analyzed by found using the following equation: $log\_trend\_i = log(\frac{x(t)-x(t-1)}{x(t-1)})$ where $x(t)$ refers to the number of

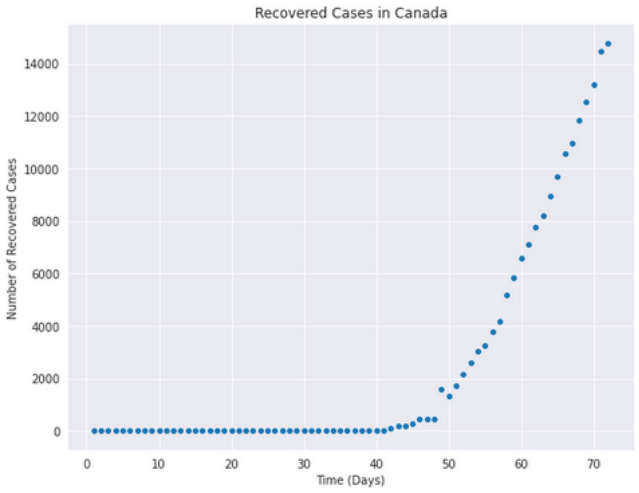

Figure 6: The number of recovered cases in Canada over time.

cases reported on the current day and $x(t-1)$ refers to the number of cases one day in the past. For this study, the trend and lag was computed seven days in the past. In addition to these created features, the population data for Canada was also added which was obtained from the Population Dataset created by Prabhu.[5] From the population dataset, the population size, density and median age was extracted, and combined with the COVID-19 dataset. This process was applied all three datasets of confirmed, recovered and fatality datasets. The final cleaned and modified dataset was then split into 80% training set and 20% testing set.

4 METHOD

4.1 SIR Model

The SIR model provided the tools to analyze the spread of directly transmitted COVID-19 virus. The model was analyzed in two different scenarios. The first scenario consisted the disease transmission rate (β) being 50% which means that how often does a susceptible person become affected when they come in contact with an infected person. The recovery rate (γ) was chosen to be $1/D$ where D was picked to be 5 days because 5 days is the average incubation period for COVID-19.

The second scenario analyzed the outcome when the disease transmission rate was reduced down to 30% and the recovery rate was kept same as 5 days. In both of these scenarios, the total population was set as the current Canadian population, the infectious count was set to the number of confirmed cases reported as of April 23, 2020, and the recovered/deceased count was set to the number of recovered cases added with number of fatality cases. The number of susceptible cases was set as the remainder of Canadian population which was neither infected or recovered or deceased. The model was solved using the SciPy library's integration function for set of ODEs.[8]

4.2 Curve Fit

Another question of this study whether the actual follows an exponential curve or a logistic curve. So, the goal here was to fit both of the curve and determine which fit was a better one. The measurement metric used here was mean squared error and R^2 value. The following exponential and logistic equations were used to fit the data:

$$f(x) = a * e^{b*x+c} \tag{2}$$

$$f(x) = \frac{L}{1 + e^{-k*(x-x_0)} + b} \tag{3}$$

These equations were optimized on the dataset using the SciPy library's curve fit function which optimizes the fit by minimizing the least squares error.[8]

4.3 Linear Regression

Linear Regression is a supervised training model which allowed the prediction of future number of reported cases for COVID-19 viral disease. As stated earlier, the dataset was feature engineered by creating new features and combining with the population dataset. Then the linear regression model was trained on 80% of the data and was tested on 20% of the data. The measurement metrics for the model were mean squared error and R^2 value.

5 RESULTS & DISCUSSION

5.1 SIR Model

The first scenario of SIR model provided the results as shown in **Figure 7**. It was found that when the disease transmission rate is 0.5 and recovery rate is 1/5, the number of infectious cases increase until day 24 where it encompasses approximately 23% of the total population and then it decreases down to almost zero given that the number of recovered or deceased cases increased over time. The susceptible population decreases quickly until approximately day 35 and after that it flattens out where it stays level around 10% of the total population. The number of recovered/deceased cases increase rapidly as the viral disease spreads which then starts to flatten out at approximately day 50 where it encompasses approximately 89% of the total population.

