5-4-2022

Two Methods of Forecasting the Spread of Disease Using R_0 Coefficient and Gradient-Descent

Dinh Song An Nguyen
College of Saint Benedict/Saint John's University, anguyen002@csbsju.edu

Two Methods of Forecasting the Spread of Disease Using R_0 Coefficient and Gradient-Descent

Dinh Song An Nguyen

The College of Saint Benedict and Saint John's University

4th May 2022

Two Methods of Forecasting the Spread of Disease Using R_0
Coefficient and Gradient-Descent

Approved by:

Dr. Heather Amthauer
Thesis Advisor
Professor of Computer Science

Dr. Peter Ohmann
Faculty Reader
Professor of Computer Science

Dr. Anne Sinko
Faculty Reader
Professor of Mathematics
Chair, Department of Mathematics

Dr. Imad Rahal
Professor of Computer Science
Chair, Department of Computer Science

Lindsey Gunnerson Gutsch, M.S.
Program Director, Director of Undergraduate Research

Contents

List of Figures

List of Algorithms

Abstract

In 2020, COVID-19 became a global pandemic and has been negatively impacting the world ever since. One of the ways to help control the spread of disease is to forecast its growth. Forecasting the spread of disease, though a daunting task, is necessary when a pandemic happens as it could help create social policies that could potentially mitigate the effects of the disease. This research aims to provide two disease prediction methods: the R_0-based method and the gradient-descent method. The R_0-based method produces short-term predictions by simulating the mobility of residents and utilizing the R_0 coefficient. The gradient-descent method maps the linear regression model onto non-linear model in order to fit the exponential growth of the disease. Experimental results show that the R_0-based method is accurate at forecasting during pandemic outbreaks. The gradient-descent method is able to study the spread of epidemics on a city-to-city scale through transport network, but with less accuracy than the R_0-based technique. [1]

[1] I offer my sincerest gratitude to my advisor, Dr. Heather Amthauer, and my committee member, Dr. Peter Ohmann and Dr. Ann Sinko, for their mentorship as well as their invaluable ideas and feedbacks. Special thanks to Lindsey Gutsch and the Undergraduate Research Department for funding me through the Summer Thesis Fellowship.

1 Introduction

Despite numerous efforts to mitigate its effect, infectious disease remains one of the most serious causes of death in human history. One of the fastest ways for infectious disease to spread around the world is through traveling. Research shows strong correlations between traveling, both internationally and domestically, and the spread of disease [1]–[5]. Infectious diseases such as H1N1 swine influenza and severe acute respiratory symptoms (SARS) were shown to infect many people around the world within only weeks. For the past two years, the coronavirus disease, otherwise known as COVID-19 or SARS-CoV-2, started in Wuhan, China, has spread to 216 countries, infecting over 38 million people and killing over 1 million [2]. Aside from the health effects, a pandemic could also take a toll on the economy. The 1994 bubonic plague in Inida recorded 52 deaths in total but resulted in over 1 billion USD economic loss [4]. As a result of the death toll and economic upheaval, there has been growing in epidemiological studies and modeling. Analyzing and predicting the impact of infectious diseases can assist in contact tracing, creating intervention policies, and developing prevention strategies. This research focuses on the coronavirus disease specifically.

COVID-19 is an infectious disease caused by the coronavirus named SARS-CoV-2, first appeared in China at the end of 2019 and soon became a global pandemic. Predicting the spread of disease is important in developing prevention strategies to counter the disease. However, like weather prediction, predicting the spread of disease is an incredibly difficult task due to the non-linear dependent nature of the pathogens as well as multifarious factors that affect the spread of disease. Spread of disease is the result of interactions between pathogens, humans, and surrounding environment. In fact, pathogens exist and evolve in ideal environments like any other organism. If the human body provides favorable conditions, pathogens will be able to reproduce exponentially. Though complicated, it is necessary to develop mathematical models to understand and produce prediction regarding the disease in order to develop long-term or short term prevention strategies and policies for effective control.

1.1 SEIR Model

The Susceptible-Exposed-Infected-Recovered model, otherwise known as SEIR model, has been used to model after the spread of disease in past research [6], [7] and recently used to predict the spread of COVID-19 in community [8].

The SEIR model is based on the assumption that population of the community, denoted as N, never changes throughout the entire pandemic period and consists of the following groups:

- S: people who are susceptible to COVID-19;
- E: people who are exposed to COVID-19;
- I: people who are infected by COVID-19 and can spread the disease;
- R: people who have recovered from COVID-19 or died because of it.

Since N is a constant, without loss of generality, we can write

$$S + E + I + R = N. \tag{1}$$

Suppose:

- Ratio of birth and death is the same and is μ.
- Average incubation time is α^{-1}.
- Average infecting period is γ^{-1}.
- An infected individual who has recovered can't get infected anymore.
- Interaction coefficient β is a time related function.

From this, we have the following system of differential equations:

$$\frac{dS}{dt} = \mu - \beta(t)SI - \mu S \tag{2a}$$

$$\frac{dE}{dt} = \beta(t)SI - (\mu + \alpha)E \tag{2b}$$

$$\frac{dI}{dt} = \alpha E - (\mu + \gamma)I. \tag{2c}$$

R is calculated using equation 1.

Using the Euler method, we will approximate $S(t), E(t), I(t)$, and $R(t)$ at time $t > 0$ when we know the values in the beginning.

1.2 R_0 Value

R_0, otherwise known as R-naught, is the reproductive value of an infectious disease. It can also be used to approximate how many people can get infected from an infected individual. This value is important because law-makers and politicians use it to determine timeline for the pandemic as well as policies.

The reproductive value R_0 is the average number of people getting infected from an infected person. High R_0 means that there will be a lot of people getting infected during the pandemic period. Conversely, low R_0 means that less people will get infected over time. However, infection will continue if there is no vaccine to counter the disease.

The concept of R_0 was first introduced in anthropology [9]. In epidemiology, R_0 is used to count the number of infected individuals. For experts, R_0 can be a valuable asset. However, the process of approximating, calculating, explaining as well as applying R_0 is a complicated task. The simplicity of R_0 and its corresponding interpretation in relation to infectivity hide the complexity of this value. During the pandemic period, R_0 is used in complex mathematical models and is developed using different set of assumptions [10].

1.3 Research Motivation

SARS-CoV-2 has appeared in Vietnam since late 2019, but only became a deadly pandemic in Viet Nam in May 2021. Information about SARS-CoV-2 positive cases for each district in Ho Chi Minh city is updated everyday on the website of the Center for Application of Geographical Information Systems. Additionally, Ho Chi Minh city Department of Science and Technology also provides information regarding COVID-19 in each wards of Thu Duc, Ho Chi Minh City [11]. Moreover, COVID-19 infection has been getting worse in California.

This research aims to provide short-term prediction for SARS-CoV-2 positive cases in Thu Duc city, an industrial-concentrated area in Viet Nam, and in counties of Southern California.

1.4 Data availability

All datasets regarding Thu Duc's population, district population, and COVID-19 infection in 2021 as well as California's COVID-19 positive cases used in this research is available on this GitHub page or this following url: `https://github.com/ndsongan/undergraduate-thesis-2022`.

2 Short-term prediction of SARS-CoV-2 positive cases based on reproductive value R_0

In this research, we will build appropriate mathematical models that allow short-term predictions of infected cases caused by SARS-CoV-2 using real data collected daily. The inputs for the models are the data regarding positive cases in each ward of Thu Duc district. The outputs of the model are the number of positive cases for each ward of Thu Duc in each day. Models can also be adapted and applied to predictions in other geographical regions and cities if provided proper data.

2.1 Methodology

Data regarding the COVID-19 pandemic that is uploaded daily on the website of the Center for Application of Geographical Information Systems as well as well as Thu Duc's official website is the main datasource that we will use to construct mathematical models to predict COVID-19 status in Thu Duc. However, because the websites only provided daily update of COVID-19 positive cases without any other information, except for Thu Duc's official website which has data regarding people who have recovered from SARS-CoV-2 but no data regarding people who have died from it, the SEIR model is inapplicable for Thu Duc.

In order to predict COVID-19 cases in the short term, we will use the reproductive value R_0. The difference between this method compared to other prediction methods is that R_0 is not a constant, but rather it evolves throughout time based on each ward in Thu Duc. Additionally, in order to determine whether the time period is fit for short-term predictions, a Markov chain model was built based on the daily collected data.

2.2 Appropriate R_0 value

In the past 2 years, ever since the COVID-19 pandemic first appeared in Wuhan, China, researchers suggested that R_0 ranges between 2.2 to 2.7 depending on the region [12]. This means that a person who has COVID-19 will be able to transmit the disease to an average of 2.2 to 2.7 other people.

Approximating the R_0 value is a complicated task. In this research, we supposed that R_0 has different values based on each region, and changes over time during the pandemic period. The R_0 coefficient depends on many factors. We considered the following factors for our research:

- The first factor is population density, denoted as D_d, is the density of region d. The higher the density, the more people living within the 1 km^2 area, which makes the risk of spreading the disease higher.
- The second factor is population, denoted as P_d, is the population of area "d". We supposed that under the condition of social distancing, the number of residents who moves from area d to area d' is $\frac{1}{1000}$ of the population of the area. Ω_d is the set of these moving residents. Since we do not need to specify which resident is which, ω_d is just an arbitrary representation in order to perform equations calculating R_0 and F_0.
- The third factor is the number of people infected by COVID-19, denoted as F_0, which was provided since social distancing policies began.

From this, R_0 of each region d is a function over time, with day as the unit, and depends on the original value of R_0, $R(t = 0)$ or $R(0)$. This is a recursive function that will be discussed in details in the next section.

