Estimating the effects of non-pharmaceutical interventions on COVID-19 in Canada

Author
Olivia Z. Shi

Supervisor
Pr. David Stephens
Pr. Christian Genest

McGill University
Master thesis
Student ID: 260738642

June 30, 2021

1 Abstract

We aim to estimate the effects of non-pharmaceutical interventions such as government interventions on COVID-19 in Canada. The public interventions included in our model are Closures/openings and social distancing. Short-term modelling forecasts provide time-critical information for decisions on containment and mitigation strategies. A major challenge for short-term forecasts is the assessment of key epidemiological parameters and how they change when first interventions show an effect. By combining Susceptible-Explored-Infectious-Recovered (SEIR) model with a Negative Binomial regression model, we analyzed the time dependence of the effective growth rate of new infections. Focusing on COVID-19 spread in Canada, we detected change points in the effective growth rate that correlate well with the times of publicly announced interventions. The key aim of these interventions is to reduce Rt, a fundamental epidemiological quantity that represents the average number of infections generated at time t by each infected case throughout their infection. Thereby, we could quantify the effect of interventions and incorporate the corresponding change points into forecasts of future scenarios and case numbers. Our code is freely available and can be readily adapted to any country or region.

2 Introduction

The novel coronavirus (COVID-19) was widely reported to have first been detected in Wuhan (Hebei province, China) in December 2019. It very quickly spread to other countries within and outside of Asia. At present, over 164 million cases of infected individuals have been confirmed in over 180 countries with more than 3.4 million deaths. In Canada, over 1.34 million cases of individuals have been confirmed with over 25,045 deaths. It spread much more easily than the severe acute respiratory syndrome (SARS) virus.

Similar research includes: Analysis in Italy and Spain using SIR and log-linear model(Jeffrey Chu, 2021).

In 2020, Non-pharmaceutical interventions (NPIs) are the only option to delay and moderate the spread of the virus. Confronted with the worldwide COVID-19 epidemic, most governments had to take decisions under rapidly changing epidemiological situations, despite (at least at the very beginning of the epidemic) a lack of scientific evidence on the individual and combined effectiveness of these measures degree of compliance of the population and societal impact. A wide range of nonpharmaceutical interventions (NPIs) has been deployed, including stay-at-home orders and the closure of all nonessential businesses. Recent analyses show that these large-scale NPIs were jointly effective at reducing the virus's effective reproduction number Rt (1), but it is still largely unknown how effective individual NPIs were.

In Canada, the last year has seen severe financial stress for businesses due to lockdowns, customer reluctance, and travel restrictions. The two tiers of government have put in place schemes to mitigate the stress. But have they been effective? Acquiring knowledge of the most effective NPIs would allow stakeholders to timely implement interventions to combat a

resurgence of COVID-19 or any other future respiratory outbreak. Because many countries rolled out several NPIs simultaneously, the challenge arises of disentangling the impact of each intervention. In this research, we use various statistical techniques to model and analyze the effectiveness of government interventions. We used a Negative Binomial regression to model the individual effects.

The main contributions of this paper are: i) to use Negative-Binomial and SEIR models to model the incidence of COVID-19 in Quebec, Canada; ii) to provide estimates of basic measures of the effectiveness of the NPIs of COVID-19 in Quebec, Canada; iii) to investigate the predictive ability of our models and provide simple forecasts for the future incidence of COVID-19 in Quebec, Canada.

3 Data

The three main source we collected our data are Government of Canada website [1], Canadian Institute of Health Informatics (CIHI) [2], and Institut national de santé publique du Québec (INSPQ) [3].

The data used in this analysis consists of the daily and cumulative incidence (confirmed cases) of COVID-19 for Canada and its 10 autonomous provinces, from 1st January 2020 to 18 May 2021, inclusive.

3.1 Effective Reproduction Number

The effective reproduction number, Rt, tracks the number of other people a single infected person is likely to infect. The INSPQ's Rt values have a 10-day lag because they depend not on the date of a reported positive test results but the likely date a person was infected. The numbers are updated weekly, Maheu-Giroux said, because day-to-day trends aren't as meaningful as trends established over longer periods of time. During a technical briefing for journalists Thursday, Maheu-Giroux cautioned that Rt has shortcomings, and needs to be considered in concert with other measurements, but it nonetheless provides another way to understand the status of the pandemic.

