COVID-19 DATA ANALYSIS BASED ON THE SEIR MODEL IN NOVA SCOTIA

by

Xing Wang

Submitted in partial fulfillment of the requirements
for the degree of Bachelor of Science Combined Honours in Statistics and Economics

at

Dalhousie University
Halifax, Nova Scotia
December 2021

Table of Contents

Abstract

Novel coronavirus (COVID-19) is a coronavirus of genus that causes a disease with predominantly inflammatory lung lesions, leading to a series of respiratory disease symptoms similar to SARS, such as dyspnea, respiratory distress syndrome or septic shock, increasing the probability of admission of patients to the intensive care unit (ICU) and morbidity and mortality. Since the discovery of COVID-19 in Canada in March 2020, the outbreak has spread rapidly. In order to quickly contain the spread of the disease, the country has introduced various initiatives, such as: delaying the return to work for companies, delaying the opening of schools, restricting travel, and quarantining at home. In this paper, we attempt to study the epidemic condition in Nova Scotia by collecting the relevant data from March 2020 to July 2021 and fitting them with the widely-used SEIR model. We analyze two different waves that occurred during this period in Nova Scotia separately, and we compare them with each other based on the estimated basic reproduction number R_0, and other parameters. Our results show that the two waves of epidemics that Nova Scotia has experienced are well captured by the SEIR model, yielding reasonable values of R_0 for both waves with the third wave worse than the first.

Acknowledgements

I would like to take this opportunity to express my gratitude to my supervisor Dr. Lam Ho, whose significant support and encouragement helped me excel and reach my academic goals. I have been very fortunate to conduct research under your guidance. Thank you.

Chapter 1

Introduction

The battle between humans and infectious diseases has been waged from ancient times to the present. From the ancient Black Death to the recent AIDS, and from influenza A to SARS, it can be said that the history of human society is a history of the struggle against infectious diseases. Now, once again human being is facing a serious pandemic called coronavirus disease 2019 (COVID-19) which is caused by the SARS-CoV-2 coronavirus. Its first confirmed case appeared in December 2019 in Wuhan, the capital city of Hubei Province, China [1, 2]. The disease has since spread globally, leading to an ongoing epidemic that has become one of the deadliest epidemics in human history. As of October 11, 2021, more than 237 million confirmed cases and more than 4,851,000 deaths have been reported worldwide [3], and it is still spreading with several variants emerging, including the most recent variant named Omicron [4].

The occurrence and development process of these infectious diseases is influenced by multiple factors such as natural environment, population structure and epidemic prevention interventions, and has always been a hot spot and difficult research area in the field of epidemiology. Mathematical and statistical models have long been used to study the transmission of viruses because of their quantitative and scientific merits, as well as their precise and reliable qualities. Among them, the susceptible-infected-recovered (SIR) model, known as SIR model has been widely used in modeling and analysis of various infectious diseases [5]. It classifies population subject to a infectious disease into three major types with their dynamics characterised by time-dependent differential equations.

Although SIR model has achieved some success, it has many defects, one of which is that the fact that many infections diseases have a certain incubation period is not taken into account. During the incubation period, exposed individuals are able to infect others without having any symptoms. Thus, scientists have further developed the SIRS model, the SEIR model and other more realistic infectious disease models based on the SIR model [5]. In this paper, we focus on the susceptible-exposed-infected-recovered (SEIR) model, which is based on the SIR model with an ”E” (Exposed) phase to better describe COVID-19.

In order to suppress the quick spread of COVID-19, since last March countries around the world have declared partial lockdown or complete lockdown [6, 7, 8, 9]. In Canada, different provincial governments have also taken this step to control COVID-19 [10, 11, 12], and with these important measures, we did witness the decay of daily new cases although new variants of virus are still hitting hard across Canada and globally, creating one wave after another [3, 4, 13].

In the second chapter of this thesis, I focus on explaining the data collected from Nova Scotia government from March 15, 2020 to July 31, 2021. Also, the SEIR model is introduced, and all the relevant parameters in this model are explained. In the third chapter, the collected data are fitted by the SEIR model using the method of nonlinear least squares, yielding best fitted parameters to determine the basic reproduction number R_0 for COVID-19 in Nova Scotia during this period. The comparisons between the first wave and the third wave in Nova Scotia are also discussed in the third chapter. The chapter 4 contains the conclusion of our study and the future work to be done.

