

<THE DISSERTATION TITLE>

By

Shuming Zhou

1611144

Supervised By

Nanlin Jin

A DISSERTATION

Submitted to

Xi’an Jiaotong-Liverpool University

in partial fulfillment of the requirements

for the degree of

MASTER OF RESEARCH

<2022.11.23>

ABSTRACT

This dissertation research the prediction model of COVID-19 both in quantitative and qualitative approaches. The policy factors are involved to see if it have any influence on the prediction models. Several models has been made into software codes such as adapted genetic algorithm and linear regression algorithm. The results shows that the policy factors can influence the accuracy of some prediction models. It is concluded that with the policy attributes involved, the linear regression model will greatly increase its accuracy. Also, the qualitative adapted genetic algorithm is better than the previously researched one.

DECLARATION

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions or writings of another.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

Shuming Zhou

ACKNOWLEDGEMENTS

I want to express my gratitude to Nanlin, Jin. She is my supervisor and tutor. She helped me a lot of this dissertation. I also want to express my gratitude to Jie Zhang for her giving me such a chance to do this research. I really appreciate her advice. I also appreciate to all the XJTLU lecturers, I learned a lot from those courses. I also want to thank my parents who support all of my research and life.

**TABLE OF CONTENT**

[LIST OF TABLES vi](#_Toc95215279)

[LIST OF FIGURES vii](#_Toc95215280)

[Chapter 1. Introduction 1](#_Toc95215281)

[Chapter 2. Literature Review 2](#_Toc95215282)

[Chapter 3. Research Methodology 3](#_Toc95215283)

[Chapter 4. Results 4](#_Toc95215284)

[Chapter 5. Analysis 5](#_Toc95215285)

[Chapter 6. Discussion 6](#_Toc95215286)

[Chapter 7. Conclusions 7](#_Toc95215287)

[REFRENCES CITED 8](#_Toc95215288)

[Appendix A. Xxxxx 9](#_Toc95215289)

[A.1 Xxxxxxx 9](#_Toc95215290)

LIST OF TABLES

Page

[Table 1 Attribute of UK daily cases 14](#_Toc120450594)

[Table 2 Simulating adapted GA parameters 24](#_Toc120450595)

[Table 3 Needed columns of Coronanet Project 27](#_Toc120450596)

[Table 4 Policies listed in Coronanet Project 28](#_Toc120450597)

[Table 5 Attributes of daily cases from Our World in Data 29](#_Toc120450598)

[Table 6 UK pattern distance 37](#_Toc120450599)

[Table 7 China pattern distance 37](#_Toc120450600)

[Table 8 Chinese Linear Regression results 38](#_Toc120450601)

[Table 9 Chinese Ridge Regression result 38](#_Toc120450602)

[Table 10 UK Linear Regression result 38](#_Toc120450603)

[Table 11 UK Ridge Regression result 38](#_Toc120450604)

LIST OF FIGURES

Page

[Figure 1 Typical genetic algorithm 7](#_Toc120373147)

[Figure 2 Oxford daily new cases result 11](#_Toc120373148)

[Figure 3 Oxford daily cases during period 12](#_Toc120373149)

[Figure 4 John Hopkins Chinese government daily cases description 14](#_Toc120373150)

[Figure 5 Punn, Sonbhadra and Agarwal's machine learning algorithm result 14](#_Toc120373151)

[Figure 6 SIR model result（infected population over time） 16](#_Toc120373152)

[Figure 7 Agent-based model prediction of covid-19 （infected population over time） 18](#_Toc120373153)

[Figure 8 Game theory simulating of COVID-19 infected 19](#_Toc120373154)

[Figure 9 python code importing sklearn library 28](#_Toc120373155)

[Figure 10 Adapted GA simulating UK result 30](#_Toc120373156)

[Figure 11 Adapted GA simulating China result 31](#_Toc120373157)

[Figure 12 China real data（infected population over time） 32](#_Toc120373158)

[Figure 13 UK real data（infected population over time） 32](#_Toc120373159)

[Figure 14 The tendency graph of the Oxford research 33](#_Toc120373160)

[Figure 15 Chinese Linear Regression over time 35](#_Toc120373161)

[Figure 16 Chinese Ridge Regression over time 36](#_Toc120373162)

[Figure 17 UK Linear Regression over time 37](#_Toc120373163)

[Figure 18 UK Ridge Regression over time 38](#_Toc120373164)

# Introduction

## Background

### Covid 19 situation

From 2020, COVID-19 has been a worldwide pandemic. There are many researches related to that topic. Each country takes their own policies and strategies to deal with this pandemic. Thus, it will be helpful to build a system which can help government to forecast the developing of COVID-19. According to WHO [1], the current global infected population is 626,337,158. COVID-19 is mutated from SARS-CoV, it can bring different symptoms such as Tinnitus, gingivitis, sudden hearing loss, Bell’s palsy [2] .

COVID-19 also brings many social issues such as social distancing. Social distancing avoids people from offline meeting which will bring high social costs [3].

Thus, it is useful to predict the development of COVID-19. The prediction results can be used to formulate new policies to decrease the influence of COVID-19 to the public. Thus, predicting the COVID-19 is the focus of this dissertation.

### Brief introduction of previous research knowledge

目前的研究分为以下几个方向

对新冠疫情的类型分三大主体：对新冠疫情下的人民的研究(和上面相同)，对新冠疫情本体的研究，比如病毒本身的研究(https://www.sciencedirect.com/science/article/pii/S1567134820302537)医疗方面(https://www.sciencedirect.com/science/article/pii/S1521661620303181)以及对新冠疫情发展的预测(file:///Users/mac/Documents/GitHub/SAT405\_program/exercise/Prediction\_COVID\_19.pdf)，以及对政府的研究：政策的研究(https://www.nature.com/articles/s41562-020-0909-7)。（这三个）

对人民的研究主要是对疫情的心理和生理的评估，和本研究关系不大，而对病毒本身的研究以及医疗方面的研究涉及具体的医学和生物学知识，不在本文的探讨范围之内，其次就是对疫情的未来发展的趋势，这是对本文极其相关的点，具体也会在下文探讨。

对政府的研究主要是对NPI的研究，（https://www.sciencedirect.com/science/article/pii/S0048969721005982）NPI代表了具体的政策实施。当然也有对政策经济代价的研究，比如(https://www.medrxiv.org/content/10.1101/2020.03.26.20044552.abstract)，提到了中英两种模式下的经济代价。

再具体一些，对于这些研究，很多都是对单个因素的研究，比如单纯探讨疫情下人的心理生理状况，以及单独对npi的研究，其中也有很多是两个要素综合起来研究的，比如oxford的论文(file:///Users/mac/Documents/GitHub/SAT405\_program/405%E6%96%87%E7%AB%A0.pdf)，探讨病毒传播率以及NPI的政府政策，以及

<https://www.jmir.org/2020/9/e21419/>

这个文章主要是探讨政策和人民的态度。

当然最和本研究主题相关的，是政府政策以及病毒传染率这两个要素。对个人态度选择而言并不在本topic的论述范围之内。

There are several research directions towards COVID-19 prediction. The SIR with its derivatives (SEIR), the agent-based model, heuristic algorithm (genetic algorithm), machine learning and deep learning algorithm and game theory.

Below is the brief introduction of the models of these research directions.

#### SIR model

According to Smith and Moore[4], SIR represents different population in the COVID-19 pandemic.

S = S(t) = susceptible population

I = I(t) = infected population

R = R(t) = recovered population

This model uses these 3 parameters to construct equations and formulations to simulate the development of the COVID.