3.2 Actual Cases

Notice the outliers in the QC graph. The outliers is not because they have a lot of cases on that given day, rather it is because the recorder realized that they had missed some cases hence added them all on one day. To achieve a more accurate result, we need the actual count of data instead of the reported count, the data after fixing the initial data entry mistakes.

We have considered introducing stochastic to our data, or smooth the data.

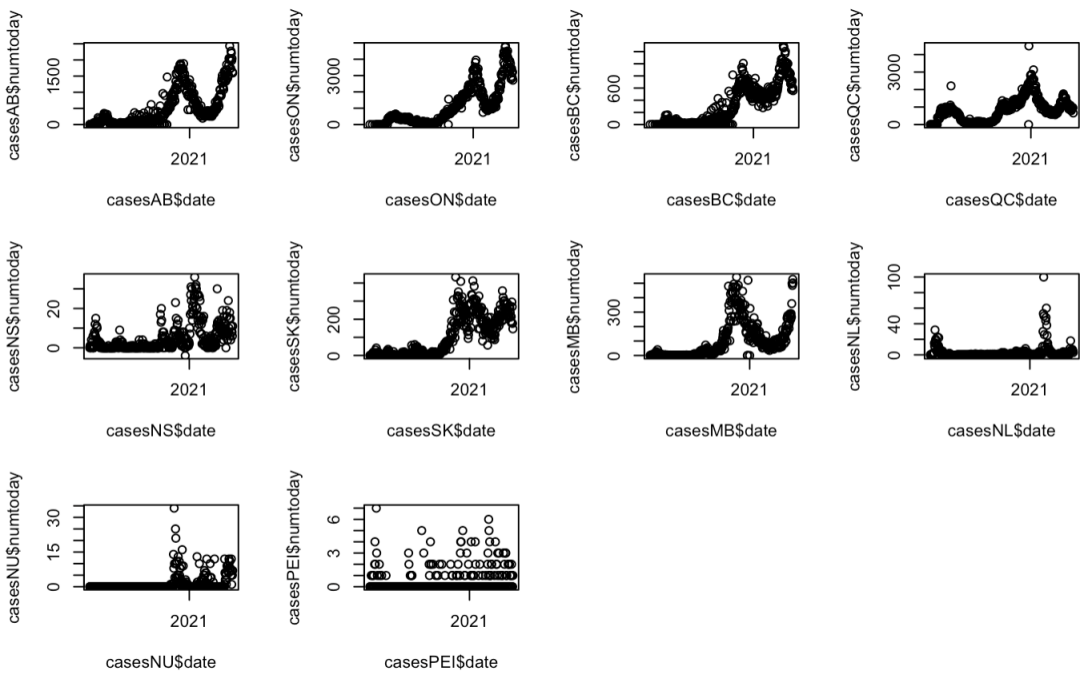


Fig1. Daily cumulative incidence of the 10 provinces in Canada for the period of 21/02/2020 to 28/03/2020, inclusive.

3.3 Data Processing Softwares

R studio software Version 1.2.5001 was used to import, clean and process the data. Anylogic was used to implement the simulation and calibration of the SEIR model. Anylogic is a dynamic modeling application based on java, which allows the user to combine different techniques and approaches such as differential equations, discrete events and agent based systems.

For the GLM part, I imported and cleaned the COVID cases data from Statistics Canada [1] and the intervention data from CIHI [2]. Then, I joined the two datasets based on dates. Thus I get a table with dates, cases per day, and intervention categories. I fitted a GLM using this table.

For the SEIR part, I imported the Rt data from INSPQ [3] using R and uploaded it into Anylogic. I used R to calculate the mean values for death rate and recovery rate based on the data provided by Statistics Canada.

3.4 Interventions data

We got the data of Government interventions from Health Canada. In the following figure, the solid dot means impose a restriction, the open dot signifies ease a restriction.