Chapter 2

Material And Methods

2.1 Data

The data used in this thesis are directly taken from statistics published by the Government of Nova Scotia, Canada [14]. We collected data about daily active cases from March 15, 2020 to July 31, 2021 in Nova Scotia province. Also, a graph associated with these data points is presented in Fig. 2.1. From this graph, we can tell that there are three waves that occurred during this period but because the second wave was very insignificant, we only focus on the first and third waves. We fit these two waves with the SEIR model which is introduced in the next section.

2.2 Method

In this section, I introduce and explain two main methods that are used in this research project. The first one is the SEIR model, and the second one is nonlinear least square method for fitting the data with the model.

2.2.1 The SEIR Model

The SEIR model is a compartmental model based on SIR model which was first proposed in 1927 by Kermack and McKendrick [15]. In SEIR model, population is classified into four compartments: susceptible, exposed, infected, and recovered. A susceptible person is one that is not infected with the disease and is capable of becoming infected. An exposed person is one that is contracted with the disease and has not yet exhibited any symptom but is able to infect susceptible individuals (i.e. in latent period). An infected person is one that is infected with the disease, and has shown obvious symptoms. A recovered person is not infected with the disease and cannot become infected in the future. There are some basic assumptions we need to make in order to write down differential equations that describe the dynamics of

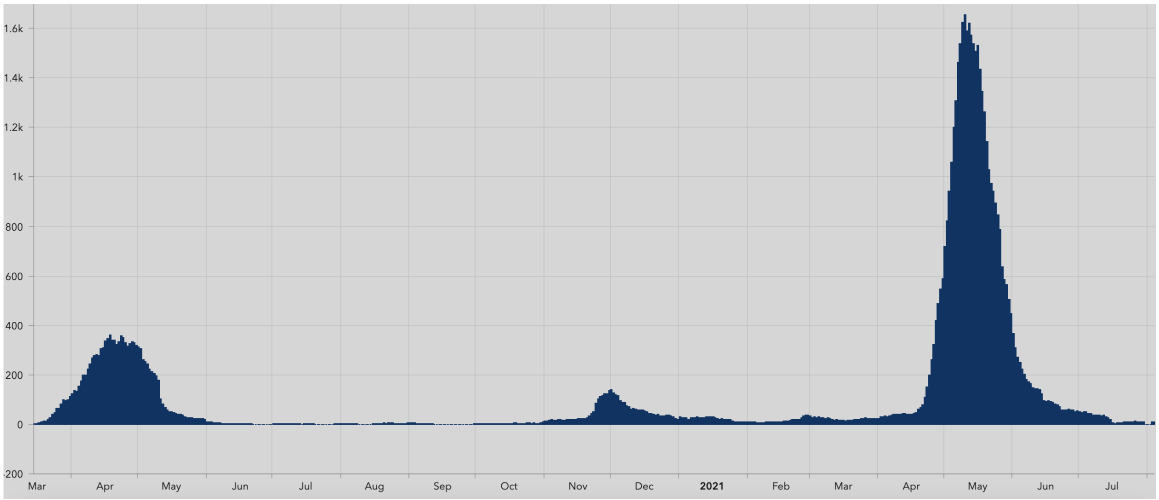

Figure 2.1: The daily active cases in NS from March 15, 2020 to July 31, 2021. In total, there are 503 days that contribute to this graph. They are directly taken from the Nova Scotia COVID-19 Dashboard [14]