2.3 Markov Chain Model for Predicting the Stopping Point

Definition (Discrete-time Markov chain): Let S be a finite set. Random sequence $\{X_0, X_1, \dots\}$ takes values from S is called a Markov chain if

$$\begin{aligned} Pr(X_{n+1} = j|X_0 = x_0, X_1 = x_1, \dots, X_n = i) &= Pr(X_{n+1} = j|X_n = i) \\ &= Pr(X_1 = j|X_0 = i) = p_{ij} \\ \text{for all } x_0, x_1, \dots, i, j \in S \text{ and } n \in N. \end{aligned} \tag{3}$$

S is index set or state space of the chain. When $X_n = i$, we often say chain is at state i at time n. Markov chain is defined like above is called discrete-time Markov chain, homogeneous and finite state space.

Definition (Transitional matrix): Let Markov chain $\{X_0, X_1, \dots\}$ have finite state S, for $|S| = k$. Matrix $P \in \mathbb{R}^{k\times k}$ is the transitional matrix of chain if

$$P_{ij} = p_{ij} = Pr(X_1 = j|X_0 = i). \tag{4}$$

The sum of of all elements in each column of matrix P is 1:

$$\Sigma_{i=1}^{k} p_{ij}, \forall 1 \le i \le k. \tag{5}$$

Definition (Distribution vector): Let random variable X from set of finite states $S = \{1, 2, \dots, k\}$. Vector $\pi \in \mathbb{R}^k$, with

$$\pi_i = Pr(X = i), \forall 1 \leq i \leq k, \tag{6}$$

is the distribution of X.

Let there be a Markov chain $\{X_0, X_1, \dots\}$. Distribution of X_n is called distribution of Markov chain at time n, and distribution of X_0 is called initial distribution of Markov chain.

In this section, we only provide the main concepts of Markov chain without proving them. Details about proofs regarding Markov chain can be reviewed in linear algebra or statistics textbooks [13].

Proposition Let there be Markov chain $\{X_0, X_1, \dots\}$ with transitional matrix P and π_t is the distribution of chain at time n. Distribution of X_0 is also called initial distribution of chain.

$$\pi_{t+n} = P^n \pi_t, \forall t, n \in \mathbb{N}. \tag{7}$$

In other words, we have

$$\pi_n = P^n \pi_0, \forall n \in \mathbb{N}. \tag{8}$$

Definition (Limiting distribution): Let there be Markov chain $\{X_0, X_1, \dots\}$ with transitional matrix P and distribution π with the property

$$\lim_{x \to \infty} P^n_{ij} = \pi_j, \forall 1 \leq i, j \leq k. \tag{9}$$

is called the distribution limit of a chain.

Proposition Let π be the distribution limit of chain $\{X_0, X_1, \dots\}$. The following statements are true:

(i) For all initial distribution and states j, $\lim_{n \to \infty} Pr(X_n = j) = \pi_j$.

(ii) For all initial distribution α, $\lim_{n \to \infty} P^n \alpha = \pi$.

(iii) $\lim_{x \to \infty} P(X_n = j) = \pi_j$.

Note:

- A chain may not have limiting distribution, but if it does, it only has one limiting distribution.

- When a chain has limiting distribution π, we can think of π_j as the long-term probability of chain at state j, meaning the probability of chain at state j after a long time does not depend on initial distribution.

Definition (Stationary distribution): Let there be Markov chain $\{X_0, X_1, \dots\}$ with transitional matrix P. If distribution π has the property

$$\pi = P\pi \tag{10}$$

then π is called the stationary distribution of chain.

Note:

- A Markov chain may or may not have one or many stationary distribution.
- When a chain falls into a stationary distribution, it will stop at that distribution, meaning if there exists an n such that $\pi_n = \pi$ is the stationary distribution, then $\pi_t = \pi, \forall t \geq n$.
- The limit distribution, if there is any, is a stationary distribution.

Definition (Regular Matrix): Matrix $P \in [0,1]^{k \times k}$ is called regular if there exists an $n \in \mathbb{Z}$ such that P^n consists of all positive element,

$$P^n_{ij} > 0, \forall i,j | 1 \leq i,j \leq k. \tag{11}$$

Proposition A chain $\{X_0, X_1, \dots\}$ has transitional matrix P. If P is regular, then chain has limiting distribution.

3 COVID-19 short-term prediction formula

3.1 Formula for predicting COVID-19 infection

F_0 is the notation for people who was tested positive for COVID-19. Given F_0 value of area d, our task is to find F_0 in area d for the next day.

In the case of Thu Duc, each ward is considered an independent unit. Therefore, we denote area d as ward d of Thu Duc. The formula to calculate F_{0_d} on day t is the following:

$$F_{0_d}(t) = F_{0_d}(t-1)(1 + \alpha + R_{0_d}(t)). \tag{12}$$

In which

- α is the interaction coefficient. This coefficient is approximated based on economic and cultural significance of the place where infection happens. With the cultural circumstance of Vietnam, usually 2 to 3 generations of family live together. We supposed this coefficient is 0.001, meaning a person on average will interact with $\frac{1}{1000}$ of residents in the area in which they live.

- $R_{0_d}(t)$ is the coefficient for risk of infection spreading in ward d. The formula for this coefficient will be established in the next section.

3.2 R_0 Formula

We will create formula for risk coefficient in area d. This coefficient is the highest at R_{0_d} of district d. We also suppose that each resident of ward d will carry the risk coefficient R_{0_d} when they leave their ward.

- If ward d's resident goes to ward d' and the risk coefficient $R_{0_{d'}} < R_{0_d}$, then this resident will have an effect on coefficient $R_{0_{d'}}$ of ward d' and α times the difference $R_{0_d} - R_{0_{d'}}$.

- Conversely, if $R_{0_{d'}} > R_{0_d}$, then the risk of resident from ward d increases and when they will affect R_{0_d} coefficient once they return. So, R_{0_d} of ward d will be updated, taking into account the average risk of infection of the residents who go to other wards and return.

α denotes the interaction coefficient, which is the probability of a resident from one ward interact with residents from other wards within the surveyed area.

We have the formula for R_0 as follow:

1. First, R_{0_d} is set for day 0, $x = 0$:

$$R_{0_d}(x=0) = R_{0_{min}} + \frac{D_d - D_{min}}{D_{max} - D_{min}}(R_{0_{max}} - R_{0_{min}}). \tag{13}$$

   In which $R_{0_{min}} = 2.2, R_{0_{max}} = 2.7$, and $D_{0_{max}}, D_{0_{min}}$ are the greatest and smallest population density of Thu Duc.

2. For $x \geq 0$, R_0 is updated with the following formula:

   2.1. A person i the risk value of $R_{d_i} = R_{0_{d'}}(x-1)$ goes from region d' to region d. If the risk value of region d, $R_{0_d}(x-1)$ is less than $R_{0_{d'}}(x-1)$, then region d will be affected by person i. Conversely, if risk value of region d is greater, then region d will not be affected at all, but person i will be affected by the risk factor from region d. We have the following formula:

$$R_{0_d}(x) = R_{0_d}(x-1)+\alpha(R_{0_{d'}}(x-1)-R_{0_d}(x-1)), \text{ if } R_{0_d}(x-1) < R_{d_i}, \tag{14}$$

and

$$R_{d_i} = R_{d_i} + \beta(R_{0_d}(x-1) - R_{d_i}), \text{ if } R_{0_d}(x-1) > R_{d_i}. \tag{15}$$

2.2. Let R_{d_i} be the cumulative risk value of person $i \in |\Omega_d|$ after they travel to nearby regions. Once $|\Omega_d|$ people return to region d, the R_0 coefficient of region d is updated with the following formula:

$$R_{0_d}(x) = R_{0_d}(x-1) + \alpha(\frac{1}{|\Omega_d|}\Sigma_{i\in\Omega_d}(R_{d_i} - R_{0_d}(x))). \tag{16}$$

3. From equations 13, 14, 15, and 16, we have the following formula for R_0:

$$R_{0_d}(x+1) = \begin{cases} R_{0_{min}} + \frac{D_d - D_{min}}{D_{max}-D_{min}} \cdot (R_{0_{max}} - R_{0_{min}}), & \text{if } x = 0. \\ R_{0_d}(x) + \alpha(R_{0_{d'}}(x) - R_{0_d}(x)), & \text{if } R_{0_d}(x) < R_{0_{d'}}(x). \\ R_{0_d}(x) + \alpha(\frac{1}{|\Omega_d|}\Sigma_{i\in\Omega_d}(R_{d_i}(x) - R_{0_d}(x)), & \text{otherwise.} \end{cases} \tag{17}$$

We define α as the interaction coefficient of the person and the districts that they visit. We set $\alpha = 0.001$. We also define β as the interaction coefficient of the district and the people that visit that district. We set $\beta = 0.01$. The α and β coefficients are established through experimentation.

4. Lastly, the number of infected people in each region d is updated using the following equation:

$$F_{0_d}(x) = F_{0_d}(x-1)(1 + \alpha R_{0_d}(x)). \tag{18}$$

3.3 COVID-19 Predicting Algorithm Using Random Walk on Graph

In order to predict the total COVID-19 cases using the R_0 coefficient, we will apply some ideas from the random walks on graphs method [14]. In this algorithm, we will represent each ward of Thu Duc as a vertex in a completed, undirected, and unweighted graph $G = (V, E)$ in which, each vertex $v \in V$ is a ward of Thu Duc, and $(v, v') \in E$ is a bird-flight path from district v to district v', and vice versa. We represented the city as a complete graph because even though districts can be adjacent, residents can use vehicle to travel without making any stop, meaning they do not interact with the intermediate districts as much. The algorithm predicting total COVID-19 cases is showed in algorithm 1.