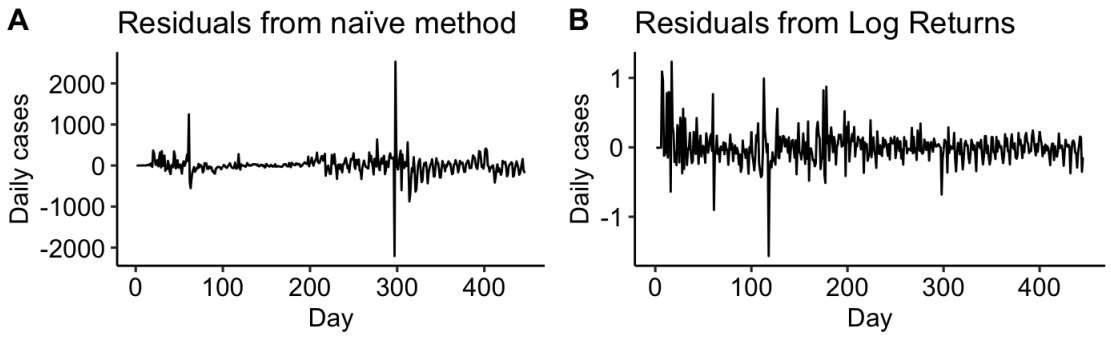
3.5 Time Series

Moving averages are a simple and common type of smoothing used in time series analysis and time series forecasting.

We used time series methods to process our counts data. First, we looked at the acf and pacf for numtoday. The acf has a significant lag 1 so we fit a MA(1) model. We used the ARIMA function in R with degree of differencing to be 1.

We then decided to look at the log returns. Log return is the log number of cases for today subtract the log number of cases for the previous day. It has a slowly decreasing acf, and a pacf of 1 significant lag. Hence we fit a ARIMA(0,0,1) model. We fit the acf again and realized that it has a seasonal effect with period 7 and adjusted it. The residual of the fitted model looks like figure B.

As shown in the graph that the residuals from the naive method and from the log returns method looks the same, which confirmed our trying. We successfully limited the residuals to around -1 and 1.



The equations for the time series are:

$ARIMA(p,d,q) * (P,D,Q)_s$

p=0,d=0,q=1; P=1,D=0,Q=0;S=7

$$\Phi(B^S)X_t = \Theta(B)\Pi(B^S)Z_t$$
$$(1 - \Phi B^7)X_t = (1 - \Theta_1 B)Z_t$$
$$X_t = \Phi X_{t-7} + Z_t - \Theta Z_{t-1}$$

4 Methods

Generalized Linear Models (GLMs) are models in which response variables follow a distribution other than the normal distribution. A Poisson regression model is a GLM that is used to model count data and contingency tables. The output Y (count) is a value that follows the Poisson distribution. It assumes the logarithm of expected values (mean) that can be modelled into a linear form by some unknown parameters. We generated a Poisson Regression model based on the resulted dataset. In the regression model, the baseline is "None", which means when no interventions is significant.

A loglinear model for the expected rate has the form $log(\mu_i/t_i) = \sum \beta_j x_{ij}$. Because $log(\mu_i/t_i) = log(\mu_i) - log(t_i)$, the model makes the adjustment $-log(t_i)$ to the log link of the mean. This adjustment is called an offset.

4.1 Negative Binomial

The Poisson and negative binomial regression models are widely used for modelling discrete count data where the count takes a non-negative integer with no upper limit, while the data is highly skewed. The negative binomial regression has the added advantage of being able to deal with the problem of overdispersion. From our result of the Poisson Rregression, we noticed that the residual deviance massively larger than the degree of freedom, which means overdispersion. Therefore, we chose to implement the negative binomial model.

However, the residual plot does not look random, which means there are significant confounding effects not captured by the model. Stephen Chan (2020) pointed out that "The best fitting count regression model for modelling the number of new daily COVID-19 cases of all countries analysed was shown to be a negative binomial distribution with log link function"

We are aware that the COVID symptoms takes up to 14 days to develop. Therefore, we incorporated a lag variable into our model that takes input from 1 to 14. To incorporate the effect of lag, we adjusted the cases date to incorporate a delay of 1 to 14 days and generated tables for each lag using a for loop.