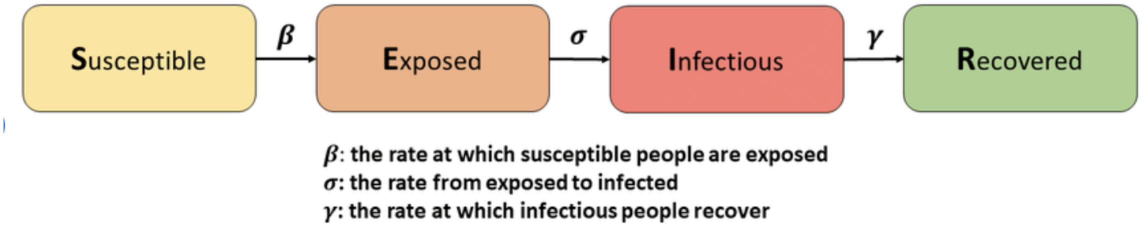
Susceptible
β
Exposed
σ
Infectious
γ
Recovered
β: the rate at which susceptible people are exposed
σ: the rate from exposed to infected
γ: the rate at which infectious people recover

Figure 2.2: : A compartmental diagram of the SEIR model. Parameters β and γ represent the rate of infection and recovery, and σ^{-1} is the average latent period.

these four quantities. These assumptions are listed as follows:

- the total population is fixed, and there is no deaths and new births during the period we study.

- a recovered individual cannot become infected again nor can it become susceptible.

- the individuals in population can be treated identically to one another at least based on their behaviours.

- there is only one-way flow from susceptible to recovered, as shown in Fig. 2.2.

With above-mentioned assumptions, we are able to write down a set of differential equations that capture the dynamics of four quantities: $S(t)$, $E(t)$, $I(t)$, and $R(t)$ which stand for the number of susceptible, exposed, infected, and recovered individuals in the population at given time t, respectively. Mathematically, they satisfy the following equations:

$$\frac{dS}{dt} = \frac{-\beta SI}{N}, \tag{2.1}$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E, \tag{2.2}$$

$$\frac{dI}{dt} = \sigma E - \gamma I, \tag{2.3}$$

$$\frac{dR}{dt} = \gamma I, \tag{2.4}$$

where non-negative parameters β and γ represent the rate of infection and recovery, and σ^{-1} is the average latent period. We should note that $N = S + E + I + R$ is a constant as the total population is assumed to be stable.

There is another important parameter worth mentioning. It is called the basic reproduction number, defined as $R_0 = \beta/\gamma$. This was first used in SIR model by Richard MacDonald in 1952 [16] to quantify the average number of contacts by an infectious individual with others before the individual recovers but it is also the same in SEIR model. Hence, the basic reproduction number R_0 can be used to estimate the level of transmission of a disease in SEIR model. If $R_0 > 1$, it means this disease is likely to become a pandemic in the population, but not if $R_0 < 1$. In general, bigger R_0 is, harder it is to control this disease.

2.2.2 Nonlinear Least Square Method

We use the SEIR model to generate a function that describes the number of the infected plus the number of the recovered for a given period. This functions contains four independent parameters: the number of initial exposed hosts E_0, β, γ, and σ. Our goal is to find the best fit for the data from Fig. 2.1 by varying these four parameters. In order to do so, we adopt the least square method. The function generated from the SEIR model is a vector with each entry corresponding to a certain time point, and then we perform the subtraction between this vector and the data vector, which gives us a new vector. After transposing this new vector and multiplying it with itself, we obtain a scalar function which is the summation of all terms with each corresponding to the square of the difference between the model-generated value and data.

We need to minimize this nonlinear scalar function using available packages in R. We choose to use ”optim” function to perform this task, which allows us to minimize the objective function while putting boundaries on the parameters.

Chapter 3

Analysis

In this chapter, I present the results for the first wave and the third wave of COVID-19 in NS, and compare these two waves based on the calculated basic reproduction number R_0 and discuss some of important facts as well.

3.1 The first wave in NS

For the first wave fitting, we use data points from March 15 to June 16, 2020 based on Fig 2.1. Since the initial number of exposed hosts is unknown to us, we set a reasonable range for E_0 in $(0, 50)$, and then we optimize the outcomes using "optim" function in R for each fixed value of E_0, which gives us a set of values for β, γ, and σ. After looping through all values of E_0, we obtain the optimal values, which are $E_0 = 49$, $\beta = 6.673655$, $\gamma = 6.394006$, and $\sigma = 4.217392$. Then, plugging these values into the SEIR model, we plot a graph showing the infected curve based on real data and a curve based on the SEIR model, which is presented in Fig. 3.1 where the horizontal axis ranging from 1 to 94 corresponds to the dates from March 15 to June 16, 2020, and the vertical axis indicates the number of active cases.