Algorithm 1 Prediction Algorithm (Part 1)

```
Input
    $G = (V, E)$                       graph represented area of infection
    $D = \{D(v), v \in V\}$            list of population density of ward $v \in V$
    $P = \{P(v), v \in V\}$            list of population of ward $v \in V$
    $days$                             number of days that needs to be predicted
    $R_{0_{min}}$                      lower bound of $R_0$ coefficient
    $R_{0_{max}}$                      upper bound of $R_0$ coefficient
Output
    $F_0 = \{F_{0_v}(t)\}$             matrix of number of COVID-19 positive cases
    in ward $v \in V$ from day $1 \leq t \leq days$.
$maxx \leftarrow Max(D)$
$minn \leftarrow Min(D)$
Initialize $R_0(v), \forall v \in V$ using equation 13.
$X[v] \leftarrow P[v]/100, \forall v \in V$        ▷ suppose only 1% of population in ward
$v \in V$ will travel out of $v$.
$t \leftarrow 1$
```

Algorithm 2 Prediction Algorithm (Part 2)

```
for $p \leftarrow 0$ to $length(X(v)$ do
    for $v \leftarrow 0$ to $length(V)$ do        ▷ Suppose $p$ is the 0.1% of the population that is infected.
        if $random < 1/1000$ then
            $R(p) = 1$
        else
            $R(p) = R_0(v)$
        end if
        $visit\_nodes = random(1, 5)$.        ▷ nodes to be visited
        $path = \{random(v_i) \in V, 1 \leq i \leq visit\_nodes$
        Update $R_0[v'], v' \in path$ using equation 14.
        Update $R_0(p)$ using equation 15.
    end for
    Update $R_0(v)$ using equation 16.
    Recalculate $F_{0_v}(t)$ using equation 18.
    $t \leftarrow t + 1$
end for
```

3.4 Regular Markov Chain and Forecasting Period

Formula predicting COVID-19 positive patients, F_{0_d}, is an increasing function because $R_{0_d} \geq 0$. Therefore, a regular Markov chain is used to calculate meaningful forecasting period.

Base on figure 1, we define the 4 states $S = \{S_1, S_2, S_3, S_4\}$ for:

- S_1: Number of cases decreases. It is considered decreasing when the percentage of new cases increases less than 2%, $S_1 \leq 0.02\Sigma_d F_{0_d}$;
- S_2: Number of cases is stable, $0.02 < S_2 \leq 0.08\Sigma_d F_{0_d}$;
- S_3: Number of cases increases, $0.08 < S_3 \leq 0.11\Sigma_d F_{0_d}$;
- S_4: Number of cases strongly increases, $0.11\Sigma_d F_{0_d} < S_4$.

During the period from July 24^{th} to August 16^{th}, we created the transitional matrix as showed in table 1.

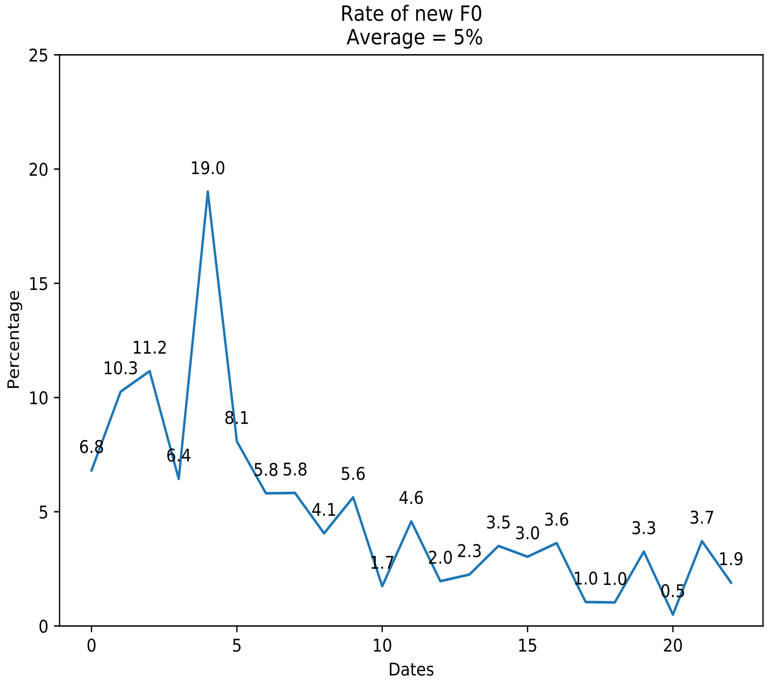


Figure 1: Rising percentage of new F_0 cases from July 23^{th} to August 16^{th}. The percent of new COVID-19 positive cases rise 5% on average.

P_0	S_1	S_2	S_3	S_4
S_1	0.16667	0.83	0	0
S_2	0.42857	0.42857	0.07143	0.07143
S_3	0	0.5	0	0.5
S_4	0	0.5	0.5	0

Table 1: Transitional Matrix $S = \{S_1, S_2, S_3, S_4\}$ created by using infection data from July 23^{th} to August 16^{th}.

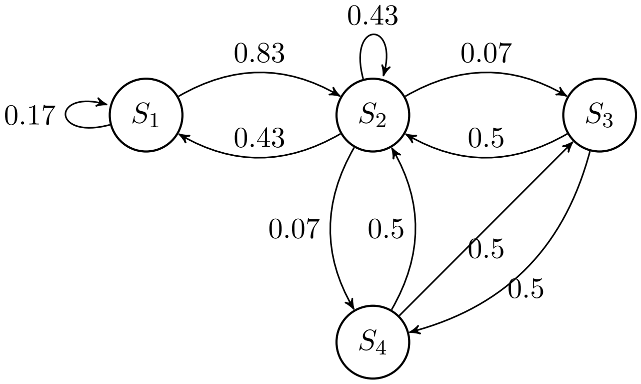


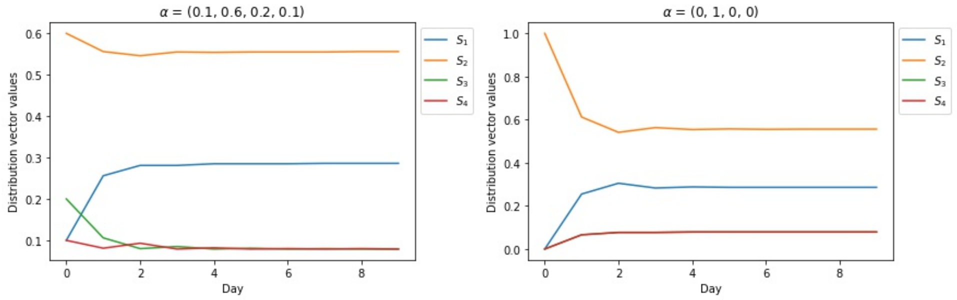
Figure 2: Another representation of the Markov chain in table 1. An arrow from state S_1 to state S_2 represents the probability of state S_1 transitioning to state S_2.

P	S_1	S_2	S_3	S_4
S_1	0.16667	0.8333	0.000001	0.000001
S_2	0.42857	0.42857	0.07143	0.07143
S_3	0.000001	0.499999	0.000001	0.499999
S_4	0.000001	0.499999	0.499999	0.000001

Table 2: Regular Matrix $S = \{S_1, S_2, S_3, S_4\}$ after one iteration.

P	S_1	S_2	S_3	S_4
S_1	0.276329	0.59149	0.064761	0.064761
S_2	0.303911	0.539402	0.076304	0.076304
S_3	0.23433	0.539402	0.050694	0.175694
S_4	0.23433	0.539402	0.175694	0.050593

Table 3: Regular Matrix $S = \{S_1, S_2, S_3, S_4\}$ after 3 iterations.



(a) $\alpha = (0.1, 0.6, 0.2, 0.1)$ (b) $\alpha = (0, 1, 0, 0)$

Figure 3: Distribution vector (a) $\alpha = (0.1, 0.6, 0.2, 0.1)$ and (b) $\alpha = (0, 1, 0, 0)$ in 10 days.

Figure 3 shows the states of the distribution vectors $\alpha = (0.1, 0.6, 0.2, 0.1)$ and $\alpha = (0, 1, 0, 0)$ in 10 days. Notice in figure 3b, there are only 3 lines represents S_1, S_2, and S_4. The line for S_3 is not missing, but rather, it locates under S_4 because S_3 increases at the same amount as S_4. As seen in both cases, the states begin to converge after 3 days of the Markov process. Therefore, in order to get the most optimal and accurate prediction, we should only predict within a 3 days period. New data must be updated after 1 to 3 days.

3.5 Example

In regards to how the R_0 based method works, consider the following scenario.

Suppose we have a complete graph of a city with five nodes representing five districts as shown in figure 4. For this example, we make the following assumptions:

- There are 2 people traveling from district 1. One person travels to district 4 and 5, and the other travels to district 2 and 3. At the end of the iteration, they return to their original district.
- Each district is assigned an R_0 value.
- $\alpha = 0.05$
- $\beta = 0.01$

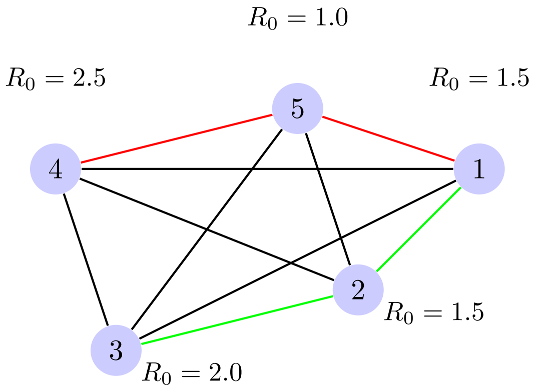


Figure 4: Two people travel from district 1 to other districts. First person's travel path is $\{1, 5, 4\}$, denoted in red. Second person's travel path is $\{1, 2, 3\}$, denoted in green. The R_0 values of each district are indicated next to the nodes.