#### Agent-based model

There are many agents in the model and each agent will have its states and actions. For example, Shamil et al. [5] defines each person as an agent and susceptible to COVID-19. Each agent will have 5 different states: HNASD to show if they are infected or healthy. Also, each agent will have their own profession and each profession will have corresponding tasks.

Agents are associated with groups based on their tasks and will interact with each other. There are 5 different groups to separate these agents: stay at home, commute, work or attend school, attend event stay at the hospital. In each group, the agents will be allocated to another group in order to realize the transmission in order to change their states.

#### Genetic algorithm

The genetic algorithm is used to let the population keep evolving and use fitness function to select the good one in order to make the population change over period. For example, according to the oxford paper, the “population” is the policies. The initial policies combination will evolve over time and it can then get the newly changed combinations of the policy.

GA (genetic algorithm) is a sort of algorithm which is used to simulate the natural selection process by Darwin according to Mathew [6]. Each individual will meet selection in natural environment. Those who does not adapt to the environment will eliminate while those fit the environment will survive. There are also mutation in the evolutionary process, the mutation will bring the offspring will traits which their predecessor do not have. Parents can crossover to bring their traits to their offspring. The GA has the characteristics described above: mutation, selection and inheritance. Each individual in GA will be represented as a genome. That is, a string of characters or numbers, just like the gene in the natural environment. Typically, the characters of the string are numbers, for example, each bit is 0 or 1. For example, if there is a need to represent the drink in the market, the drink can be accessed by 3 criteria: has sugar or not, has mineral substance or not, has fat or not. Therefore, the purified water can be represented as 0-1-0 which means it does not have any sugar but it has mineral substance and do not have any fat. A specific GA can be described like the flow chart below:

图形用户界面, 图示

描述已自动生成

Figure 1 Typical genetic algorithm

There are many turns in the entire GA loop process. Firstly, each string of characters will be initialized to get the current population, each string will be set as the combination of 0s and 1s. Secondly, selection will be made, those who does not fit will eliminate but those fit will survive. The GA has a function to judge if an individual adapts to the environment or not, that is the fitness function. The input of the function is the string of each individual. In figure 1, the string 3 is eliminated. After selection is the crossover process. String 1 and string 2 do such a process and generate the new offspring: new string 1 and new string 2. In order to describe the process of the crossover, it is very convenient to make an example: assuming that string 1 =1-0-0-0-0-1, string 2 =1-1-1-0-1-0. Then the new offspring will be new string = 1-0-0-0-1-0. The 1-0-0 at start is from the top 3 bit of string 1, 0-1-0 is from the bottom 3 bit of string 2. The last stage is mutation. In this stage, each string will have a chance to mutation. For example, if string =0-0-1-1-0-1, typically the mutation process will make each bit of the string to mutate. The third of the string is 1, then after mutation, the 1 will be changed to a value which is not the original value 1. Since here each bit can only be 0 or 1, thus it can only be 0 after mutation. Each bit will have a mutation probability to decide if that bit will change or not. After mutation, that loop is over and turn to the next loop.

#### Machine learning and Deep learning

Machine learning related to predict COVID-19 data are mainly supervised learning. To be specific, it is mainly the regression not the classification since the output data of the model is continuous values.

Deep learning is mainly about building the neural network, the input data will be fed into the model, each node of the neural network will use a function and the weight of the signal to do the calculating.

#### Game theory

The players of game theory is the most significant factors. Each player will have his/or strategies and each strategy will have a corresponding paid-off. The paid-off can be simply seen as the money got. It also have minus values and plus values which means get profit or pay for something.

## Research gap

Based on the previous introduction of the previous research and the forward literature review, there are lack of policy factors and real time data involved in simulating and predicting the COVID-19. However, the policy factor is a significant factors to be considered. For example, the re-opening economy strategy of UK will cause loosing the prevention of COVID-19 [7]. By comparing the new cases of UK and China and their anti-virus policies, a hypothesis can be made: the policy of a country will influence their own COVID-19 development [8]. Thus, it is necessary to take policy into consideration.

## Topic Of Research Paper

Based on the brief introduction of previous research knowledge, the topic of this dissertation can be driven. This paper is mainly focusing on the policy factors in predicting the COVID-19 development, to be specific, its daily new cases. The question to be solved is how the policy will influence the prediction of Covid-19 daily cases. In other words, if the policy factor is introduced, will prediction of Covid-19 daily cases be more accurate or less accurate? How much accuracy or how much influence on the prediction result will be? Thus, new models with policy factors will be introduced to answer this research question and its accuracy will be used to evaluate the answer.

# Literature Review

## Oxford Genetic algorithm by Vie [9]

### Data

There is no real data used as input.

### Method

It uses genetic algorithm. By using the genetic algorithm, the changing of the policy combination can be simulated during the entire process.

### Results

There are 3 different situations in the output.

1. Both the virus and the policy will evolve
2. Only virus evolve
3. Only policy evolve

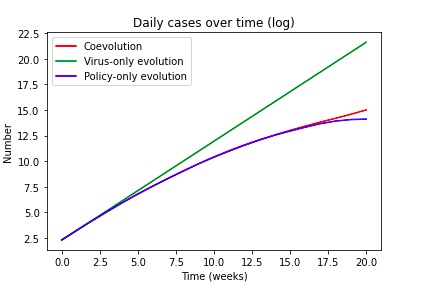


Figure 2 Oxford daily new cases result

### Gaps

It is obvious that there is no real time data. Therefore, when looking at its results and comparing to the real data, there is a huge difference between them.

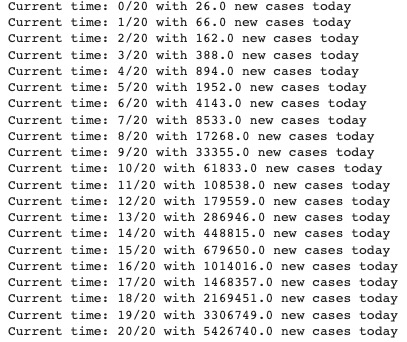


Figure 3 Oxford daily cases during period

This is the simulated data given by Oxford algorithm. It is obvious that in the simulated results, the new cases each time will only increase but not decrease. That is different from the reality.

The reason may be the parameter of this model is not set correctly. It said the parameter can be easily changed to get different result. However, it is still useful to use genetic algorithm as a simulating tool for qualitative research to predict the tendency of Covid-19.

## Machine learning and deep learning algorithm by Punn, Sonbhadra and Agarwal [10]

### Data

The data is achieved from the github website by Johns Hopkins whiting school of engineering [11]. The webpage claims its data is from Johns Hopkins University with different countries. For example, the attributes of data from UK daily cases are shown in the data below.

Table 1 Attribute of UK daily cases

|  |
| --- |
| areaType |
| areaName |
| areaCode |
| Date |
| newCasesBySpecimenDate |
| cumCasesBySpecimenDate |
| new-FirstEpisodes-BySpecimenDate |
| cum-FirstEpisodes-BySpecimenDate |

It claims that it is valid since it is directly extracted from government website.

Actually, most of the data resource is government website so the data quality is good.

However, there are also some problems with this github website. Since this website is just a collection of these datasets, it does not do any reorganizing or improvement. Different countries have different table columns and are not unified, it may cause problems when doing unified data pre-processing. For example, the dataset of China has such fragment:

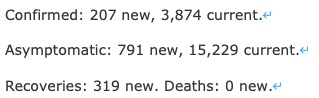


Figure 4 John Hopkins Chinese government daily cases description

That is different from the UK dataset. And according to that paper, only 6 parameters of the dataset columns will be used: province/state, country/region, last update, confirmed, death and recovered cases.