The second scenario for SIR model provided the results shown in **Figure 8**. This time it was found that with transmission rate of 0.3 and recovery rate of 0.2, the virus spread at a much slower rate. It was found that the number of infectious cases keep rising until day 55 where it encompasses 6% of the total population and then it starts to drop to zero. But at the same time the number of recovered/deceased rise until day 95 where approximately 55% of the total population. Interestingly the number of susceptible cases decrease gradually until day 90 where approximately 41% of the total population is susceptible and stays susceptible going forward. This is an interesting find because the viral disease does not fully disappeared.

5.2 Curve Fit

The curve fit of **Equation 2** can be visualized in **Figure 9** and the curve fit of **Equation 3** can be visualized in **Figure 10**. From visual

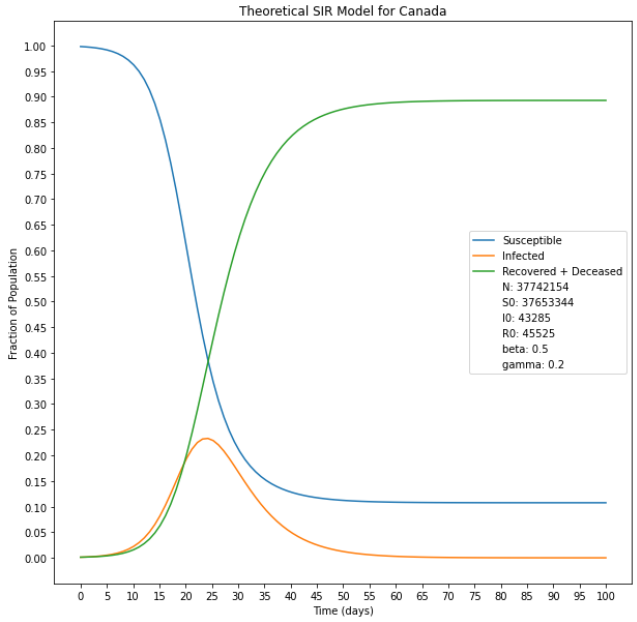

Figure 7: The SIR model with $\beta = 0.5$ and $\gamma = 1/5$.

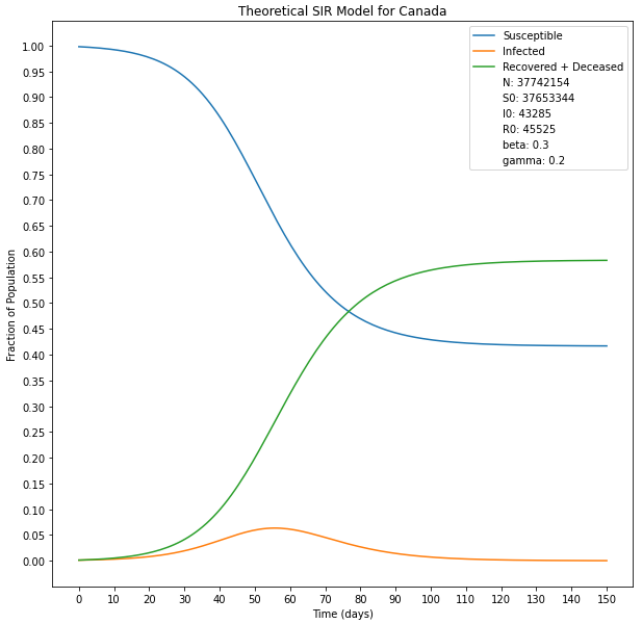

Figure 8: The SIR model with $\beta = 0.3$ and $\gamma = 1/5$.

analysis, it was found that logistic curve fit matches the actual number of confirmed cases more than exponential curve fit. It was found that the logistic curve fit had lower mean square error of 408411 when compared with the exponential curve fit which had mean squared error of 3190059. Also, the R^2 value for logistic curve was found at 0.99 and the R^2 value for exponential curve was found at 0.97. By comparing these measurement metrics, it can be stated that logistic curve is a better fit when compared with exponential curve. Another interesting discovering was that according to **Figure 10**, Canada has passed the inflection point which is represented by the red dot, and is approaching the flattening of the curve soon.

