Chenxi Yang, Yi Tang
DATA 583
April 13, 2020

Data 583 Report

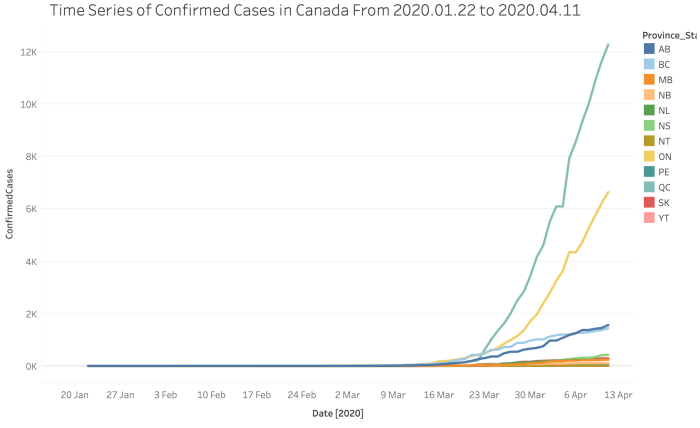
COVID-19 Virus Infection Prediction

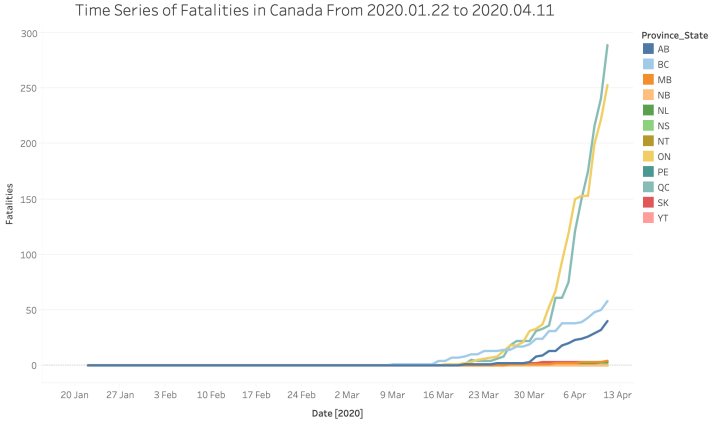
Currently, COVID-19 has been spread globally since the beginning of 2020, and as a result of this pandemic outbreak of COVID-19, many countries are affected severely by this situation economically and politically. Therefore, to assist medical and governmental institutions, it is necessary to predict the future trend of the number of confirmed cases and deaths.

All data related to this project are retrieved from Kaggle, which contains the number of confirmed cases and fatalities of 184 countries across the world from January 22, 2020, to April 11, 2020. In this project, we mainly focus on data of 10 provinces and 2 territories of Canada and predict the future trend of each of them until 31 May 2020.

I. Recent Situation Discussion

In the beginning, we analyze the number of confirmed cases and fatalities of each province. According to two plots below, lines of provinces that have a relatively higher number of confirmed cases and fatalities, especially Ontario and Québec, show the exponential curve.





II. Method Selection

Our objective is to find the best method for prediction. We choose 3 kinds of methods for comparison, including quasipoisson, linear and ARIMA models with different forms of parameters and degrees. The total dataset is divided into training data (22 January to 31 March) and test data (1 April to 11 April). The mean square errors of all models of both confirmed cases and fatalities are listed below, which are supposed to be as small as possible.