### Method

There are both deep learning and machine learning algorithms in that paper. These algorithms are implemented by python “sklearn” library, such as SVR, DNN, LSTM and PR.

### Evaluation

The mean squared error (MSE) is the most widely used objective function and root mean square error (RMSE) as a metric function for evaluating the regression models.

### Results

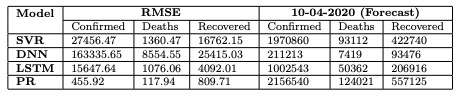


Figure 5 Punn, Sonbhadra and Agarwal's machine learning algorithm result

### Gaps

The evaluation methods are not enough. For example, the R2 value is not included as the input. Also, the model does not include the parameters related to policy. In other words, policies corresponds to COVID-19 are not taken into consideration.

However, deep learning methods and machine learning methods are good to predict the new cases.

## SIR model by Wang et al. [12]

### Data

Data is from China CDC.

### Method

Using SIR model to predict the future data.

### Evaluation

However, it is now hard to evaluate the model. In that paper, the author put the results into the website to indicate his predicted data and real data there to see their differences. But, the website now is dropped and no data can be found there.

### Results

It uses a graph to show the result, it is actually only a vision since it provides 2 cases: the loosening case and strict case.

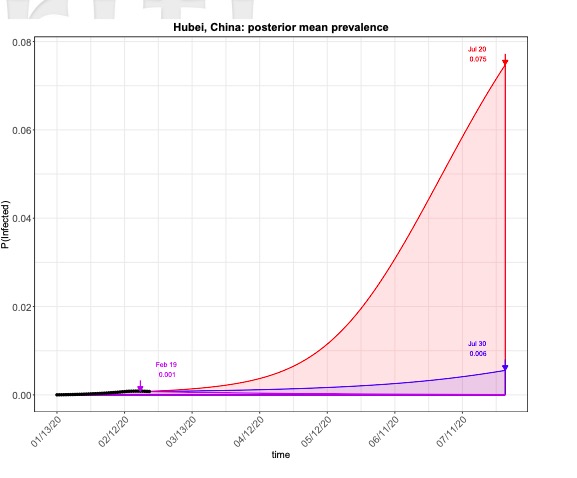


Figure 6 SIR model result（infected population over time）

### Gaps

It is obvious that it uses many Math tools so it is based on solid theory. However, it is quite complicated because of using not only SIR but also Markov chain. Also, it does not pay attention to the policy factor.

## Agent-based model by Shamil et al. [5]

### Data

2 categories of data are used: the Location-specific data and the Physiological data.

#### Location-specific data

1. demographics of the inhabitants in a particular city (i.e., education, employment, life expectancy, percentage of individuals having different professions, and the nature and timing of various tasks performed by the people) the data related to the number of transports and the average family size.
2. the data related to COVID-19 disease, its spread among the population, and the intervention measures taken by the authorities. These include the number of infections in the city and the day of the announcement of restrictive policies or awareness measures.

#### Physiological data

The probability of a person coughing and sneezing, touching contaminated objects, coming into physical contact with others, or washing hands is also an important parameter of our model, which would differ based on whether a person is at work, home, or hospitalized.

### Method

Since it is agent-based model, then the concept of agent should be explained. Each person in this model is an agent and susceptible to COVID-19. The states of each agent are 5 in total: healthy, asymptomatic, infected, symptomatic, dead or recovered.

Each agent is associated with a family and is assigned to 4 generic professions: healthcare workers, students, service holders and unemployed. Each profession has tasks represented as T.

### Evaluation

It uses its prediction data and the real data to make a plot to compare them in the time series

### Results

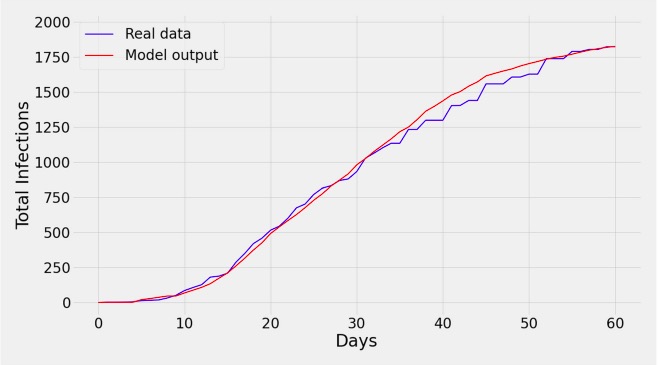


Figure 7 Agent-based model prediction of covid-19 （infected population over time）

It is obvious that the predicted data and the real data is close to each other

### Gaps

The overall structure and the model and the data in use is concise and coherent and creative. However, in the result part, it only shows 60 days of data. The length of the daytime is not enough.

## Game theory by Kabir and Tanimoto[13]

### Data

As the author wrote, the project has no additional data

### Method

It uses the behavioural model which means each person in this model will have to choose 2 strategy: comply with stay-at-home or resist doing that. The stay-at-home order has some economic cost which is the classic attribute of game theory. Thus, by summing up the population, each strategy of the person can be calculated. Also, the SEIR model is also used to calculate the quarantined and infected population. It is then so-called SEQIHR(susceptible-exposed-quarantined-infected-hospitalized-recovered) model.

### Evaluation

Since there is no real time data, there is no evaluation methods involved, all it has is showing the result.

### Results

The result is about the predicted total infected population and the time series.

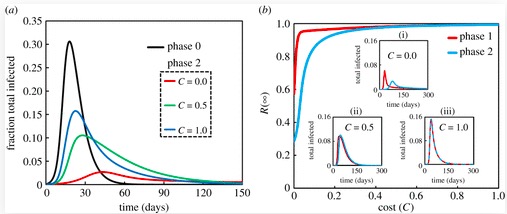


Figure 8 Game theory simulating of COVID-19 infected

### Gaps

It is just like the genetic algorithm of Oxford, there is no comparison of the real time data. However, there is a good point of this model. In the complex parameters of this model, it mentions the paid-off of each strategy, the economic factor is included which is different from Oxford dissertation.

## Summary of these research

In summary, there are 2 different prediction directions: the accurate ones which can be used to predict the new infected population each day and the other one which just shows the tendency of the new infected population. In order words, one type is qualitative and another one is quantitative. The policy factors are mentioned on some of the research papers especially the Oxford one. However, the Oxford one does not pay attention to the economy costs of the policies.

In order to do the prediction, the dataset has to contains the policy data. According to the previous review, the Agent-based model is quite comprehensive thus making another agent-based model is not meaningful. The game theory model only pays attention to the personal paid-off of the strategy but the policy part is implicit: the 2 strategies of each person, compliance and non-compliance can be influenced by government. However, each person has their own values and decision, so their strategies can be influenced other factors. But the economic factors can be taken into consideration.

The SIR model doesn’t pay attention to policies, and it is so complicated.

The machine learning policy is coherent with evaluation methods and the testing and training models there with different ML and deep learning algorithms in python. However, it still does not have any policy factors.

The Oxford genetic algorithm pays attention mainly to policy outcomes and its infection population prediction is deviated from real time data. Also, it is better to invoke economic factors to this algorithm to make it more validate.

# Research Methodology

According to the literature review, there will be 2 different methods, both qualitative research and quantitative researches will be involved.

In order to judge how much the policy factor will influence the prediction of COVID-19, both tendency and accuracy of the prediction model should be inspected.

## Qualitative research

Qualitative research is to predict the tendency of the COVID-19 pandemic to see if the model can explain the development of COVID-19.

### Data

Because it is qualitative research, the main focus will be checking the tendency of the simulating development corresponds to the tendency of real data. Therefore, no additional data will be used or processed.