When a person travels from district 1 to district 5, since district 1's R_0 value, denoted as R_{0_1} is greater than that of district 5's, R_{0_1} remains the same but R_{0_5} will increase at the end of the iteration. However, when that person travel to district 4, since $R_{0_1} < R_{0_4}$, R_{0_4} will remains the same but R_{0_1} will increase. The R_{0_1} is calculated using equation 17.

$$\begin{aligned} R_{0_1} > R_{0_5} : R_{0_1} &= R_{0_1} + 0 \\ R_{0_1} < R_{0_4} : R_{0_1} &= R_{0_1} + \alpha(R_{0_4} - R_{0_1}) \\ &= 1.5 + 0.05(2.5 - 1.5) = 1.55 \end{aligned}$$

Similarly, when the other person travel from district 1 to district 2 and 3, the R_{0_1} is calculated using equation 17.

$$\begin{aligned} R_{0_1} = R_{0_2} : R_{0_1} &= R_{0_1} + 0 \\ R_{0_1} < R_{0_3} : R_{0_1} &= R_{0_1} + \alpha(R_{0_3} - R_{0_1}) \\ &= 1.5 + 0.05(2.0 - 1.5) = 1.525 \end{aligned}$$

When both persons return to district 1, the total R_{0_1} is calculated using equation 16:

$$\begin{aligned} R_{0_1} &= R_{0_1} + \alpha((1.55 - 1.5) + (1.525 - 1.5)) \\ &= 1.5 + 0.05(0.05 + 0.025) = 1.50375 \end{aligned}$$

At district 5, since $R_{0_1} > R_{0_5}$ when district 1's resident visit district 5, district 5's R_0 value is computed using equation 15:

$$R_{0_5} = R_{0_5} + \beta(1.5 - 1.0) = 1.0 + 0.01(1.5 - 1.0) = 1.005$$

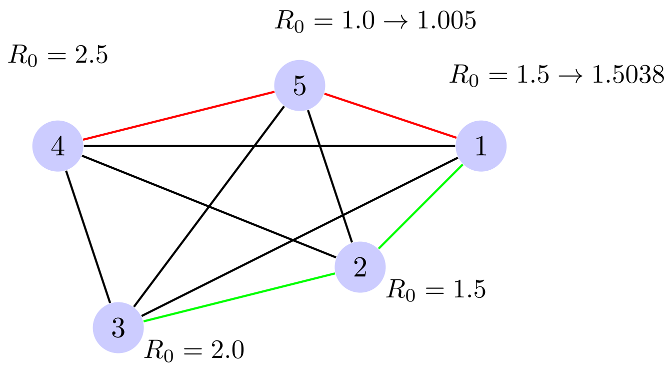


Figure 5: The updated graph after 1 iteration. The R_0 values of district 1 and district 5 are increased due to the travel and the interaction of district 1's residents.

4 Gradient-Descent Prediction

In addition to the R_0 based prediction method, we also developed another prediction method that combines

4.1 Linear Regression

Regression is one of the most popular tools to analyze epidemiology due to its accessibility and simplicity, though it can get complicated as need be. Linear regression used to explicate the relationships between explanatory and response variables in diseases, thus analyzing the risks factors that the variables might have [15], [16]. Binomial regression such as logistic regression is used as a classification method [17], [18]. Regression is also used to forecast the growth of diseases by combining multiple variables [19]. Regression methods is an universal tool that can be applied to any epidemic, depending on the goals of the study. In this research, we will focus specifically on linear regression and multiple linear regression and utilize them to forecast the growth of COVID-19.

Regression refers to the task of predicting a continuous quantity. The goal of regression is to find a general model that fits the given data. Linear regression is a stochastic model that assumes the relationship between the input variable (x) and the output variable (y) to be **linear**. When there are more than one input variable, the model is called multiple linear regression. Linear regression assumes the form of equation 19.

$$y = w_0 + w_1 x. \tag{19}$$

In equation 19, y is the dependent variable while x is the independent variable with $w_0, w_1 \in \mathbb{R}$.

Definition (Independent variable): In function $f(x)$, the independent variable, denoted as x, is the variable that does not depend on other variables. It is used as input value for a function.

Definition (Dependent variable): In function $f(x)$, the dependent variable, denoted as $y = f(x)$, is the variable that depends on other variables, in this case, the independent variables. It is called the output of a function.

For multiple linear regression, we will have n independent variables and 1 dependent variable, e.g. multiple inputs and one output. Then, our regression equation will be like equation 20.

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n. \tag{20}$$

4.2 Gradient-Descent

Gradient-descent is defined as the following:

Definition (Gradient-Descent): an algorithm to find the local minimum or local maximum of a function. It is used to minimize the cost of a loss function [20].

4.2.1 Notations

Let

- $x_1(t)$: number of infected of ward 1 on day t^{th}.
- $x_2(t)$: number of infected of ward 2 on day t^{th}.
- $\ldots$
- $x_n(t)$: number of infected of ward n on day t^{th}.

4.2.2 Prediction equations

Let there vector $x(t) = \{x_1(t), x_2(t), \dots, x_n(t)\}$ be the vector of infections in the city on day t^{th}. Suppose we have a weight matrix $W = [w_{ij}]$ with $1 \leq i, j \leq n$ and $B = [b_i]$ for $1 \leq i \leq n$. We predict infection rate for day $(t+1)^{th}$ by computing the following equations:

$$y_j = w_{1j}x_1(t) + w_{2j}x_2(t) + \cdots + w_{nj}x_n(t). \tag{21a}$$

$$x_j(t+1) = (h(y_j) + b_j)x_j(t) \tag{21b}$$

In which $h(x)$ is a tangent hyperbolic logistic equation:

$$h(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}. \tag{22a}$$

$$h'(u) = 1 - h(x)^2. \tag{22b}$$

4.2.3 Methodology

We established some learning definitions for the prediction algorithm.

- $st = x(t)$ and $st1 = x(t+1)$ be the vectors of infections at day t^{th} and $(t+1)^{th}$, respectively.
- $z = (z_1, \dots, z_n)$ is the prediction of the model.
- $e = \frac{1}{2}\Sigma_{k=1}^{n}(z_i - st1_i)^2$ is the total error of predictions.
- $\Delta = \frac{de}{dw_{ij}}$ is the derivative of the error function.

Then we calculated the weight w_{ij} using the formula:

$$w_{ij} = w_{ij} + \alpha\Delta. \tag{23}$$

From the definitions, we have the following derivative equations:

$$\Delta = \frac{de}{dw_j} = \frac{de}{dz_j}\frac{dz_j}{dy_j}\frac{dy_j}{dw_j}. \tag{24a}$$

$$\frac{de}{dz_j} = z_j - st_j. \tag{24b}$$

$$\frac{dz_j}{dy_j} = h'(y_j)st_i. \tag{24c}$$

$$\frac{dy_j}{dw_j} = st_j. \tag{24d}$$

From equations 22, we have

$$\Delta = (z_j - st_j)h'(y_j)(st_j)^2. \tag{25}$$

4.3 Linear prediction functions

Let $x(t) = \{x_1(t), x_2(t), \ldots, x_n(t)\}$ be the infection vector of n states or cities on day t^{th} for $t \geq 0$. Suppose we have matrix $W = [w_{ij}]$ with $1 \leq i, j \leq n$ and matrix $B = [b_i]$ for $1 \leq i \leq n$. Then, we can prediction infection rate for day $(t+1)^{th}$ by computing the following equations:

$$x(t+1) = x(t)W + B \tag{26a}$$

or

$$x_i(t+1) = w_{1i}x_1(t) + \cdots + w_{ni}x_n(t) + b_i \tag{26b}$$

Each element in $W = \{w_{ij}\}$ can be determined using multi-variable regression analysis with vector $x(t) = (x_1(t), \ldots, x_n(t))$ being the independent variable and vector $x_i(t+1)$, a part of vector $x(t+1)$, being the dependent variable. Therefore, we need to perform multi-variable regression n times.

Suppose for the i^{th} regression iteration, we have

$$x_i(t+1) = a_{0i} + a_{1i}x_1(t) + \cdots + a_{ni}x_n(t), \tag{27}$$

with

$$\begin{aligned} w_{1i} &= a_{1i}, a_{2i}, a_{3i}, \ldots, a_{ni} = a_{ni}, \\ b_i &= a_{0i}. \end{aligned} \tag{28}$$

Then, we have matrix $W_i = \begin{pmatrix} a_{1i} \\ \vdots \\ a_{ni} \end{pmatrix}$ is the i^{th} column of matrix W and

$$W = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}, \; B = \begin{bmatrix} a_{01} \\ \vdots \\ a_{0n}. \end{bmatrix} \tag{29}$$

4.4 Non-linear prediction functions

The functions in the previous section will perform well with linear data. However, epidemic data tends to be non-linear. We can non-linearize the predictions by using the following equations:

$$y_j = w_{1j}x_1(t) + w_{2j}x_2(t) + \cdots + w_{nj}x_n(t). \tag{30a}$$

$$x_j(t+1) = (h(y_j) + b_j)x_j(t) \tag{30b}$$

In which $h(x)$ is a type of logistic function called the tangent hyperbolic function, defined by function 31.