There are two important messages worth noticing: first, the first wave is almost captured by the SEIR model as the curves overlap well with each other. Second, the estimated average basic reproduction number $R_0 = \beta/\gamma = 1.0437$ is a decent value as it is only slightly greater than 1, quite in accordance with the actual situation in NS. However, the found latent period $\sigma^{-1} = 0.237$ is quite off the typical range of latent period of COVID-19, which lies between 1 and 14 days [17, 18, 19]. This could be due to the errors that come from the reported cases and the actual infected cases, and here the parameters β, γ, and σ are all time-independent but it is likely that they are time-dependent [20, 21], varying throughout the spread, which fails to be captured by the current model we use.

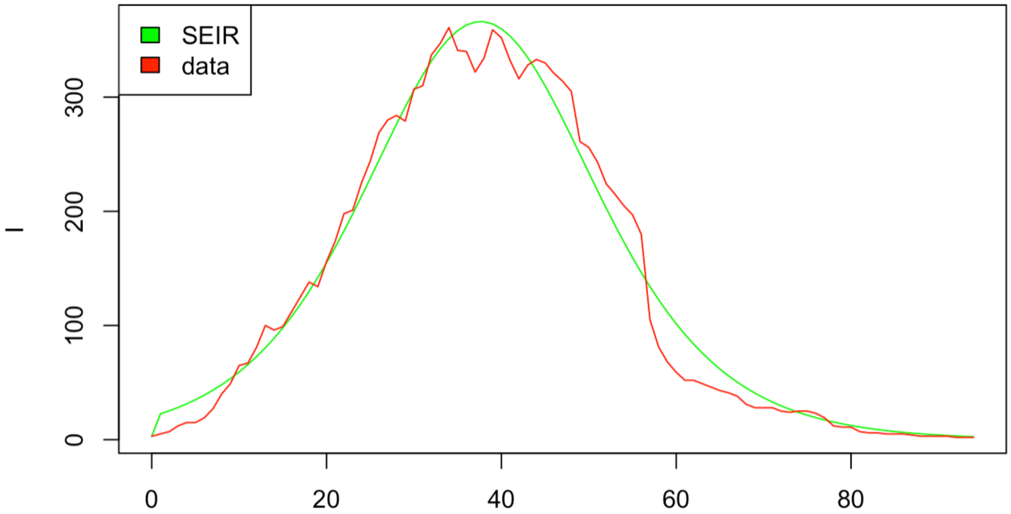

Figure 3.1: The optimized fitting of the first wave using the SEIR model, where the parameters E_0, β, γ, and σ are optimized to produce the best fit. They take the values $E_0 = 49$, $\beta = 6.673655$, $\gamma = 6.394006$, and $\sigma = 4.217392$. The estimated basic reproduction number $R_0 = \beta/\gamma = 1.0437$.

3.2 The third wave in NS

For the third wave fitting, we use data points from April 16, 2021 to July 25, 2021 based on Fig 2.1. Similarly, as we are unaware of the initial number of exposed hosts, we set a reasonable range for E_0 in $(0, 200)$, and again we optimize the outcomes using "optim" function in R for each fixed value of E_0. After looping through all values of E_0, we obtain the optimal values, which are $E_0 = 193$, $\beta = 4.709338$, $\gamma = 4.273817$, and $\sigma = 2.381285$. Then, plugging these values into the SEIR model, we plot a graph showing the infected curve based on real data and a curve based on the SEIR model, which is presented in Fig. 3.2 where the horizontal axis ranging from 397 to 497 corresponds to the dates from April 16 2021 to July 25, 2021, and the vertical axis indicates the number of active cases.