4.2 The SEIR (Susceptible-Explored-Infectious-Recovered) model

Predictive mathematical models for epidemics are fundamental to understand the course of the epidemic and to plan effective control strategies. Compartmental models can describe the spread of a disease within a population. SEIR model describes the flow of individuals through four mutually exclusive stages of infection: susceptible, exposed, infected and recovered. It can be used to compute the infected population and the number of casualties of this epidemic.

4.2.1 Simulation

We made some modifications to the standard SEIR model:
1. A standard SEIR model assumes that the recovered group can no longer be infected with the disease, which is not the case for COVID - the recovered group can still get infected again. Hence we incorporated a lostImmunity parameter and an Immuned compartment.
2. The birth rate replenishes the 'susceptibles' and allows for multi-wave phenomena if the infection rate is fixed across time, but since the infection rate is allowed to vary across time, you don't need the births.
3.We calculated the detection rate by dividing ratetested by ratetests. The first 20 values of detection rate is greater than 1, which we suspect may due to data collection errors in the early stage. We thus started using the data since April 1, 2020.

We use compartments Susceptible S(t), Exposure E(t), Infectious I(t), Recovered R(t), and Immuned I(t) to denote the number of individuals in the five groups mentioned above, as functions of time t.

The COVID-19 dynamics is modelled by the following system of four differential equation:

$$\frac{dS}{dt} = 8485000 - \underbrace{\beta I S \frac{1}{N}}_{infection} + \underbrace{\omega R}_{lostimmunity}$$

$$\frac{dE}{dt} = \underbrace{\beta I S \frac{1}{N}}_{infection} - \underbrace{\sigma E}_{latency}$$

$$\frac{dI}{dt} = \underbrace{\sigma E}_{latency} - \underbrace{\gamma I}_{recovery} - \underbrace{\mu + \alpha I}_{death}$$

$$\frac{dR}{dt} = \underbrace{\gamma I}_{recovery} - \underbrace{\omega R}_{lostimmunity} - \underbrace{\mu R}_{death}$$

$$R_0 = \beta/(\gamma + \alpha)$$

The below table shows the parameter values used in our model.

name	value	description
β	0.2	infection
ω	1/365	lost immunity
σ	1/21	latency
α	0.01	death due to infection
γ	1/14	recovery
N	8485000	total population

One objective is to find the basic reproduction number, R_0. Stability of the equilibrium occurs for $R_0 > 0$, representing an immediate recovery with no large involvement of the population. Larger values of R_0 imply a strong affection of the population according to equation.

4.2.2 Estimating Rt

Rt is the average number of secondary cases that each infected individual would infect if the conditions remained as they were at time t.

Time-Varying Reproduction Numbers. The first aim of the SEIR model is to estimate Rt. We used a generic and robust tool for estimating the time-varying reproduction number developed by XXX.

4.2.3 Calibration

For the calibration process, we used getbestpara function in Anylogic, which returns the value of the given optimization parameter variable for the best solution found thus far. It may be infeasible. The optimization objective is to minimize difference(root.InfectiousDS, root.InfectiousHistory). That is, we choose the best parameter set by minimizing the sum of squared residuals (ie. use least squares).

Since we have the data for deathRate, RecoveryRate and InfectionRate, we fix those values to find optimized values for latency, lostImmunity. After, we use the optimized values for latency and lostImmunity to simulate a model. This allows us to visualize how the model we have implemented matches the data (deaths/cases series) that we have already collected.

4.2.4 Piecewise R0

Our first attempt is to get change points from interventions and get piecewise Rt values.

Choosing change point

4.2.5 change point analysis

Change-points are abrupt variations in time series data that may represent transitions that occur between states. A posteriori change-point detection problems have been of interest to statisticians for many decades. Although, naturally, details vary, a theme common to many of them is as follows: a time-evolving quantity follows a certain stochastic model whose parameters are, exactly or approximately, piecewise constant. In such a model, it is of interest to detect the number of changes in the parameter values and the locations of the changes in time. Such piecewise-stationary modelling can be appealing since i) the resulting model is usually much more flexible than the corresponding stationary model but still parametric if the number of change-points is fixed; ii) the estimated change-points are often 'interpretable' in the sense that their locations can be linked to the behaviour of some exogenous quantities of interest.