### Simulating algorithm

The algorithm used is adapted from genetic algorithm (GA). The brief introduction of GA can be found in 1.1.2.3. Here, the typical GA is adapted according to the requirements to pay attention to policy factors and the state/population of the policy. Therefore, there is modification of the adapted algorithm used for qualitative research. The description of the algorithm is given below:

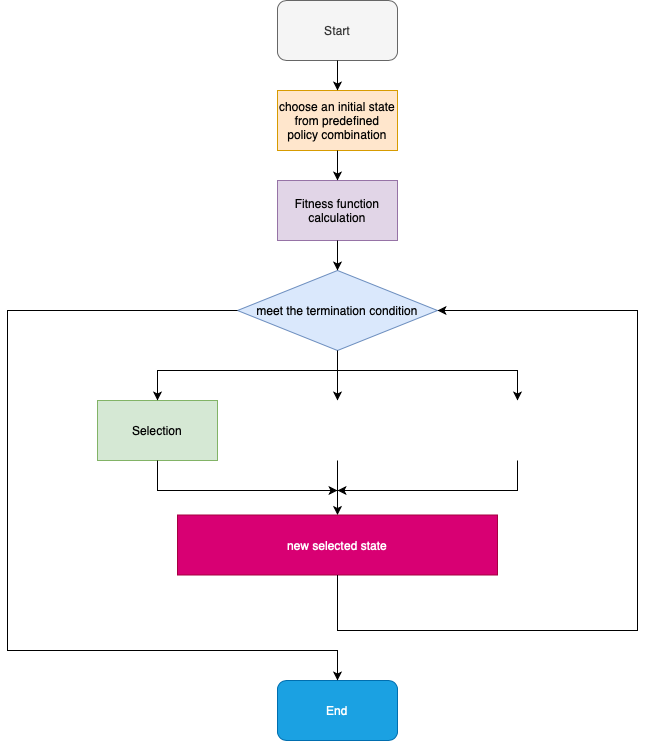


Figure 9 Adapted Genetic Algorithm

#### Differences from typical genetic algorithm

There are several points which this algorithm is different from typical genetic algorithm. Firstly, the typical genetic algorithm will generate the initial state/population randomly while the number of the individual of the initial population will be over 200. However, the adapted algorithm will only have 2 different policies be chosen: whether the government will choose to prevent COVID-19 aggressively with actions like lockdown a city or they will choose to take laissez-faire actions. Secondly, there will be only selection process without mutation or crossover process in it. Even if in typical genetic algorithm, not all of these 3 processes will be set all the time, there are always genetic algorithm with only 2 of the processes.

#### Parameters of the adapted genetic algorithm

It is suitable to use GA since it can describe several objects and their inner attributes. Also, it can simulate the transformation of the status of each object, this will be expressed in detail in the following sections. In addition, GA is much comprehensive and simple than other simulation algorithms such as SIR and Agent-based algorithm.

There are several significant parameters and functions of this algorithm.

Table 2 Simulating adapted GA parameters

|  |  |
| --- | --- |
| Attributes/Function | Explanation of the attribute/function |
| Daily new cases infected population | The new population get infected each day |
| The policy | The policy chosen by the government on that day |
| Economy baseline | Each policy will bring its economic cost, government will balance the cost and infection |
| Policy change | The current policy will change according to different factors, this is represented as fitness function |
| Isolated population | Each day some people will be isolated |

Simple pseudocodes:

1. Choose the initial policy: aggressive policies or laissez-faire polices.
2. Evaluate the fitness of the policy due to the economic costs and the daily new cases
3. Repeat on this process until termination (the given period, the ending condition met)

### Testing method

The testing method will use the tendency graph of adapted GA and the real data tendency graph. If the tendency of the real data tendency and result tendency graph are similar or the result can be used to explain the real data tendency, the reliability of the GA can be promised so that the influence of policy elements can be proved in a certain extent.

#### Pattern distance and piecewise linear representation algorithm

In order to judge the similarity of the tendency of 2 curves, the pattern distance ought to be introduced. Pattern distance is to calculate and judge if the shape of one curve is similar to another. According to WangDa and RongGang[14], if 2 curves have similar shape, their pattern distance will be small. For example, in figure 10, the curve y1 and y3 is similar since it looks like y1 move down 2 then get y3. However, y2 is not similar to y1 or y3.

图表, 折线图

描述已自动生成

Figure 10 pattern distance 3 curves

The calculating method of pattern distance is given below.

形状

中度可信度描述已自动生成

Figure 11 pattern distance example

The description takes the figure 11 as example.

Firstly, both patterns of curve S2 and S1 will be recorded.

S1 = {(1,x1),(-1,x2),(-1,x3),(0,x4)}

S2 = {(1,x1),(1,x2),(-1,x3),(-1,x4)}

The 1,-1,0 values is called m value in S1 and S2 are the slope of each curve where 1 represents the slope is positive so the tendency of curve is increase and -1 represents minus slope while 0 means the slope is 0.

After getting both S1 and S2, then the pattern distance D = . Where is the m value of each curve.

Then the calculation will be 0+2+0+1 = 3. Therefore the pattern distance of curve S1 and S2 is 3.

### Evaluation

This qualitative research account for significant economic issues. Since policy making will influence the pandemic, the cost of the pandemic is a significant economic issues for all the nation governments. It is helpful to do this qualitative research. While it is can not offer accurate results of the infected population each day, it can be used to predict the development of virus at the beginning of the outbreak of pandemic. Since at that time, data is not that much so looking at tendency is significant.

## Quantitative research

### data

#### data of the policy table

The data of the policy table is mainly from Coronanet Project [15]. However, there are so many columns (attributes) in this table and many of them are not related to the research, the detail data can be seen in Appendix. Here, the columns needed are given below.

Table 3 Needed columns of Coronanet Project

|  |  |
| --- | --- |
| Attribute | Meaning of the attribute |
| type | The name of the policy |
| |  | | --- | | date\_start | | The start date of the policy |
| |  | | --- | | date\_end | | The end date of the policy |

By looking at the type attribute, all the policies are given below.

Table 4 Policies listed in Coronanet Project

|  |
| --- |
| New Task Force, Bureau or Administrative Configuration |
| Anti-Disinformation Measures |
| Closure and Regulation of Schools |
| COVID-19 Vaccines |
| Curfew |
| Declaration of Emergency |
| External Border Restrictions |
| Health Monitoring |
| Health Resources |
| Health Testing |
| Hygiene |
| Internal Border Restrictions |
| Lockdown |
| Other Policy Not Listed Above |
| Public Awareness Measures |
| Quarantine |
| Restriction and Regulation of Businesses |
| Restriction and Regulation of Government Services |
| Restrictions of Mass Gatherings |
| Social Distancing |

#### daily cases data

The data is mainly from Our World in Data [16]. However, there are also many columns which should be eliminated in order to keep those related to the research.

Table 5 Attributes of daily cases from Our World in Data

|  |  |
| --- | --- |
| Attribute | Meaning of the attribute |
| date | Current date |
| new\_cases | New cases of infection on the day |
| total\_deaths | Total death population |
| new\_deaths | New death population |
| reproduction\_rate | R0 value of the virus |

### data pre-processing

The 2 tables of policy and cases should be combined in order to build a new table which has both policy influence and daily new cases in it. By taking this measure, with the same date, the current cases and the current policy combinations will be given as the input of the machine learning model. Below is the specific combining process.

1. All the concrete policies will be extracted from the policy table as columns to add into the cases.
2. Using the start date and end date of each policy to set the corresponding policies columns of that date into 1, it means on this date, government took these policies.