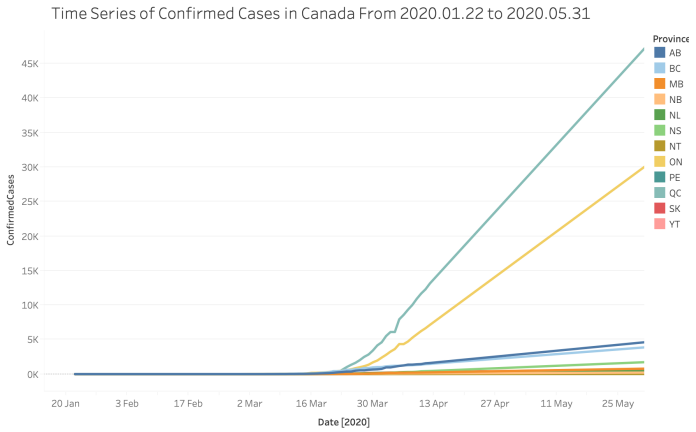
MSE of Multiple Models

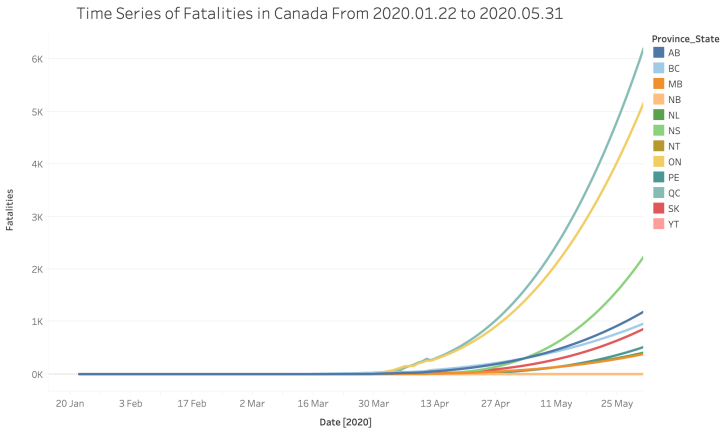
			Per Prov	Confirmed Cases	Fatalities
Quasipoisson	Logistic Regression	x	N	26677285.96	38543.76
		log(x)	N	5557214.46	1089.26
	Polynomial Regression	2nd deg	Y	4000657.2	3845.22
		3rd deg	Y	1.10967276411217E+276	Inf
	Polynomial Regression w/ Truncated Splines	2nd deg	Y	36794057.76	974772.1
		3rd deg	Y	1.10967276411217E+276	4.09549605684E+24
	Polynomial Regression w/ B-splines	1st deg	Y	24224976.75	1810.84
Linear	Regression	x	N	6809581.48	3042.17
		log(x)	N	7553420.07	3457.14
	Polynomial Regression	2nd deg	Y	1897788.42	1769.27
		3rd deg	Y	476392.19	1610.7
	Polynomial Regression w/ Truncated Splines	2nd deg	Y	79233456.78	1628.03
		3rd deg	Y	824635.33	1873.26
	Polynomial Regression w/ B-splines	1st deg	Y	1233624.11	2313.53
ARIMA	Auto-Arima		Y	127209.58	1664.98

According to the result shown above, the method that has the smallest is supposed to be chosen. For predicting the number of confirmed cases, the auto ARIMA model is chosen; for predicting the number of fatalities, the quasipoisson logistic regression with log(x) is chosen. Quasipoisson regression does well in expanding standard error while Poisson regression would lack significant parameters when counting data.

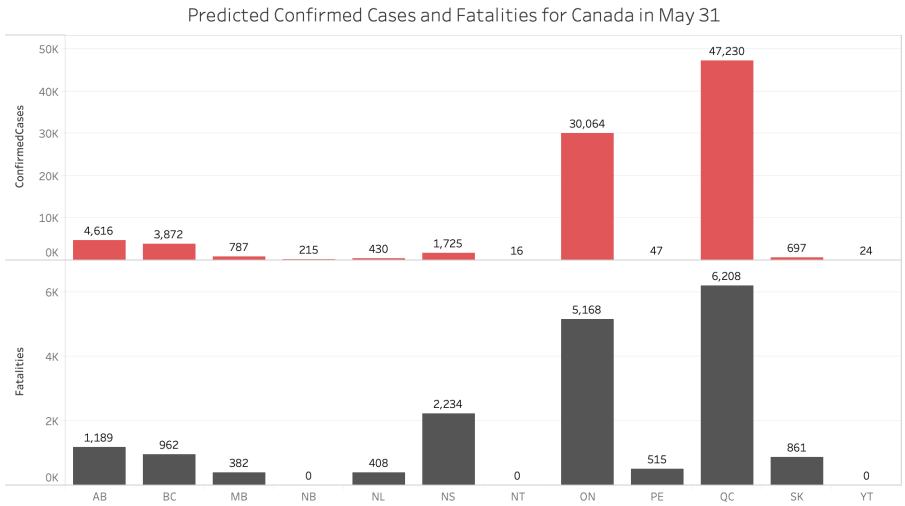
III. Prediction

Methods selected above are applied to refit models using dataset from January 22, 2020 to April 11, 2020, and use new models to predict the number of confirmed cases and fatalities from now on to 31 May 2020. All predicted amounts are shown in two line charts below.





As can be seen from the graphs, all curves depict the upward trends, which indicates within the next month, there is no considerable decrease in the amount of both confirmed cases and fatalities. Therefore, based on our current models, it is hard for the number of cases to decrease to manageable levels. However, building new models when there are more data available may result in more optimal results.



Focusing on 31 May 2020, at the end of that day, the number of confirmed cases and fatalities for each province across Canada are shown in the bar chart above. The bar chart illustrates that Québec and Ontario are still two regions that have a dominating number of confirmed cases and fatalities, which are predicted to have 47,230 and 30,064 confirmed cases, and 6,208 and 5168 fatalities respectively.