From the observed case numbers of COVID-19, we can quantify the impact of these measures on the disease spread using change point analysis. A change-point analysis is performed on a series of time ordered data in order to detect whether any changes have occurred. It determines the number of changes and estimates the time of each change.

A traditional way is to use Binary Segmentation to determine changepoints in the historical infection data. Binary Segmentation is a method for identifying changepoints in a given set of summary statistics for a specified cost function and penalty. Fryzlewicz proposed the wild binary segmentation(wbs) for consistent estimation of the number and locations of multiple change-points in data.

We also attemped to determine change point in the simulated infection data. That is, to incorporate the number of changepoint as a parameter to optimize in R, using the optim function.

4.2.6 Dynamic Rt

Our second attempt is to estimate Rt as a dynamic variable.

4.2.7 SEIRSO

We made a SEIRSO function and the 'optim' function in R to estimate the time-variant cases number of each compartment and the most optimal change-points, along with their most optimal parameters. The input of the SEIRSO function are cases count (Ccount),

death count (Dcount), Nval=38e6 , and the number of change points (K).

The SEIRSO function returns the sum of the squared difference between the actual case count (cv) and the estimated case count (It) plus the actual death count (dv) and the estimated death count (Dt). Notice here that we divided the cases by 3600 in order to match the scale of the death count.

$$\sum_{t=1}^{400}(It - cv^2)/3600 + \sum_{t=1}^{400}(Dt - dv^2)$$

The 'optim' function vary the parameters and change-points locations to get the most optimal parameter values and change-points locations by comparing the SEIRSO output.

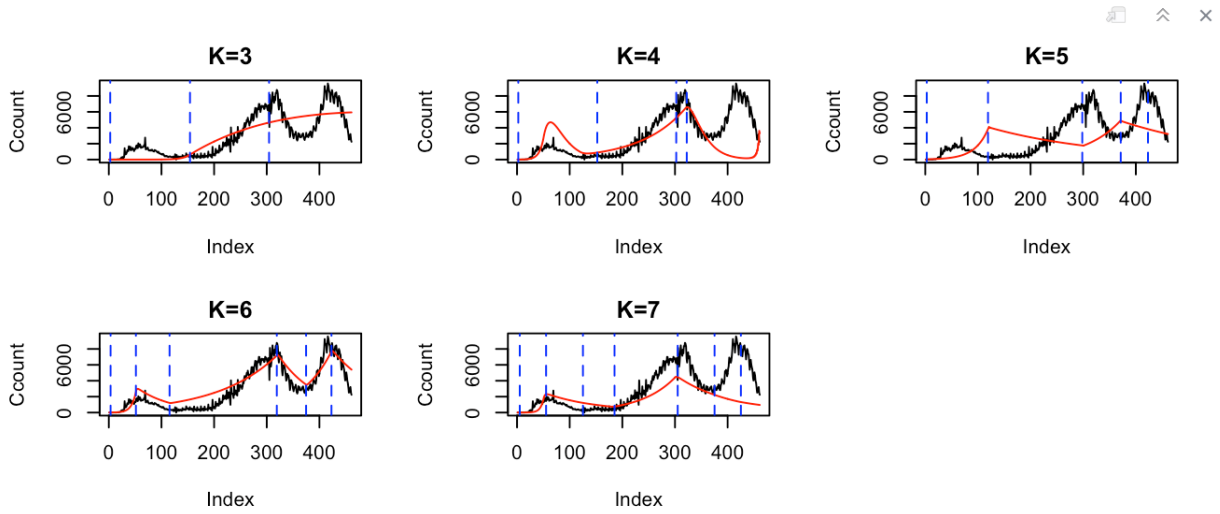


Fig. SEIRSO result

The red curve is the estimated infected cases. The black curve is the actual infected cases. The blue dotted lines indicate the change-points. We also wrote a function that computes the difference between the estimated cases and actual cases plus the difference between the estimated deaths and actual deaths. The result are as follows:

diffcd(Ccount,Dcount,I3) = 7208687655
diffcd(Ccount,Dcount,I4) = 7005677527
diffcd(Ccount,Dcount,I5) = 6971969569
diffcd(Ccount,Dcount,I6) = 8207780959
diffcd(Ccount,Dcount,I7) = 4674017615

Having 7 change-points is the most optimal by our algorithm.