### Brief introduction of ML algorithm being used

#### Why Regression

Firstly, the target of this research is using many columns such as daily dead or taking a specific policy or not to predict just one column: the daily new case today.

Also, the label of the predicted column is given. In other words, the value of today’s new daily cases is given. Thus, the model has to be supervised learning instead of un-supervised learning since the value is predefined. Also, the value of new cases each day is not a discrete value, it’s continuous value. Therefore, it should not be a classification model but needs to be a regression model.

#### Regression

Regression is actually a type of question which has a form of Regression. Regression algorithm in machine learning is the combination of several different specific regression model such as linear regression.

This research will choose linear regression. There are different regression such as linear regression, Lasso regression and ridge regression.

Since the research question is not to explore which regression algorithm is most accurate and good for predicting, it is actually testing if the policy parameter has influence on COVID-19 development.

Thus, normal regression model and ridge regression are used and the result can be seen in Chapter 4.

#### Linear Regression

In linear Regression, there will be both independent and dependent variables. Independent variable is represented as (i =1,2,3) while dependent variable is represented as [17].

The function of the relationship between xi and y is

where i = 1,…,n.

While represents the random error.

By feeding the data into the model, both Y and Xi will be given so that the coefficient like will be adjust to get Y to be more close to actual . is the original value while Y is predicted value.

In order to train model efficiently, it uses RSS to represents the differences between Y and . RSS = . The formulation can be transformed into

By solving the equation, then .

In order to get , has to be non-singular matrix. However, this requirement cannot always be satisfied. For example, if the tasks has so many columns which is even more than the row number, then will not be a non-singular matrix, then ‘s determinant will be close to 0, so there will be more than one satisfy this equation.

#### Ridge Regression

In order to solve the problem mentioned in the previous one, then Ridge regression is involved, by adding a into the , then the can be gotten. In order words, Ridge regression is a variant of the ordinary linear regression

#### The brief introduction of sklearn library

Sklearn is a python library built for machine learning algorithms. It provides the template for Ridge regression and ordinary linear regression.

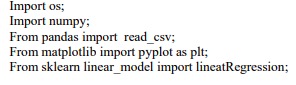


Figure 12 python code importing sklearn library

In order to use them, the importing part should be included.

Detailed information for sklearn for these 2 regression can be found in the official webpage [18].

The regression approaches are trained, tested and used for prediction on real data using the mentioned column.

### Testing Method

There are many evaluation values of these 2 regression algorithms such as R2 and RMSE. These values should be used to evaluate the result of the algorithm and be introduced below. Also, these values of the results will be compared with results in past researches.

#### R2

R2 = ESS/TSS.

Where ESS = and TSS =

If R2 is close to 1, then the prediction is good if R2 is close to 0 [19].

#### RMSE

RMSE =

According to Davide, Matthijs and Giuseppe [20], if RMSE is close to 0, then the model is good, if it gets to infinity, then it is bad.

The real dataset will be separated into training set and testing set. The root\_mean\_squared\_error and R2 will be introduced to judge the accuracy of the model. Also, the plots will be generated as a result.

The python code can be found in Appendix.

### Evaluation

This model can be used for government accurate forecasting of future infected population. Similar to the evaluation of qualitative research, quantitative research is also significant to national economy.

# Results

Qualitative research

The simulating results are shown below.

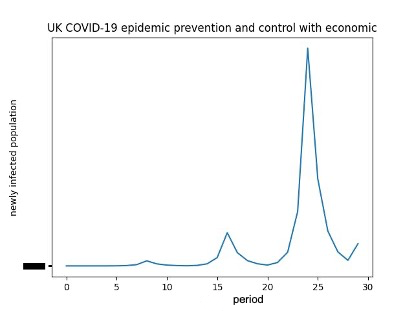


Figure 13 Adapted GA simulating UK result

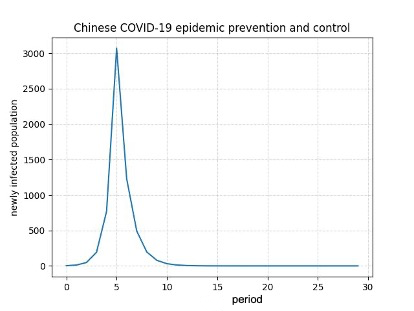


Figure 14 Adapted GA simulating China result

Below are the real data.

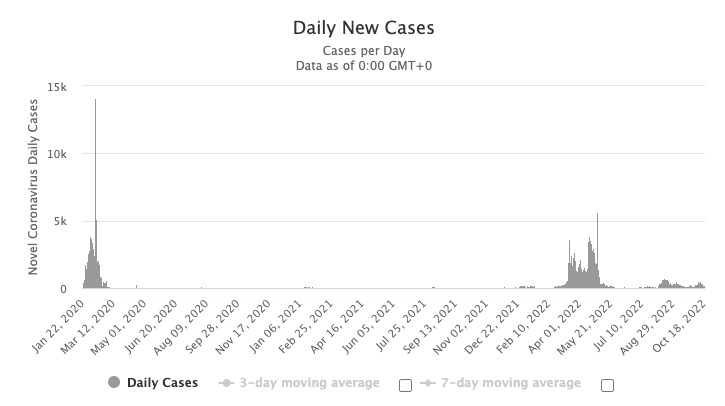


Figure 15 China real data（infected population over time）

UK

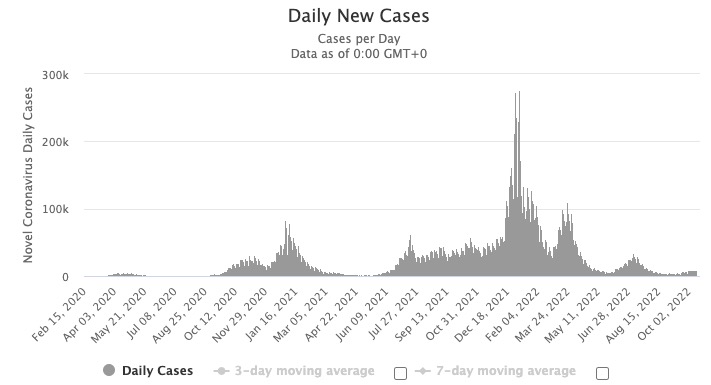


Figure 16 UK real data（infected population over time）

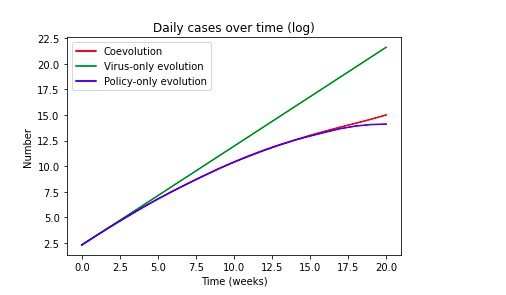


Figure 17 The tendency graph of the Oxford research

By calculating the pattern distance of both Chinese cases and UK cases, the results is given below.

|  |  |  |
| --- | --- | --- |
|  | Real data vs. Oxford result | Real data vs. adapted GA |
| Pattern distance | 5430852 | 151585 |

Table 6 UK pattern distance

|  |  |  |
| --- | --- | --- |
|  | Real data vs. Oxford result | Real data vs. adapted GA |
| Pattern distance | 5432583 | 9517 |

Table 7 China pattern distance

Quantatitive research

Both the training and testing data are shown below.