$$h(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}. \tag{31a}$$

$$h'(u) = 1 - h(x)^2. \tag{31b}$$

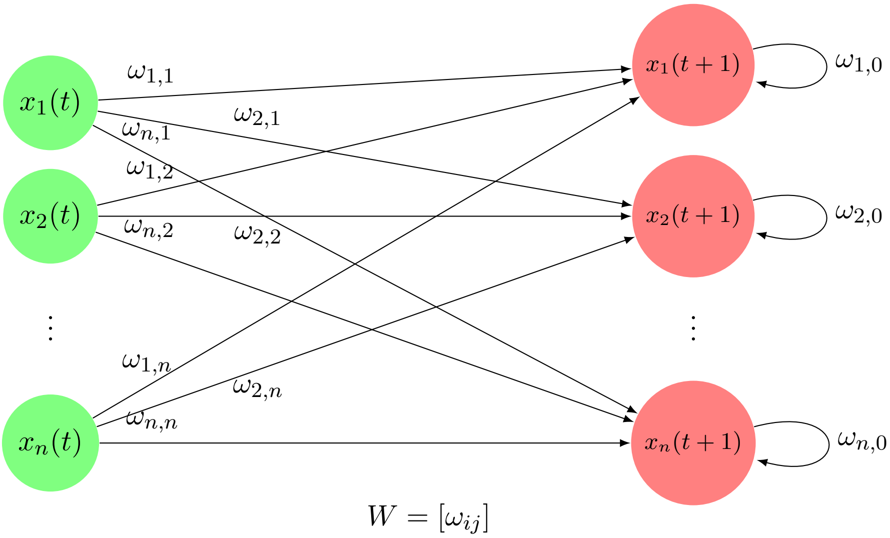


Figure 6: Non-linear prediction model. The method uses $x(t)$ to predict $x(t+1)$ using the weight matrix W determined by applying gradient-descent method.

After determining the weights, we can compute the infection prediction for day $t{+}1$ using infection data from day t with a system of linear regression

$$\begin{aligned}
x_1(t+1) &= \omega_{1,1}x_1(t) + \omega_{1,2}x_2(t) + \cdots + \omega_{1,n}x_n(t) + \omega_{1,0}. \\
x_2(t+1) &= \omega_{2,1}x_1(t) + \omega_{2,2}x_2(t) + \cdots + \omega_{2,n}x_n(t) + \omega_{2,0}. \\
&\vdots \\
x_n(t+1) &= \omega_{n,1}x_1(t) + \omega_{n,2}x_2(t) + \cdots + \omega_{n,n}x_n(t) + \omega_{n,0}. \qquad (32)
\end{aligned}$$

Equations 32 can also be rewritten as matrix calculation

$$x(t+1) = Wx(t) + B \qquad (33)$$

$$\begin{bmatrix} x_1(t+1) \\ x_2(t+1) \\ \vdots \\ x_n(t+1) \end{bmatrix} = \begin{bmatrix} \omega_{1,1} & \omega_{1,2} & \dots & \omega_{1,n} \\ \omega_{2,1} & \omega_{2,2} & \dots & \omega_{2,n} \\ \vdots & & \ddots & \vdots \\ \omega_{n,1} & \omega_{n,2} & \dots & \omega_{n,n} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} + \begin{bmatrix} \omega_{1,0} \\ \omega_{2,0} \\ \vdots \\ \omega_{n,0} \end{bmatrix}$$

Figure 6 is the visual representation of how the system of equations (32) and the matrix calculation 33 works. Infection from county x_1 on day t will have some effects (i.e. weight) on infection from county x_1 itself as well as counties x_2, x_3, and so on.

4.5 Gradient-Descent prediction algorithm

Having established the prediction equation and the learning rules, we have the algorithm 3 for COVID-19 prediction using gradient-descent.

Algorithm 3 Weight matrix gradient training algorithm

Input: Dataset $\Omega = \{x(t) = (x_1(t), \ldots, x_n(t)), 1 \leq t \leq n\}$
Output: Weight matrix W and column vector B

```
for $i, j \leftarrow 0$ to $n$ do
    $W_{ij} \leftarrow Random(-1, 1)$.
end for
for $epoch \leftarrow 0$ to $1000$ do
    for $t \leftarrow 0$ to $N - 1$ do
        $st \leftarrow x(t)$
        $st1 \leftarrow x(t+1)$
        for $j \leftarrow 1$ to $n$ do
            $yj \leftarrow 0$
            for $i \leftarrow 0$ to $n$ do
                $yj \leftarrow yj + w[i][j] \cdot st[i]$
            end for
            $zj \leftarrow (h(yj) + b[j] \cdot st(j)$
        end for
        $\Delta \leftarrow \alpha \cdot (yj - st1[j]) \cdot st[j] \cdot h'[j]$
        for $i \leftarrow 0$ to $n$ do
            $w[i][j] \leftarrow w[i][j] + \Delta$
        end for
        $b[j] \leftarrow b[j] + \Delta$
    end for
end for
return $W, B$
```

4.6 Example

For example, consider 3 counties with the following infection data

- $x_1(t) = 14155$.
- $x_2(t) = 2792$.
- $x_3(t) = 3074$.

We then randomly assigned the weights for each county. The weight represents the effects that one county has on another. In this example, the effect that $x_1(t)$ has on $x_1(t+1), x_2(t+1), x_3(t+1)$ are $0.623, 0.918, 0.9652$, respectively. Similarly, the weights of $x_2(t)$ are $0.239, 0.3732, 0.645$ and the weights of $x_3(t)$ are $0.742, 0.196, 0.24585$. The intercepts assigned to $x_1(t+1), x_2(t+1), x_3(t+1)$ are $0.302, 0.947, 0.3273$, respectively. Figure 7a shows how the weights work on each county.

From figure 7a, we can form the following system of linear regression equations

$$
\begin{aligned}
x_1(t+1) &= 0.623x_1(t) + 0.239x_2(t) + 0.742x_3(t) + 0.302 \\
x_2(t+1) &= 0.918x_1(t) + 0.3732x_2(t) + 0.196x_3(t) + 0.947 \\
x_3(t+1) &= 0.9652x_1(t) + 0.645x_2(t) + 0.24585x_3(t) + 0.3273
\end{aligned}
$$

The weights in figure 7a are not going to produce a very accurate results as they are generated randomly and linear regression tends to be linear instead of logarithmic or exponential, which is the usual rate of change for the spread of disease. In order to minimize the errors that those weight produce, we apply the gradient descent method shown previously in section 4.4. After one iteration, we obtain the following system of linear equations

$$
\begin{aligned}
x_1(t+1) &= 0.622x_1(t) + 0.238x_2(t) + 0.742x_3(t) + 0.301 = 15960 \\
x_2(t+1) &= 0.917x_1(t) + 0.3731x_2(t) + 0.196x_3(t) + 0.945 = 5158 \\
x_3(t+1) &= 0.9651x_1(t) + 0.645x_2(t) + 0.24583x_3(t) + 0.327 = 3848
\end{aligned}
$$

Similarly, after two iterations, we have the following system linear regression equations

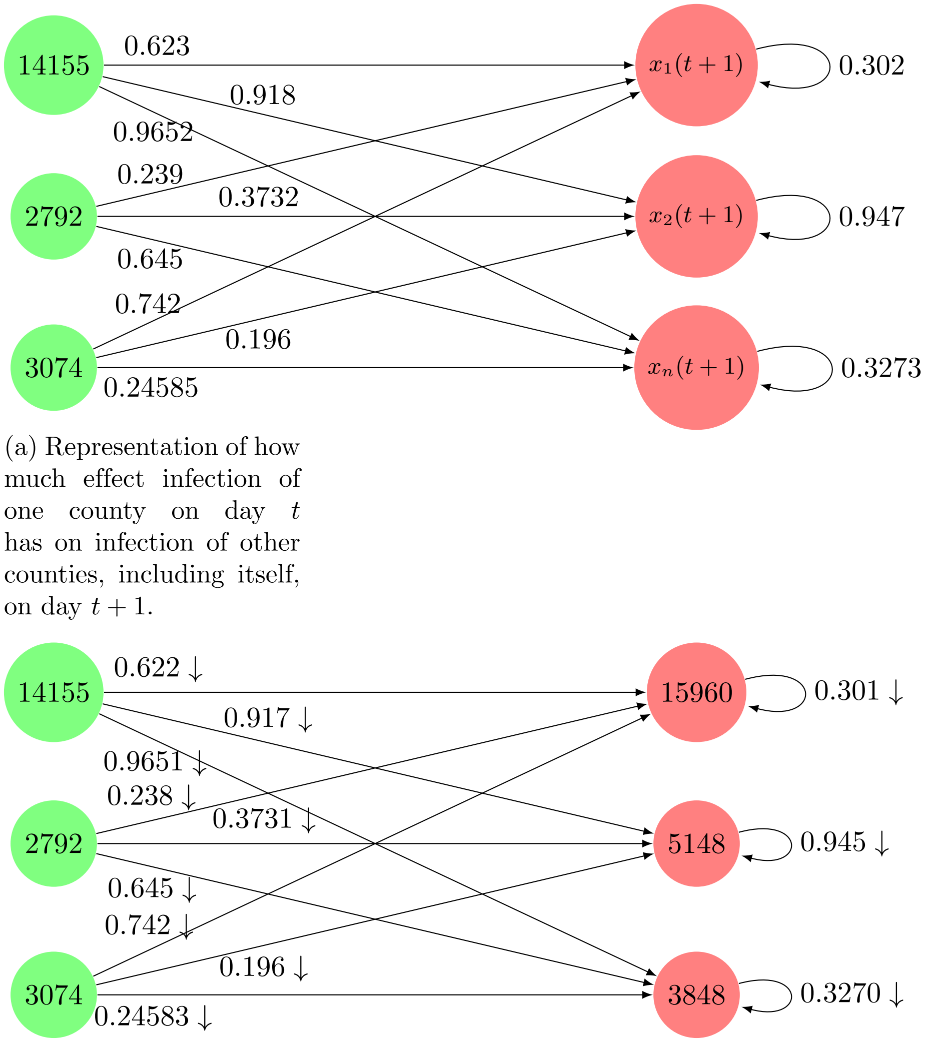
$$
\begin{aligned}
x_1(t+2) &= 0.6219x_1(t+1) + 0.239x_2(t+1) + 0.7419x_3(t+1) + 0.299 = 15934 \\
x_2(t+2) &= 0.9175x_1(t+1) + 0.373x_2(t+1) + 0.6457x_3(t+1) + 0.944 = 5145 \\
x_3(t+3) &= 0.719x_1(t+1) + 0.1963x_2(t+1) + 0.2458x_3(t+1) + 0.3273 = 3847
\end{aligned}
$$

5 Prediction Experimentation in Thu Duc City

Having established the calculation system for R_0 and the estimation algorithm, we will now predict the rise of COVID-19 in Thu Duc, Ho Chi Minh City.