5 Results

5.1 Negative Binomial Model

```
Coefficients:
                                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                                     6.476e+00  6.993e-02  92.615  < 2e-16 ***
lagged_table$numactive                          2.404e-05  4.082e-06   5.888  3.9e-09 ***
lagged_table$Intervention.categoryCase management -1.777e-01  1.926e-01  -0.922   0.3563
lagged_table$Intervention.categoryClosures/openings 5.841e-02  9.679e-02   0.603   0.5462
lagged_table$Intervention.categoryDistancing    7.674e-02  1.439e-01   0.533   0.5938
lagged_table$Intervention.categoryHealth services 3.902e-02  1.538e-01   0.254   0.7997
lagged_table$Intervention.categoryHealth workforce -2.794e-01  1.120e-01  -2.495   0.0126 *
lagged_table$Intervention.categoryPublic information 3.572e-02  1.405e-01   0.254   0.7994
lagged_table$Intervention.categoryTravel        4.171e-01  2.149e-01   1.941   0.0523 .
lagged_table$Intervention.categoryTravelease    1.396e-01  2.026e-01   0.689   0.4910
lagged_table$Intervention.categoryVaccine       6.738e-01  2.621e-01   2.571   0.0101 *
```

Figure 1. Negative Binomial Regression Model without considering lag

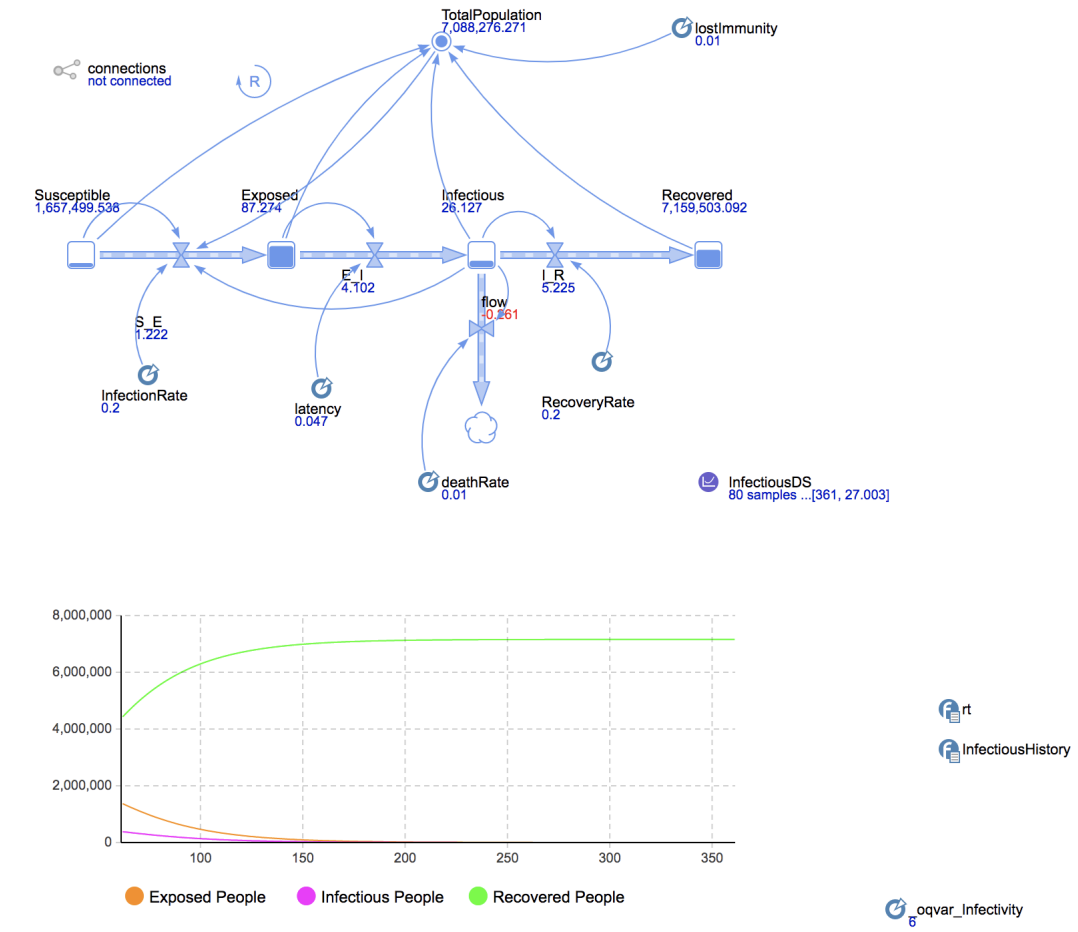
A negative estimate value means that it is reducing cases, a positive estimate value means that it is increasing cases, the magnitude of the effect is captured by the size of the estimate. The "Closures/openings" category has an estimated value of -1. "Distancing" has an estimated value of -6, which means that distancing is very effective in reducing cases. From Figure 1. we can see that vaccine has a positive estimate, meaning that the vaccine correlates with increasing cases after 14 days. Closures/openings, Distancing and State of emergency are effective in reducing cases. The p-value for every intervention is significant because they actually play a role in the count.