Table 8 Chinese Linear Regression results

|  |  |
| --- | --- |
| Root\_mean\_squared\_error | 2.6491336683344714e-21 |
| r2\_score | 1.0 |

Table 9 Chinese Ridge Regression result

|  |  |
| --- | --- |
| Root\_mean\_squared\_error | 1582.8848411985034 |
| r2\_score | 0. 9998761890261062 |

Table 10 UK Linear Regression result

|  |  |
| --- | --- |
| Root\_mean\_squared\_error | 3.913360982361367e-10 |
| r2\_score | 1.0 |

Table 11 UK Ridge Regression result

|  |  |
| --- | --- |
| Root\_mean\_squared\_error | 39.78548530806811 |
| r2\_score | 0.999876189 |

Also, the graph of the real time data and the prediction data is given below.

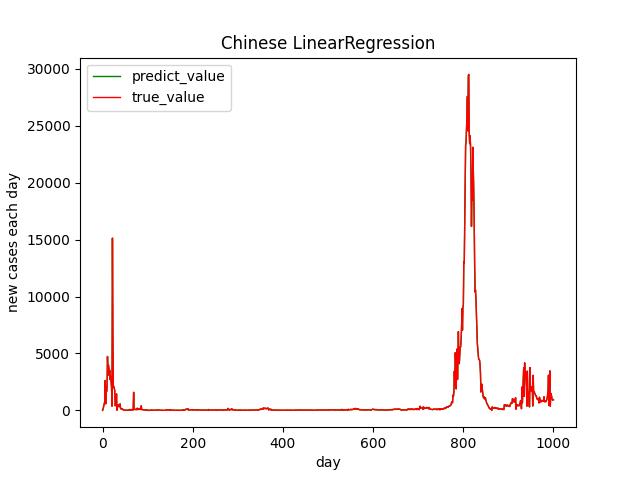


Figure 18 Chinese Linear Regression over time

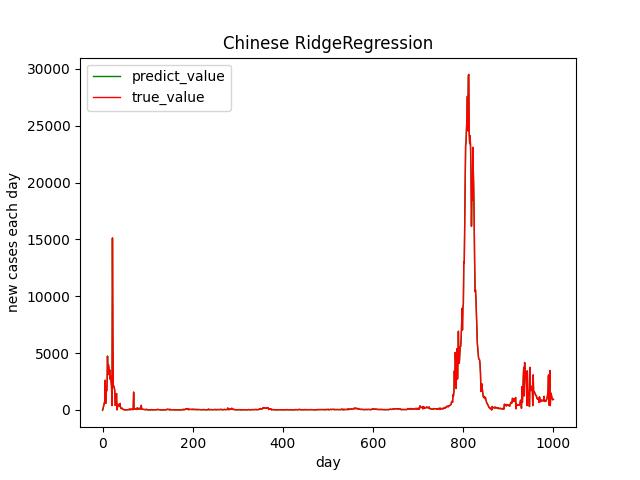


Figure 19 Chinese Ridge Regression over time

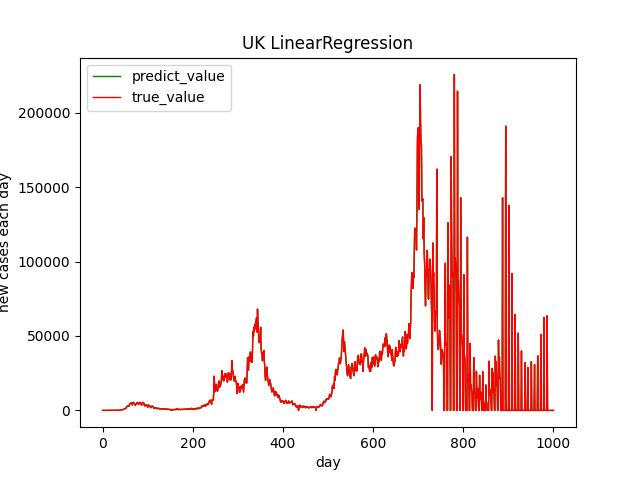


Figure 20 UK Linear Regression over time

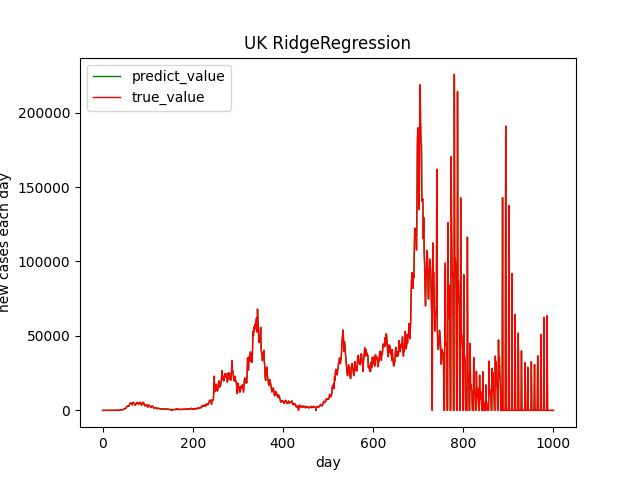


Figure 21 UK Ridge Regression over time

# Analysis

Qualitative approach

### Critical analysis on methodology

The entire components of qualitative approach are given below. Firstly, there is no data for qualitative approaches. The policy parameters are involved to build a GA model which is adapted from Oxford GA. The goal of this qualitative approach is to see if the model will present results which corresponds to the real tendency of daily new cases.

Since there is no data given, the reliability is not high enough. The tendency given in daily new cases is drawn from government website. Therefore, its reliability is enough more than non-official institution. However, the best quality data is tracking all the infected and death data each day which cannot be entirely fulfilled. Because of the testing efficiency and its costs, governments may not completely test every person in the nation.

The genetic algorithm is useful to simulate the transformation of a specific target so when the policy parameter involved, the GA will simulate its transformation.

Since other algorithms are applied by other researches on this topic and get reliable results, there is no need for deeper application of these algorithms such as multi-agent algorithm.

### Critical analysis on experimental design

The experimental design is used is Before-and-after without control design.

Here, the test group is the prediction models, datasets and models’ tendency given by other researchers.

The treatment is invoking the policy factors. The level of the phenomenon before treatment is the tendency of the researchers’ models and the level of phenomenon after treatment is the tendency of the models in this dissertation.

### Critical analysis on initial results

The tendency of China COVID-19 development of simulated result and real data is similar: from the beginning, the new daily cases start from 0 with a exponential growth and then down immediately. The curve looks like a single wave.

The tendency of UK COVID-19 development of simulated result and real data is also relatively similar: with many waves together while the peak value of some waves are lower than other waves.

However, the data is extracted from government website. There is possibility that the real infected people each day do not do any COVID-19 nucleic acid test so their data is not recorded.

Therefore, here an assumption is made: if the number of real infected people is close to the data extracted from government website, then the tendency of the simulated is close to the real data.

However, it cannot be inferred that this is caused by adding policy factors into the adapted genetic algorithm. It is much better than the simulation of Oxford model, but there may be other qualitative research which have a better description of tendency but without adding policy factors.

Quantitative approaches

### Critical analysis on methodology

The entire components of quantitative approach are given below. Firstly, the data is extracted from government website. The policy parameters are involved to build regression models. The goal of this quantitative approach is to see if the model will present results which have a better accuracy than previous researchers’ models.

### Critical analysis on experimental design

It is just as the Qualitative approach except that the tendency should be accuracy.

### Critical analysis on initial results

Firstly, compared with previous results, the RMSE of linear regression is greatly improved from 15647.64 to 2.6491336683344714e-21. Apart from policy factors, both the linear regression models of this dissertation and the previously researched are invoking python’s sklearn library. The data are both from government website. Thus, the only differences between these 2 datasets are the attained date. The data here is 2022 while the previous research was ended in 2020.