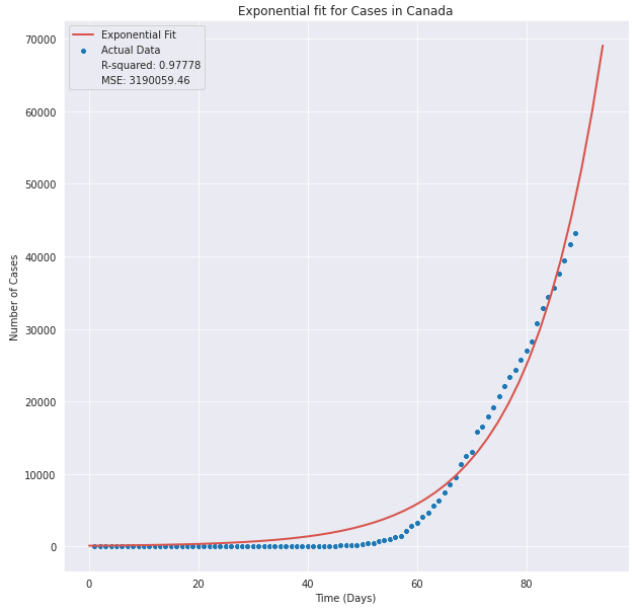

Figure 9: The exponential curve fit to the confirmed cases in Canada.

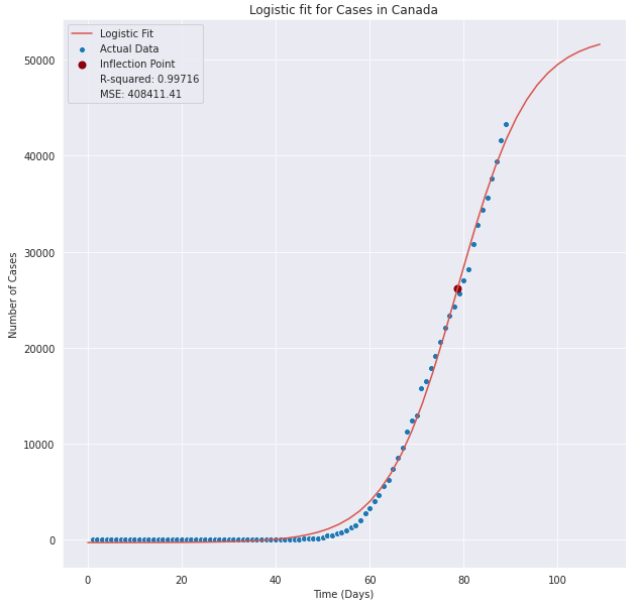

Figure 10: The logistic curve fit to the confirmed cases in Canada.

5.3 Linear Regression

Before applying linear regression, the data for confirmed, recovered and fatality cases was converted to the log scale because it will allow the model to get a better fit. This can be seen in **Figure 11**, **Figure 12** and **Figure 13** as the data can be seen to follow somewhat linear trend. Another interesting observation is that **Figure 11** shows a downward decrease in number of confirmed cases which may indicate that Canada may be close to flattening the curve soon. The model was trained on 80% of the data and tested on 20% of the data. The results obtained after applying linear regression to confirmed, recovered and fatality cases is summarized in **Figure 14**. The results showed that linear regression allowed to achieve low mean squared error and high R^2 score which indicates that this is a good model for this data. Therefore, it can be used to make future prediction on how the COVID-19 cases will increase in Canada.

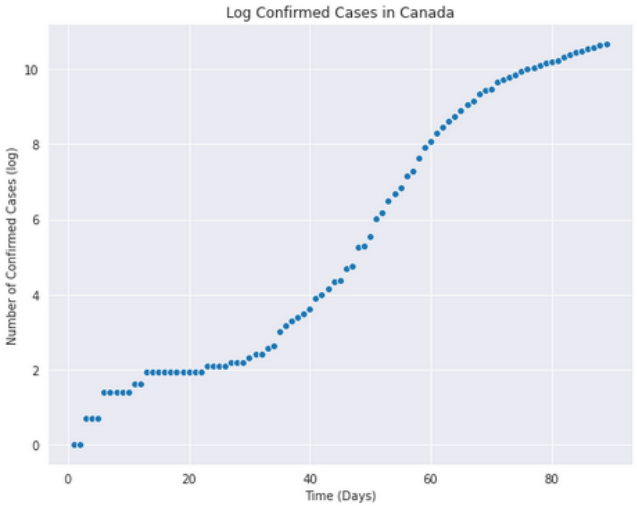

Figure 11: The number of confirmed cases in Canada over time on the log scale.