Thu Duc is a municipal city, covering approximately 211.56 km^2, under the administration of Ho Chi Minh City, Vietnam. It was formed near the end of 2020 by combining three districts together: District 2, District 9, and Thu Duc District. It is an industrial concentrated area and comprises of many important businesses of Ho Chi Minh City. This research, supported by the GIS Center of the Science and Technology Department of Ho Chi

Figure 7: Two iterations of gradient-descent. After each iteration ,the weights as well as the infection rate changes slightly.



(a) Representation of how much effect infection of one county on day t has on infection of other counties, including itself, on day $t+1$.

(b) After applying one iteration of gradient descent, the weights of each county changes slightly. In this example, all the weights decrease slightly after one iteration, however, infection from each county increases.

Minh City, aims to predict the total number of COVID-19 positive cases in each ward of Thu Duc.

Thu Duc consists of 34 wards: An Khanh, An Loi Dong, Thao Dien, An Phu, Binh Trung Tay, Binh Trung Dong, Cat Lai, Thu Thiem, Thanh My Loi, Hiep Phu, Long Binh, Long Phuoc, Long Truong, Long Thanh My, Phu Huu, Phuoc Binh, Phuoc Long A, Phuoc Long B, Tang Nhon Phu A, Tang Nhon Phu B, Truong Thanh, Tan Phu, Binh Chieu, Binh Tho, Hiep Binh Chanh, Hiep Binh Phuoc, Linh Chieu, Linh Dong, Linh Tay, Linh Trung, Linh Xuan, Tam Binh, Tam Phu, and Truong Tho. In this research, An Loi Dong and Thu Thiem were not considered because the number of COVID-19 cases in those two wards became stable before the observed period.

5.1 Data Collection

Data regarding the daily total COVID-19 positive cases in Thu Duc was published on Thu Duc's website: https://thuduc-covid.hcmgis.vn. Figure 8 shows an excerpt of the website and how it works.

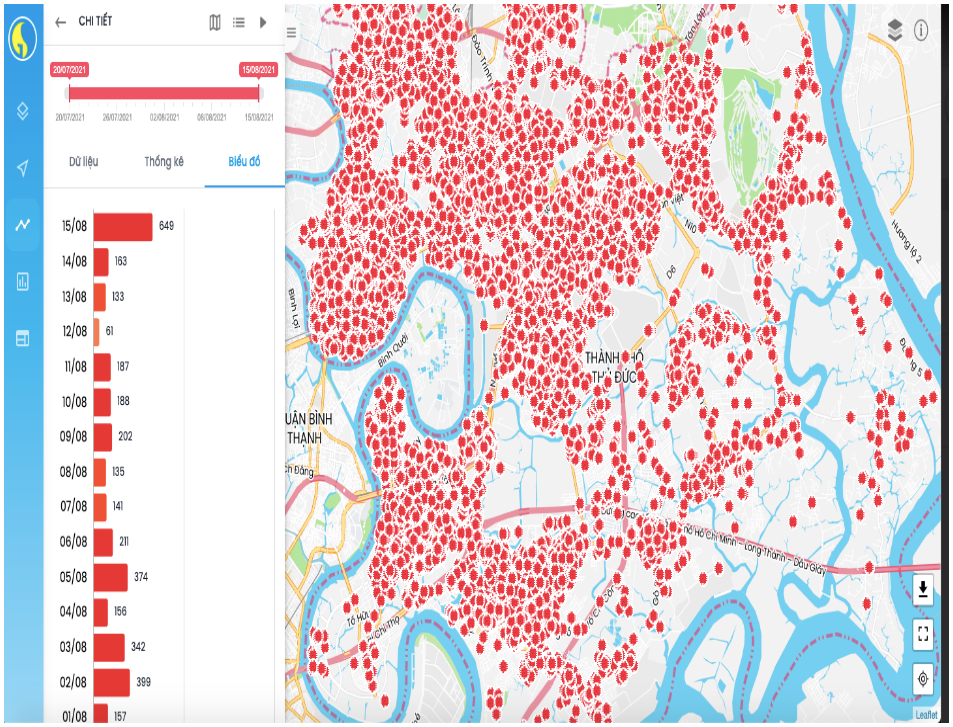


Figure 8: An excerpt of Thu Duc's GIS website https://thuduc-covid.hcmgis.vn. On the right is the map of Thu Duc. The red dots represent places that have COVID-19 positive residents. On the left is the data of total COVID-19 positive cases in each ward.

5.2 Results from using R_0 method

5.2.1 Results Based on Daily Prediction

Using the population density of the wards in Thu Duc, the reproductive value R_0 of each ward is initialized using equation 15 with the range $[1.2, 2.7]$. This range is based on the reported R_0 values of COVID-19 from other studies [21], [22]. After that, R_0 is updated by replicating the activities of residents in the ward of Thu Duc, with the current total number of cases in each ward being the inputs, and using equation 16, 17, 18 and the interaction coefficients $\alpha = 0.0001$ and $\beta = 0.1\alpha$. Figure 9 is the detailed comparison between predicted data and confirmed number of COVID-19 cases from each ward of Thu Duc on August 10, 2021. Figure 10 compares the total number of cases in Thu Duc from July 24, 2021, the day when the city's officials implemented social distancing policy, to August 10, 2021, roughly a week before Thu Duc locked down.

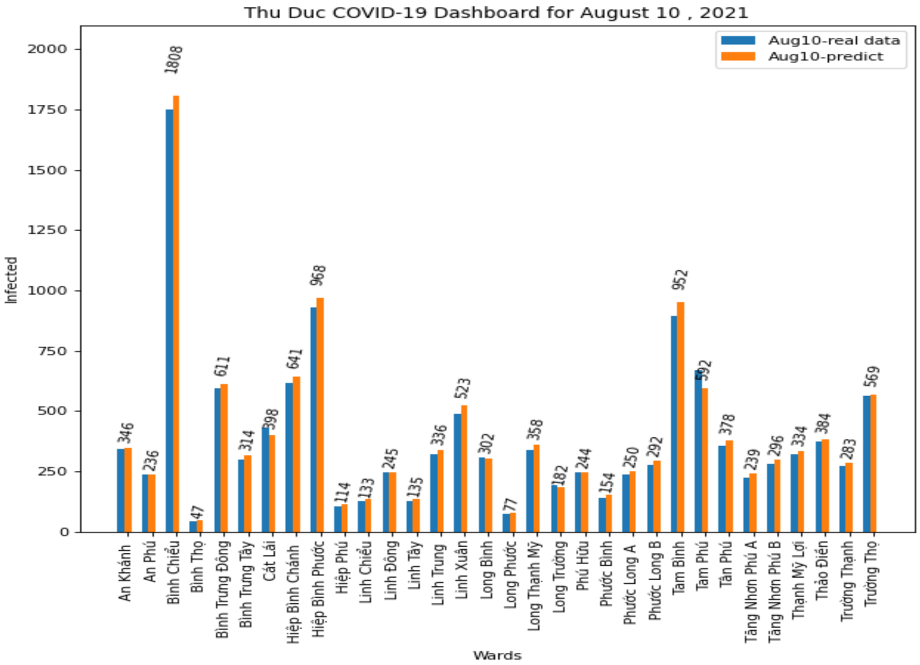


Figure 9: Comparison of predicted and real data for each ward of Thu Duc on August 10^{th}, 2021.

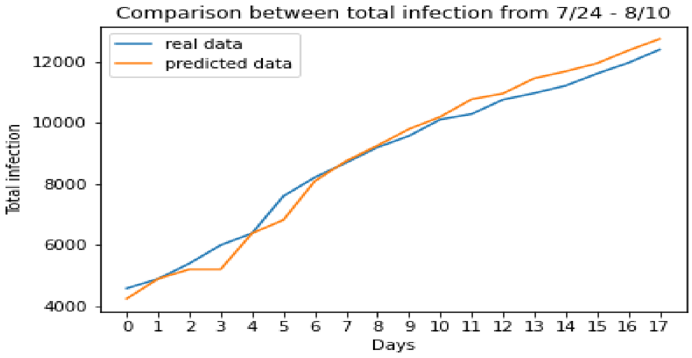


Figure 10: Comparison of predicted and real data for all of Thu Duc from July 24^{th} to August 10^{th}, 2021.

5.2.2 Optimal Forecasting Period using on Markov Chain

In order to perform long-term, daily prediction as in Section 5.2.1, we need to determine the most optimal time frame for prediction, meaning after how many days do we need to update the data again to continue predicting. For this reason, we implemented a simple Markov Chain to determine the optimal time frame.

6 COVID-19 Prediction for California Counties

6.1 Data Collection

Data regarding COVID-19 cases in each county and city of California is provided on the official website of California Health and Human Services Open Database [23]. We study and predict the infection data from January 1, 2021 to March 31, 2021 of the following counties: Imperial, Kern, Los Angeles, Orange, Riverside, San Bernadino, San Diego, San Luis Obispo, Santa Barbara, Ventura, and Yuba.