Also, we find that the third wave is well captured by the SEIR model as the curves overlap well with each other. The estimated basic reproduction number $R_0 = \beta/\gamma = 1.1019$ is larger than that of the first wave, which is quite consistent with the actual situation in NS. The third wave was much more severe than the first one as much more people got infected and it reached its peak at faster rate than the first wave

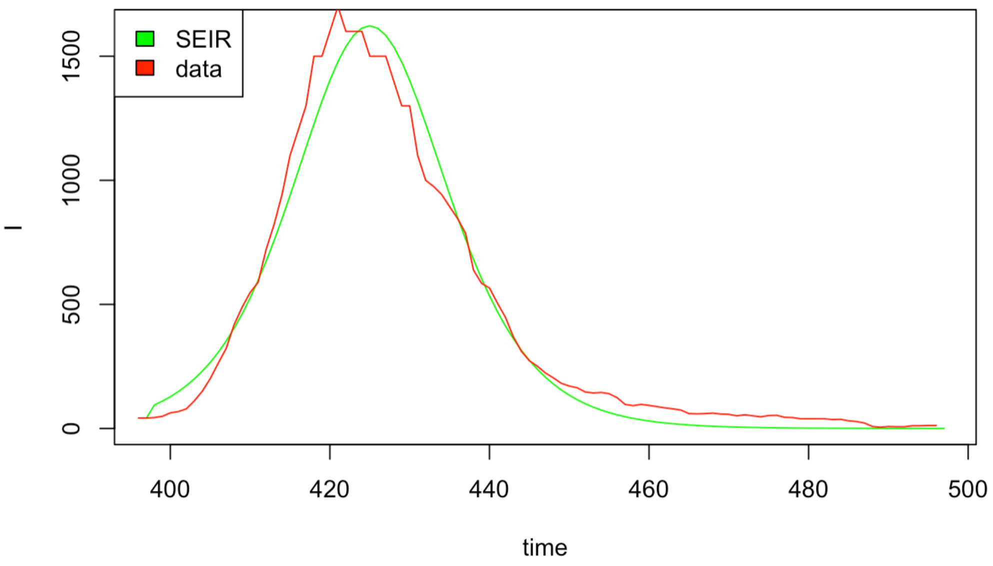

Figure 3.2: The optimized fitting of the third wave using the SEIR model, where the parameters E_0, β, γ, and σ are optimized to produce the best fit. They take the values $E_0 = 193$, $\beta = 4.709338$, $\gamma = 4.273817$, and $\sigma = 2.381285$. The estimated basic reproduction number $R_0 = \beta/\gamma = 1.1019$.

if we compare Fig.3.1 and Fig.3.2. Moreover, the latent period for the third wave is computed to be 0.4199, which is still off the typical range of latent period of COVID-19 but better than that of the first wave. This could be due to the fact that during the third wave, much more people went for testing and the sample size is much bigger than the that of the first wave. However, there still exists some errors, which could be that parameters β, γ, and σ may be time-dependent [20, 21], and the death is not included in calculations [22, 23]. In addition, for both waves the estimated basic reproduction number R_0 is an average value as it varies during the transmission, and it is also much affected by the government regulations such as lockdown and shutdown of public places, which reduce the transmission rate [21].

Chapter 4

Conclusion and Outlook

In this work, we used the well-known SEIR model to fit the collected COVID-19 data from March 15, 2021 to July 31, 2021. We used a nonlinear least square method implemented in R language to find the optimal values for β, γ, σ, and E_0, which yields the infected host curve that well matches the corresponding curve based on the data. Moreover, the calculated basic reproduction number R_0 for the first wave is smaller than that of the third wave, which is in agreement with the actual situation in Nova Scotia. The estimated latent period is off from the typical range for both waves, which could be further improved by making parameters in the SEIR model time-dependent [20, 21] or considering other compartmental models such as the SIRD model [22, 23].

One of the interesting work left for the future could be to apply the same model to the data from other provinces such as Alberta or Ontario where the COVID situation was far severe. We can estimate the basic reproduction number and the late period for these provinces and compare the results to what we obtained here for NS. In addition, making comparisons between the different compartmental models for the COVID in NS is also worth pursuing in near future. As mentioned before, having time-dependent parameters in various models is likely to better characterize the actual situation for COVID, thus making it another interesting direction to investigate.

Bibliography

[1] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The lancet*, 395(10223):507–513, 2020.

[2] Juan Yang, Xinhua Chen, Xiaowei Deng, Zhiyuan Chen, Hui Gong, Han Yan, Qianhui Wu, Huilin Shi, Shengjie Lai, Marco Ajelli, et al. Disease burden and clinical severity of the first pandemic wave of covid-19 in wuhan, china. *Nature communications*, 11(1):1–10, 2020.