However, to make a more precise comparison, the additional experiment has been made. This experiment uses the exact same setting of this research model: python sklearn with policy factors involved but uses the data no later than 2020 just like the previously researched one. However, a better RMSE value is still attained.

Therefore, adding the policy factors into the linear regression model will increase the accuracy.

Problems during implementation

At first, it is hard to choose which algorithm is good for implementation. There are several requirements for the algorithms to be used. Firstly, the algorithm must be not complex. Secondly, the policy factors can be integrated into the algorithm. Third, the algorithm haven’t been deeply applied by other researchers.

In order to fulfill the requirements listed above, much work has been done such as literature review and code writing. At last, the GA and regression algorithm is choosen as the proper one to be used.

This problem is the most significant part of this dissertation since it’s the core of experiment. By reviewing different algorithms and writing codes to see the algorithms’ effect, this problem is finally addressed.

# Discussion

It can partly answer the research question. Adding the policy attributes can increase the accuracy of quantitative model especially linear regression. Adding policy attributes can increase the effect of simulating algorithm: genetic algorithm.

With the assumption made: if the daily infected people are all recorded, then the qualitative can be more precise. And other quantitative can add policy attributes in order to see whether the results are better or not.

# Conclusions

In conclusion, the results means that , it is significant since it offers a new angle for the future work: add the policy attributes into the model is important.

Critical reflection

### Reflection on design

The design is partly as expected. Due to the limitation of tools and time, the design is constructed to get a limited result which can partly answer the research question.

For example, in the quantitative research, regression algorithms are involved and policy factors are included and the results shows better accuracy than the previous research. However, there are so many other algorithms to be tested.

### Alternative explanations for the findings

#### Qualitative research result

The result shows the adapted genetic algorithm are better than Oxford genetic algorithm when simulating the tendency of UK and China in 30 periods. However, it can only propose that Oxford genetic algorithm is not good as adapted genetic algorithm. Maybe there are other genetic algorithms which does not involve policy factors but are much better than the adapted genetic algorithm taken in this dissertation.

#### Quantitative research result

The result shows that with the policy factor given, the newly built regression algorithm is better than the previously researched regression algorithm. However, it can only promise that until the time of the dataset extracted, the policy factor regression algorithm is better than the previous one. There is possibility that in the future, more daily data will be collected and the previous algorithm is better than the policy factor algorithm.

### Scope

The scope of the study is described in Chapter 1.3. The boundary of the study is to explore the influence of policy factors in prediction models. However, after the research, the scope needs to be narrower. The scope ought to be exploring the influence of policy factors in regression models. Therefore, the conclusions given by the results will be more concise and clear: policy factors will influence the accuracy of linear regression models from 15647.64 to 2.6491336683344714e-21.

Personal development

During this dissertation project, I have learned a lot. It’s quite helpful to follow the advice given by supervisors since they are experienced researchers.

Also, critical thinking needs to be made not only when doing literature review to judge others’ work but also when writing chapters of this dissertation.

Getting familiar with data processing and software programming is also important, learning such tools can increase the efficiency of experimental implementation.

Last but not least, communicating and demonstrating skills is quite important when showing own points and express ideas to others.

Future work

According to the results and analysis, other qualitative researches which includes policy factors ought to be made to test if policy factors will influence the prediction of COVID-19.

REFRENCES CITED

[1] WHO. “WHO COVID-19 dashboard.” <https://covid19.who.int/> (accessed: Oct. 27 2022).

[2] E. Elibol, “Otolaryngological symptoms in COVID-19”. *Eur Arch Otorhinolaryngol*., vol. 278, pp. 1233–1236, April. 2021. [Online]. Available: https://doi.org/10.1007/s00405-020-06319-7

[3] V. Saladino, D. Algeri and V. Auriemma, “The psychological and social impact of Covid-19: new perspectives of well-being.” *Frontiers in psychology*, vol. 11, pp 2550, Oct, 2020, doi: 10.3389/fpsyg.2020.577684

[4] D. Smith, L. Moore. “The SIR model for spread of disease-the differential equation model”. *Convergence.* <https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model> (accessed: Oct. 27 2023).

[5] M. Shamil, F. Farheen, N. Ibtehaz, I. M. Khan and M. S. Rahman, “An agent-based modeling of COVID-19: validation, analysis, and recommendations.” *Cognitive Computation*, vol. 12, pp.1-12, Feb, 2021, doi: 10.1007/s12559-020-09801-w

[6] Mathew, T. V. “Genetic algorithm.” Report submitted at IIT Bombay, Jan. 12, 2012. [Online]. Available: <https://datajobs.com/data-science-repo/Genetic-Algorithm-Guide-%5BTom-Mathew%5D.pdf>

[7] International Monetary Fund. “policy responses to COVID-19.” <https://www.imf.org/en/Topics/imf-and-covid19/Policy-Responses-to-COVID-19#C> (accessed: Oct.27 2022).

[8] Worldometer. “COVID-19 CORONAVIRUS PANDEMIC.” <https://www.worldometers.info/coronavirus/#countries> (accessed: Oct.27 2022).

[9] A. Vie, “Modelling SARS-CoV-2 coevolution with genetic algorithms.” *arXiv*, Aug, 2021. [Online]. Available: <https://arxiv.org/abs/2102.12365>

[10] N. S. Punn, Sonbhadra, S.K. and S. Agarwal, “COVID-19 epidemic analysis using machine learning and deep learning algorithms.”  *MedRxiv*, Apr, 2020. [Online]. Available: <https://europepmc.org/article/ppr/ppr150219>

[11] Johns Hopkins Whiting school of Engineering. “COVID-19.” <https://github.com/CSSEGISandData/COVID-19> (accessed: Oct.29 2022).

[12] L. Wang, Y. Zhou, J. He, B. Zhu, F. Wang, L. Tang, M. Kleinsasser, D. Barker, M. C. Eisenberg and P. X. Song, “An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China.” *Journal of Data Science*. Nov, 2021. Accessed: Oct, 27, 2022, doi: [https://doi.org/10.1101/2020.02.29 20029421](https://doi.org/10.1101/2020.02.29%2020029421). [Online]. Available: <https://www.medrxiv.org/content/10.1101/2020.02.29.20029421v1>

[13] K. A. Kabir and J. Tanimoto,. “Evolutionary game theory modelling to represent the behavioural dynamics of economic shutdowns and shield immunity in the COVID-19 pandemic.” *Royal Society open science*. Sep, 2020. Accessed: Oct, 27, 2022, doi: <https://doi.org/10.1098/rsos.201095>. [Online]. Available: <https://royalsocietypublishing.org/doi/10.1098/rsos.201095>

[14] D. WangDa & RongGang,. “Pattern distance of time series.” *WIT Transactions on Information and Communication Technologies. 2003*. [Online]. Available: https://scholar.google.com/scholar?hl=zh-CN&as\_sdt=0%2C5&q=pattern+distance+of+time+series&oq=#d=gs\_qabs&t=1669529829254&u=%23p%3DM\_Mz6OOXMMIJ

[15] Coronanet Project. “Coronanet Research Project.” <https://www.coronanet-project.org/> (accessed: Oct.26 2022).

[16] Our World in Data “Coronavirus Pandemic (COVID-19).” <https://ourworldindata.org/coronavirus> (accessed: Nov.11 2022)

[17] A. E. Hoerl, & R. W. Kennard. “Ridge regression: Biased estimation for nonorthogonal problems”. *Technometrics*. Aug, 1970. Accessed: Oct, 29, 2022, doi: 10.1080/00401706.1970.10488634. [Online]. Available: <https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634?tab=permissions&scroll=top>

[18] scikit-learn “sklearn.linear\_model.Ridge.” <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html#sklearn.linear_model.Ridge> (accessed: Nov. 8 2022)

[19] S. Rong, & Z. Bao-Wen (2018). “The research of regression model in machine learning field.” *MATEC Web*., 2018, Vol. 176. Available: https://www.researchgate.net/publication/326121964\_The\_research\_of\_regression\_model\_in\_machine\_learning\_field

[20] Chicco, Davide, Matthijs J. Warrens, and Giuseppe Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation." PeerJ Computer Science 7 (2021): e623.