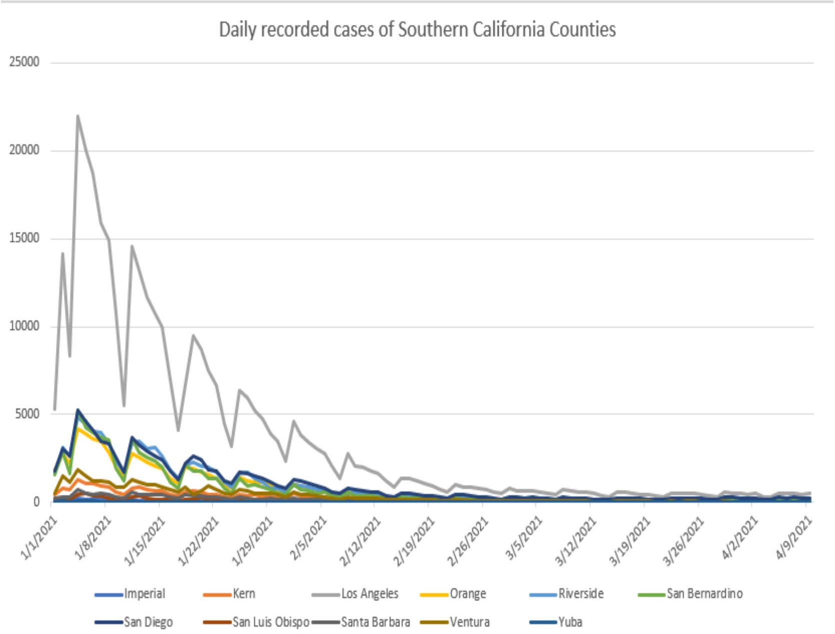


Figure 11: Daily recorded cases in counties in Southern California from 1/2/2021 to 9/4/2021.

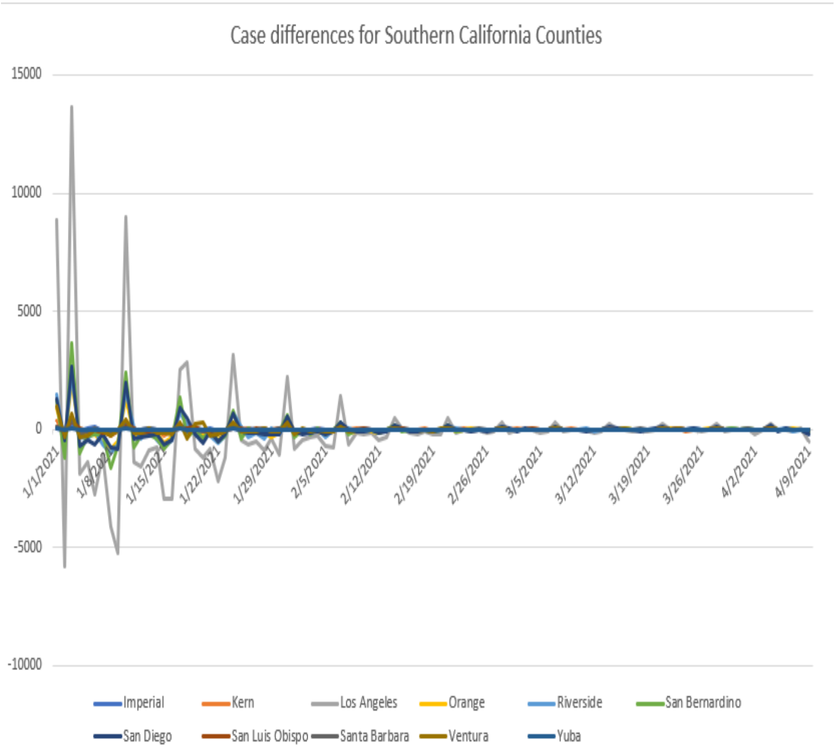


Figure 12: Difference between old and new cases in counties in Southern California from 1/2/2021 to 9/4/2021.

6.2 Prediction results from Gradient-Descent method

From using the gradient-descent method with 5000 iterations and applied it on the infection data of Southern California counties, achieve the results of

- $R2 = 0.2235384078469438$
- $RMSE = 110.25295618541874$

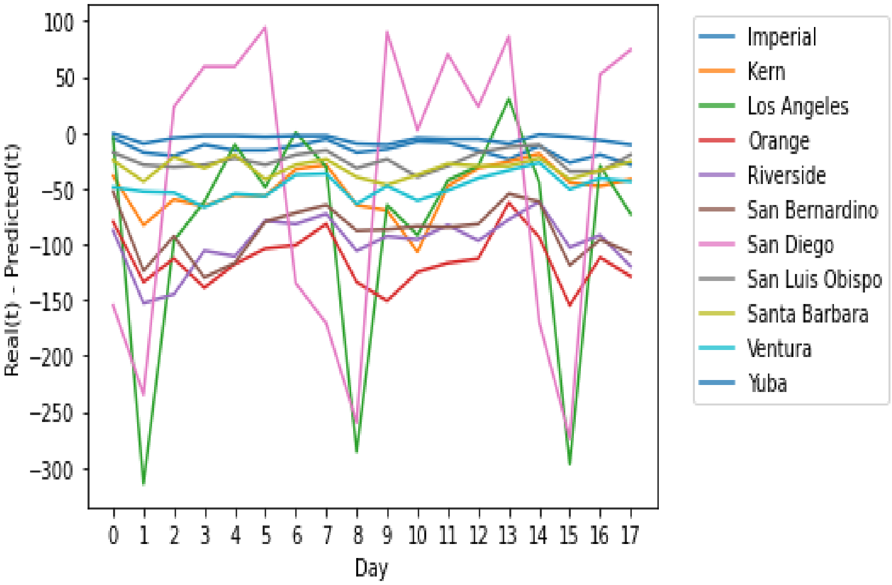


Figure 13: The difference between real infection data and predicted infection data for counties in Southern California from March 14, 2021 to March 31, 2021 when using the gradient-descent method.

Using the standard linear regression method, we were able to achieve the following results

- $R^2 = -13.830695590520959$
- $RMSE = 368.9376798609871$

7 Discussion

7.1 Result significance

This thesis aims to develops two prediction methods: a method uses R_0 coefficient to prediction infection rate and the other uses gradient-descent and multiple linear regression to determine infection. We have not compared the two methods on the same dataset, but that is because of the nature of the methods. R_0 method is intended traveling from district to district, and gradient-descent method is intended for bigger distance traveling, i.e. moving from county to county. Though both methods are shown to achieve significant results, there are some advantages and disadvantages of each methods that need to be discusses.

7.1.1 R_0-based prediction

One of the biggest advantages of this predicting method is that it is able to produce relatively significant results with only minimal data. The only data necessary for the R_0-based method to work is the R_0 value of the disease, the population and the population density of the nodes or studied area. Additionally, the method is able to compensate for what two popular disease forecasting methods, mathematical modeling and agent-base modeling, lack in.

Firstly, Mathematical modeling, while simplistic and easy to produce, is too general and too homogeneous while diseases are much more complex and complicated [2]. Mathematical modeling may provide a general outline of the disease, but will be inaccurate once the behavior of disease changes (e.g. each variant of COVID-19 has a greater R_0 value than its predecessor [24], [25]).

Secondly, Agent-based modeling (ABM) tries to simulate the heterogeneous behaviors of human and diseases, however, it requires too much calculation because ABM has to keep track of every single interaction that an agent makes [2]. Furthermore, ABM cannot incorporate real-time data which makes it difficult to predict the spread of disease in short-term.

The R_0-based method is not only simplistic and easy to produce as it only needs the daily infection rate to perform short-term prediction for spread of disease, it also applies the "random walk on graph" method, thus taking into account the mobility and movements of each agent. In addition, it can perform well with real-time data.

However, one of the main disadvantage that this method has is that it can only perform prediction in a short period of time. As previously shown in section 3.4, the R_0-based method can only produce effective results up to 3 days period. After 1 or 3 days, it needs to update with the real data. Furthermore, this method also has high time complexity, which can be troublesome when applied to larger dataset.

7.1.2 Gradient-Descent prediction

The gradient-descent method is created based on the interaction of each county within a big state. Similar to the R_0-based method, this method assumes that residents will travel from one county to another and infect or get infected by one another. Because of this assumption, this method could potentially be applied for contact-tracing and forecasting spread of disease in a travel network such as a train network or an airline network.

The accuracy scores of the gradient-descent method when applied to infection data of Southern California counties is not as good as the R_0 method when applied to Thu Duc's infection data. This is because infection is more likely to occur and spread much faster in urban areas such as a city, not big, expanded areas like a county [26]. Furthermore, infection data shows only Los Angeles county, Sand Diego, San Bernardino, and Orange county are active in spreading infection. Furthermore, as shown in figure 11 and figure 12, cases in Los Angeles county are significantly higher than cases in other counties. The other counties either spreads infection so slowly that infection data from those counties become insignificant. This heavily affects the weight matrix when training the data, which leads to inaccuracies in prediction result.

Although the prediction results are not perfect and the time complexity is high, the gradient-descent method is found to be a better predictor for spread of disease than the normal linear regression method as infection rate tends to be non-linear.

7.2 Future work

A graph is one of the most important data structures in computer science because it is incredibly useful in modeling abstractions and solving problems. Graphs play an important role in studying social networks, transportation networks, biological networks, etc [27].