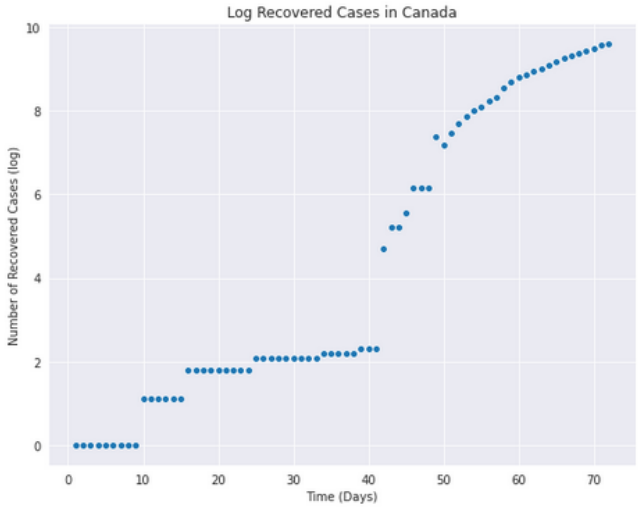

Figure 12: The number of recovered cases in Canada over time on the log scale.

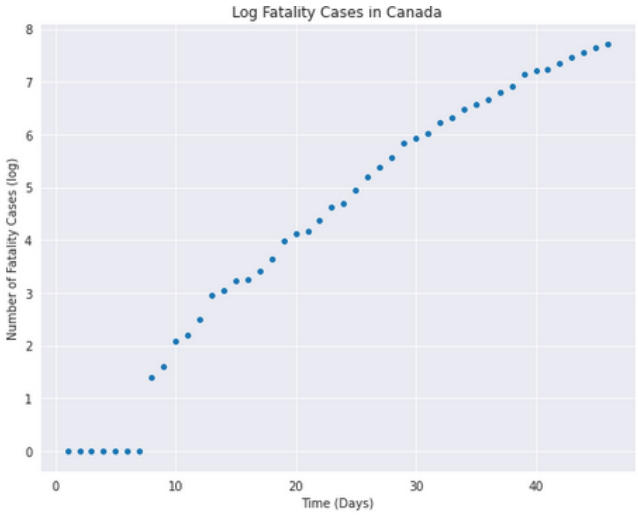

Figure 13: The number of fatality cases in Canada over time on the log scale.

6 CONCLUSION & FUTURE WORK

This paper discussed the analysis of theoretical mathematical SIR model which allowed to determine what kind of consequence would the spread of virus would have the number of susceptible, infectious and recovered/deceased population. Then the exponential and logistic curves were fit to the confirmed cases data and it was determined that logistic curve is a better fit to the data. This indicated that the actual data is likely to be following a logistic trend rather than an exponential in and this fit could be used to make future predictions on the number of confirmed cases to be expected. Lastly, linear regression was applied to all three pieces of data and it was determined to be a good model which can also be used to make future prediction and analyze the type of trend to be expected.

For future work, It would be beneficial to determine the type of distribution the data follows. In order to determine this, statistical bootstrapping can be used where randomly a sample will be picked from the distribution in question for k iterations and each iteration it will be determined how closesly does the sampled distribution matches the actual data's distribution.

REFERENCES

[1] Johns Hopkins CSSE. 2020. *2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE.* Johns Hopkins CSSE. https://github.com/CSSEGISandData/COVID-19.

[2] Y. Hu, X. Sun, X. Nie, Y. Li, and L. Liu. 2019. An Enhanced LSTM for Trend Following of Time Series. *IEEE Access* 7 (2019), 34020–34030.

[3] Government of Canada. 2020. *Coronavirus disease (COVID-19): Symptoms and treatment.* Government of Canada. https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/symptoms.html.

[4] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez. 2017. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. In *2017 IEEE Symposium on Computers and Communications (ISCC).* 204–207.

Data	Mean Squared Error	R2 Score
Confirmed Cases	0.00025	0.996
Recovered Cases	0.00187	0.990
Fatality Cases	0.00713	0.979

Figure 14: The summary of results obtained after applying Linear Regression to each of the dataset.

[5] Tanu N Prabhu. 2020. *Population by Country - 2020.* https://www.kaggle.com/tanuprabhu/population-by-country-2020.

[6] Infection Prevention and Control Canada. 2020. *Coronavirus (COVID-19).* Infection Prevention and Control Canada. https://ipac-canada.org/coronavirus-resources.php.

[7] Ronald Ross and William Heaton Hamer. 2013. The SIR model and the Foundations of Public Health.

[8] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2