APPENDICES

###### Code

Python code of linear regression and ridge regression model

import matplotlib.pyplot as plt  
import numpy as np  
from sklearn.model\_selection import train\_test\_split  
from sklearn.preprocessing import StandardScaler  
from sklearn.linear\_model import LinearRegression,Ridge  
import pandas as pd  
from sklearn.metrics import mean\_squared\_error,r2\_score  
  
def method1():  
 UKcase = "new\_dataframe-UK.csv"  
 upd = pd.read\_csv(UKcase)  
 upd = upd.loc[0:1001]  
 upd.fillna(0,inplace=True)  
  
  
 data = upd.drop('Unnamed: 0',axis=1)  
 data = data.drop('iso\_code',axis=1)  
 data = data.drop('continent',axis=1)  
 data = data.drop('location',axis=1)  
 data = data.drop('date',axis=1)  
 data = data.drop('total\_cases',axis=1)  
 data = data.drop('new\_cases\_smoothed',axis=1)  
 data = data.drop('stringency\_index',axis=1)  
  
  
 target = upd[['new\_cases']]  
  
 # print(data.columns)  
  
 x\_train,x\_test,y\_train,y\_test = train\_test\_split(data,target,random\_state=30)  
  
 transfer = StandardScaler()  
 x\_train = transfer.fit\_transform(x\_train)  
 x\_test = transfer.transform(x\_test)  
  
  
 estimator = LinearRegression()  
  
 lr = estimator.fit(x\_train,y\_train)  
  
 print("权重系数为: \n", estimator.coef\_)  
 print("偏置为: \n",estimator.intercept\_)  
  
  
 y\_predict = lr.predict(x\_test)  
 # print("预测结果为: \n", y\_predict)  
 error = mean\_squared\_error(y\_test,y\_predict)  
  
 print("正规方程-均方误差为:\n ",error)  
 print("RMSE = ", np.sqrt(error))  
  
 r2 = r2\_score(y\_test,y\_predict)  
 print("r2 = ",r2)  
  
 t = np.arange(len(x\_test))  
  
 '''  
 下面就是新的制图方式，将原本的true的data和用于训练以及预测的data一并绘入  
 true data = target  
 预测以及用于训练的数据  
 y\_test有每个index对应两个值，一个是代表的天数的，一个是代表new\_cases  
 而y\_train和y\_test结构相同  
  
 而y\_predict和y\_test实际上一一对应，但是遗憾的是,y\_predict的代表天数的  
 数据是错误的，需要调整的和y\_test一致才可以  
 '''  
  
 date = y\_test.index  
  
 y\_predict = pd.DataFrame(y\_predict)  
  
 y\_predict.index = date  
  
 '''  
 把y\_predict和y\_train融合起来  
 '''  
 y\_predict.rename(columns={0: 'new\_cases'}, inplace=True)  
  
 y\_prediction = y\_predict.append(y\_train)  
  
 y\_prediction = adjust\_y\_pred(y\_prediction)  
  
 # plt.plot(t,y\_test,'r',linewidth=1,label='y\_test')  
 # plt.plot(t,y\_predict,'g',linewidth=1,label='y\_train')  
 #  
 #  
 date = np.arange(len(target))  
 plt.plot(date, y\_prediction, 'g', linewidth=1, label='predict\_value')  
 plt.plot(date, target, 'r', linewidth=1, label='true\_value')  
 plt.xlabel("day")  
 plt.ylabel("new cases each day")  
 plt.legend()  
 plt.title("UK LinearRegression")  
 plt.savefig("UK LinearRegression.jpg")  
 plt.show()  
  
def method2():  
 UKcase = "new\_dataframe-UK.csv"  
 upd = pd.read\_csv(UKcase)  
 upd = upd.loc[0:1001]  
 upd.fillna(0,inplace=True)  
  
 data = upd.drop('Unnamed: 0',axis=1)  
 data = data.drop('iso\_code',axis=1)  
 data = data.drop('continent',axis=1)  
 data = data.drop('location',axis=1)  
 data = data.drop('date',axis=1)  
 data = data.drop('total\_cases',axis=1)  
 data = data.drop('new\_cases\_smoothed',axis=1)  
 data = data.drop('stringency\_index',axis=1)  
  
  
 target = upd[['new\_cases']]  
  
 # print(data.columns)  
  
 x\_train,x\_test,y\_train,y\_test = train\_test\_split(data,target,random\_state=30)  
  
 transfer = StandardScaler()  
 x\_train = transfer.fit\_transform(x\_train)  
 x\_test = transfer.transform(x\_test)  
  
  
 estimator = Ridge()  
  
 lr = estimator.fit(x\_train,y\_train)  
  
 print("权重系数为: \n", estimator.coef\_)  
 print("偏置为: \n",estimator.intercept\_)  
  
  
 y\_predict = lr.predict(x\_test)  
 # print("预测结果为: \n", y\_predict)  
 error = mean\_squared\_error(y\_test,y\_predict)  
  
 print("正规方程-均方误差为:\n ",error)  
  
 print("RMSE = ", np.sqrt(error))  
  
 r2 = r2\_score(y\_test,y\_predict)  
 print("r2 = ",r2)  
  
 t = np.arange(len(x\_test))  
  
 '''  
 下面就是新的制图方式，将原本的true的data和用于训练以及预测的data一并绘入  
 true data = target  
 预测以及用于训练的数据  
 y\_test有每个index对应两个值，一个是代表的天数的，一个是代表new\_cases  
 而y\_train和y\_test结构相同  
  
 而y\_predict和y\_test实际上一一对应，但是遗憾的是,y\_predict的代表天数的  
 数据是错误的，需要调整的和y\_test一致才可以  
 '''  
  
 date = y\_test.index  
  
 y\_predict = pd.DataFrame(y\_predict)  
  
 y\_predict.index = date  
  
 '''  
 把y\_predict和y\_train融合起来  
 '''  
 y\_predict.rename(columns={0: 'new\_cases'}, inplace=True)  
  
 y\_prediction = y\_predict.append(y\_train)  
  
 y\_prediction = adjust\_y\_pred(y\_prediction)  
  
 # plt.plot(t,y\_test,'r',linewidth=1,label='y\_test')  
 # plt.plot(t,y\_predict,'g',linewidth=1,label='y\_train')  
 #  
 #  
 date = np.arange(len(target))  
 plt.plot(date, y\_prediction, 'g', linewidth=1, label='predict\_value')  
 plt.plot(date, target, 'r', linewidth=1, label='true\_value')  
 plt.xlabel("day")  
 plt.ylabel("new cases each day")  
 plt.legend()  
 plt.title("UK RidgeRegression")  
 plt.savefig("UK RidgeRegression.jpg")  
 plt.show()  
  
def adjust\_y\_pred(y\_prediction):  
 *'''  
 将y\_prediction的顺序按照日期填入  
 '''* res = y\_prediction.sort\_index(axis=0,inplace=False)  
 '''  
 把res第一列的数据全部合到第二列中去  
 '''  
  
  
 return res  
  
if \_\_name\_\_ =="\_\_main\_\_":  
 method1()  
 print("#########################################")  
 method2()