Graph partitioning is defined as a problem in which we take a large graph and cut its edges, thus dividing it into several small sub-graphs. In the era of big data, large and complex graph structures are created, and graph partitioning becomes a more important problems as it reduces the level of complexity of such big structure and allows scientists to study the graph more closely. In 2017, a graph partitioning framework developed by Nazi et al. to generalize and partition big graph structure in a fast manner, even on unseen graph [28].

On the other hand, there is proof-of-concept about graph partitioning being applied to study the spread of disease and epidemiology [29]. In the future, we would like to turn the United States into a big graph network, with each state is represented by a node and any transportation connection between the states is represented by an edge. We would like to then apply graph partitioning, and lastly apply the 2 proposed methods to predict the spread of disease on a sub-graph. We would like to see how the proposed methods perform on sub-graph structure as well as how each sub-graph interact to one another.

8 Conclusion

We propose two methods of COVID-19 prediction method: the R_0-based method and the gradient-descent method.

Firstly, for the R_0-based prediction method, we represented the studied city as a complete, undirected graph structure with each node represents a district. We applied the random walk on graph to simulate the mobility of the district's residents and how disease can transmit from one resident to another based on the R_0 value. We achieve accurate results in short-term prediction of COVID-19 cases in Thu Duc city using the R_0-based method.

Lastly, for the gradient-descent method, we used gradient-descent to determine the weight matrix and utilized the current infection data to predict future infection data. We applied this method to the counties of Southern California and were able to achieve relative success. Our gradient-descent method produced higher accuracy in short-term prediction than the standard linear regression method.

Bibliography

[1] K. Khan, J. Arino, W. Hu, P. Raposo, J. Sears, F. Calderon, C. Heidebrecht, M. Macdonald, J. Liauw, A. Chan, and M. Gardam, "Spread of a novel influenza a (h1n1) virus via global airline transportation", *The New England journal of medicine*, vol. 361, pp. 212–4, Jul. 2009. DOI: 10.1056/NEJMc0904559.

[2] J. Li, T. Xiang, and L. He, "Modeling epidemic spread in transportation networks: A review", *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 8, no. 2, pp. 139–152, 2021, Transportation Planning and Operations for COVID-19 Epidemic and Other Emergencies, ISSN: 2095-7564. DOI: https://doi.org/10.1016/j.jtte.2020.10.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095756420301616.

[3] T. Hollingsworth, N. Ferguson, and R. Anderson, "Frequent travelers and rate of spread of epidemics", *Emerging infectious diseases*, vol. 13, pp. 1288–94, Oct. 2007. DOI: 10.3201/eid1309.070081.

[4] A. Tatem, D. Rogers, and S. Hay, "Global transport networks and infectious disease spread", in *Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications*, ser. Advances in Parasitology, S. I. Hay, A. Graham, and D. J. Rogers, Eds., vol. 62, Academic Press, 2006, pp. 293–343. DOI: https://doi.org/10.1016/S0065-308X(05)62009-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0065308X0562009X.

[5] A. Ceria, K. Köstler, R. Gobardhan, and H. Wang, "Modeling airport congestion contagion by heterogeneous sis epidemic spreading on airline networks", *PLOS ONE*, vol. 16, no. 1, pp. 1–17, Jan. 2021. DOI: 10.1371/journal.pone.0245043. [Online]. Available: https://doi.org/10.1371/journal.pone.0245043.

[6] D. Smith and L. Moore, "The sir model for spread of disease", *Convergence*, Dec. 2004.

[7] W. Yang, B. J. Cowling, E. H. Y. Lau, and J. Shaman, "Forecasting influenza epidemics in hong kong", *PLOS Computational Biology*, vol. 11, no. 7, pp. 1–17, Jul. 2015. DOI: 10.1371/journal.pcbi.1004383. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1004383.

[8] O. N. Bjørnstad, K. Shea, M. Krzywinski, and N. Altman, "The seirs model for infectious disease dynamics", *Nature Methods*, vol. 17, no. 6, pp. 557–558, Jun. 2020, ISSN: 1548-7105. DOI: 10.1038/s41592-020-0856-2. [Online]. Available: https://doi.org/10.1038/s41592-020-0856-2.

[9] J. Heesterbeek, "A brief history of r_0 and a recipe for its calculation.", *Acta Biotheoretica*, vol. 50, pp. 189–204, Feb. 2002. DOI: 10.1023/A:1016599411804.

[10] N. C. Achaiah, S. B. Subbarajasetty, and R. M. Shetty, "of COVID-19: Can We Predict When the Pandemic Outbreak will be Contained?", *Indian J Crit Care Med*, vol. 24, no. 11, pp. 1125–1127, Nov. 2020.

[11] Center for Application of Geographical Information System. (2021). "Thu duc covid maps", [Online]. Available: https://thuduc-covid.hcmgis.vn/#/maps/layers (visited on 07/02/2021).

[12] S. Sanche, Y. T. Lin, C. Xu, E. Romero-Severson, N. Hengartner, and R. Ke, "High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2", *Emerging Infectious Diseases*, vol. 26, no. 7, pp. 1470–1477, Jul. 2020. DOI: 10.3201/eid2607.200282.

[13] M. M. 1. Meerschaert, *Mathematical modeling*, English, Waltham, MA, 2013. [Online]. Available: https://birmingham-primo.hosted.exlibrisgroup.com/openurl/44BIR/44BIR_VU1?u.ignore_date_coverage=true&rft.mms_id=9933521527404871.

[14] L. László, "Random walks on graphs: A survey, combinatorics, paul erdos is eighty", *Bolyai Soc. Math. Stud.*, vol. 2, Jan. 1993.

[15] N. E. Breslow, N. E. Day, and E. Heseltine, "Statistical methods in cancer research", 1980.

[16] P. Royston and W. Sauerbrei, "Building multivariable regression models with continuous covariates in clinical epidemiology–with an emphasis on fractional polynomials", *Methods Inf Med*, vol. 44, no. 4, pp. 561–571, 2005.

[17] M. A. Khan, R. Khan, F. Algarni, I. Kumar, A. Choudhary, and A. Srivastava, "Performance evaluation of regression models for covid-19: A statistical and predictive perspective", *Ain Shams Engineering Journal*, vol. 13, no. 2, p. 101 574, 2022, ISSN: 2090-4479. DOI: https://doi.org/10.1016/j.asej.2021.08.016. [Online]. Avail-

able: https://www.sciencedirect.com/science/article/pii/S2090447921003385.

[18] B. Wilder, M. Charpignon, J. A. Killian, H.-C. Ou, A. Mate, S. Jabbari, A. Perrault, A. N. Desai, M. Tambe, and M. S. Majumder, "Modeling between-population variation in covid-19 dynamics in hubei, lombardy, and new york city", *Proceedings of the National Academy of Sciences*, vol. 117, no. 41, pp. 25 904–25 910, 2020. DOI: 10.1073/pnas.2010651117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2010651117. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2010651117.

[19] M. I. Vicuña, C. Vásquez, and B. F. Quiroga, "Forecasting the 2020 covid-19 epidemic: A multivariate quasi-poisson regression to model the evolution of new cases in chile", *Frontiers in Public Health*, vol. 9, 2021, ISSN: 2296-2565. DOI: 10.3389/fpubh.2021.610479. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpubh.2021.610479.

[20] H. Daume III, "The perceptron", in *A Course in Machine Learning*, (book chapter), ciml.info, 2012. [Online]. Available: http://ciml.info/dl/v0_8/ciml-v0_8-ch03.pdf.

[21] H. Salje, C. Tran Kiem, N. Lefrancq, N. Courtejoie, P. Bosetti, J. Paireau, A. Andronico, N. Hozé, J. Richet, C. L. Dubost, Y. Le Strat, J. Lessler, D. Levy-Bruhl, A. Fontanet, L. Opatowski, P. Y. Boelle, and S. Cauchemez, "Estimating the burden of SARS-CoV-2 in France", *Science*, vol. 369, no. 6500, pp. 208–211, Jul. 2020.

[22] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Y. Lam, J. T. Wu, G. F. Gao, B. J. Cowling, B. Yang, G. M. Leung, and Z. Feng, "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia", *N Engl J Med*, vol. 382, no. 13, pp. 1199–1207, Mar. 2020.

[23] California Health & Human Service Agency. (2020). "Covid-19 time-series metrics by county and state", [Online]. Available: https://data.chhs.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state (visited on 01/01/2022).

[24] Y. Liu and J. Rocklöv, "The reproductive number of the Delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus", *J Travel Med*, vol. 28, no. 7, Oct. 2021.

[25] K. Ito, C. Piantham, and H. Nishiura, "Relative instantaneous reproduction number of Omicron SARS-CoV-2 variant with respect to the Delta variant in Denmark", *J Med Virol*, Dec. 2021.

[26] J. Aguilar, A. Bassolas, G. Ghoshal, S. Hazarie, A. Kirkley, M. Mazzoli, S. Meloni, S. Mimar, V. Nicosia, J. J. Ramasco, and A. Sadilek, "Impact of urban structure on infectious disease spreading", *Sci Rep*, vol. 12, no. 1, p. 3816, Mar. 2022.

[27] A. Buluç, H. Meyerhenke, I. Safro, P. Sanders, and C. Schulz, "Recent advances in graph partitioning", in. Nov. 2016, vol. 9220, ISBN: 978-3-319-49486-9. DOI: 10.1007/978-3-319-49487-6_4.

[28] A. Nazi, W. Hang, A. Goldie, S. Ravi, and A. Mirhoseini, "GAP: generalizable approximate graph partitioning framework", *CoRR*, 2019. arXiv: 1903.00614. [Online]. Available: http://arxiv.org/abs/1903.00614.

[29] J. Hadidjojo and S. A. Cheong, "Equal graph partitioning on estimated infection network as an effective epidemic mitigation measure", *PLoS One*, vol. 6, no. 7, e22124, 2011.