[3] Covid-19 dashboard by the center for systems science and engineering at johns hopkins university. `https://gisanddata.maps.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6`.

[4] Salim S Abdool Karim and Quarraisha Abdool Karim. Omicron sars-cov-2 variant: a new chapter in the covid-19 pandemic. *The Lancet*, 2021.

[5] David JD Earn. A light introduction to modelling recurrent epidemics. In *Mathematical epidemiology*, pages 3–17. Springer, 2008.

[6] Alasdair Sandford. Coronavirus: Half of humanity now on lockdown as 90 countries call for confinement. `https://www.euronews.com/2020/04/02/coronavirus-in-europe-spain-s-death-toll-hits-10-000-after-record-950/-new-deaths-in-24-hou`.

[7] Michael Levenson. Scale of china's wuhan shutdown is believed to be without precedent. `https://www.nytimes.com/2020/01/22/world/asia/coronavirus-quarantines-history.html`.

[8] Nicola Perra. Non-pharmaceutical interventions during the covid-19 pandemic: A review. *Physics Reports*, 2021.

[9] The Lancet. India under covid-19 lockdown. *Lancet (London, England)*, 395(10233):1315, 2020.

[10] Emily Mertz. Alberta closes some non-essential business, prevents evictions as 542 covid-19 cases confirmed. `https://globalnews.ca/news/6742251/alberta-health-coronavirus-covid-19-march-27`.

[11] Office of the Premier. Ontario announces provincewide shutdown to stop spread of covid-19 and save lives. `https://news.ontario.ca/en/release/59790/ontario-announces-provincewide-shutdown-to-stop-spread-of-covid-19-and-save-lives`.

[12] Province's bars to close, restaurants limited to take-out, delivery as of thursday; gatherings limited to 50 or fewer effective immediately. `https://novascotia.ca/news/release/?id=20200317005`.

[13] S Alexandar, M Ravisankar, R Senthil Kumar, and Kannan Jakkan. A comprehensive review on covid-19 delta variant. *Int J Pharmacol Clin Res*, 5:83–85, 2021.

[14] Nova scotia covid-19 dashboard. `https://experience.arcgis.com/experience/204d6ed723244dfbb763ca3f913c5cad`.

[15] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[16] George Macdonald. The analysis of equilibrium in malaria. *Tropical Disease Bulletin*, 49(9):813–829, 1952.

[17] Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9):577–582, 2020.

[18] Conor McAloon, Áine Collins, Kevin Hunt, Ann Barber, Andrew W Byrne, Francis Butler, Miriam Casey, John Griffin, Elizabeth Lane, David McEvoy, et al. Incubation period of covid-19: a rapid systematic review and meta-analysis of observational research. *BMJ open*, 10(8):e039652, 2020.

[19] Jing Qin, Chong You, Qiushi Lin, Taojun Hu, Shicheng Yu, and Xiao-Hua Zhou. Estimation of incubation period distribution of covid-19 using disease onset forward time: a novel cross-sectional and forward follow-up study. *Science advances*, 6(33):eabc1202, 2020.

[20] Shaobo He, Yuexi Peng, and Kehui Sun. Seir modeling of the covid-19 and its dynamics. *Nonlinear dynamics*, 101(3):1667–1680, 2020.

[21] James P Gleeson, Thomas Brendan Murphy, Joseph D O'Brien, Nial Friel, Norma Bargary, and David JP O'Sullivan. Calibrating covid-19 susceptible-exposed-infected-removed models with time-varying effectivecontact rates. *Philosophical Transactions of the Royal Society A*, 380(2214):20210120, 2021.

[22] Jesús Fernández-Villaverde and Charles I Jones. Estimating and simulating a sird model of covid-19 for many countries, states, and cities. Technical report, National Bureau of Economic Research, 2020.

[23] Saptarshi Chatterjee, Apurba Sarkar, Swarnajit Chatterjee, Mintu Karmakar, and Raja Paul. Studying the progress of covid-19 outbreak in india using sird model. *Indian Journal of Physics*, 95(9):1941–1957, 2021.