The model shows that government interventions such as quarantine of contacts and isolation of cases can help halt the spread on novel coronavirus.

5.2 SEIR Model

From the SEIR model, we found an R_0 of 2.03, implying that the pandemic will persist in the human population in the absence of strong control measures.

For any parameter set we can trace out hypothetical trajectories for I(t), D(t) and R(t) by solving the system of ODEs. We cannot do this analytically, but we can do it numerically. We want to choose the best parameters to match the observed data series i(t), d(t), r(t) that are observed (daily, weekly ?) as closely as possible; this is a curve fitting exercise. We can choose the best parameter set by minimizing the sum of squared residuals (ie use least squares)



5.2.1 Piecewise Rt

timestemp	InfectionRate	DeathRate	RecoveryRate	R_0
0-150	0.011	0.04		0.275
150-350	0.573	0.03		14.325
350-450	0.112	0.03		2.8

6 Discussion

In this paper, we have provided a simple statistical analysis of the novel Coronavirus (COVID-19) outbreak in Quebec, Canada. Using data of the daily and cumulative incidence over approximately the first month after the first cases were confirmed, we have analyzed the trends and modelled the incidence and estimated the basic reproduction value using two common approaches in epidemiology — the SEIR model and a negative binomial model.

We used a data-driven approach to estimate the effects that seven nonpharmaceutical interventions had on COVID-19 transmission in Quebec, Canada between January 2020 and the end of May 2021. We found that several NPIs were associated with a clear reduction

in Rt, in line with mounting evidence that NPIs are effective at mitigating and suppressing outbreaks of COVID-19. Furthermore, our results indicate that some NPIs outperformed others. While the exact effectiveness estimates vary with modelling assumptions, the broad conclusions discussed below

7 Competing Interests

The authors declare that they have no competing interests.

8 Code Availability

Our code is available in Github $https://github.com/olixbridge/COVID_RESEARCH.git$

9 Acknowledgements

First and foremost, I would like to thank Professor David Stephens and Professor Christian Genest for supervising me on this master thesis, and for their valuable guidance. I am very grateful to Dr. Ivo Panayatov for helping me interpreting the data.

I also would like to thank my colleagues and friends: Julian Osorio for their support on interpreting the data; X and X for their support with Matlab software; A,B for offering valuable advice to improve my paper.

10 References

[1]government of canada

[2] https://www.cihi.ca/en/covid-19-intervention-timeline-in-canada

[3]INSPQ

[?] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0249037sec001

https://nccid.ca/wp-content/uploads/sites/2/2020/06/Math-Modelling-Summary˙JUNE24.pdf

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0245101sec002