Comp0031 Group Report:

Haocheng Lin, Coco Liu, Mohseen Hussain

Introduction

SARS-Cov2 coronavirus has originated from Wuhan, China in 2019. Since then, the coronavirus has spread globally where it mutated into new variants. Different epidemic models have been implemented to simulate the possible projection of the pandemic to evaluate whether the current measures are effective. The models use different criteria: simulation time, infection, and recovery probability. Several papers have already conducted studies on the differences between the epidemic modelling and the real-life cases.

This paper applies the classical SIR model to predict the number of cases and compare with the synthetic data. The comparison also helps to fitting data for projection at specific dates. In addition, external factors are considered to ensure that the simulations are made as realistic as possible for a more accurate comparison between the synthetic and simulated data.

Design

Objectives

Goal:

1. Perform a fit during an epidemic wave to predict the number of infections and determine the prediction error.
2. Identify the problems that are related to how the population groups are simulated throughout the epidemic simulations.
3. Produce several different visualizations of the predictions and the errors between the predicted and the synthetic data.
4. Apply noise to the synthetic data to make the data more authentic by including the noise as an abstraction of the external factors.

Detailed Goals

1. Problems related to the data collection ☑ 🗎
2. Scenario: perfect data and applying without noise ☑
3. Complicate the model ☑
4. Description: pseudocode for the program. ☑
5. Graph 1: SIR graph ☑ 🗎
6. Graph 2: cumulative infection data and fit ☑ 🗎
7. Logistic fit uses all data ☑ 🗎
8. Apply fittings to different dates☑ 🗎
9. Graph 3: fitting from different dates. ☑ 🗎
10. Add noise to the data points. ☑
11. Consider data points in the networks. ☒

12. Produce an average error plot. ☒
13. Compare the error with the epidemic trends. ☒
14. Written documentation for the technical solutions. 🗎

Methodology

We decided to use Python programming language with the NumPy and matplotlib libraries. Python is a common program language with plenty of prior research. Its suitable at visualizing data.

In the project, we aim to create different simulations:

- A simulation for the infected, susceptible, and recovered groups.
- Perform logistic fitting on the cumulative infected population.
- Measure the error from the sample simulation.
- Introduce random noise to the simulation to add the external factors into consideration.

The simulations will use the SIR Model: splitting the population into 3 groups (the susceptible, infected, and recovered). There are 2 transitions in the model: a susceptible individual becoming infected and an infected individual recovering.

The simulation shows how the different population groups change over time. Initially, there were an exponential increase in the number of cases. The infected population changes reflect an epidemic wave: the infected population will reach its maximum; the epidemic reaches its peak. As the infected population grows, the recovery rate increases, and more people will recover than infected so the infected population will decrease and eventually the epidemic will end.

Pseudocode

A simulation for the infected, susceptible, and recovered groups.

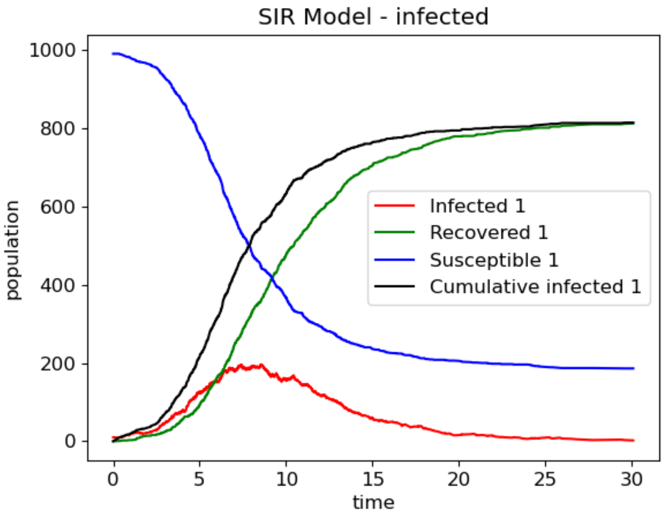
The simulation is a CTMC SIR Model with the initial parameters:

1. β, and γ constants: infection and recovery constant
2. Initial sample population: 1000
3. Simulation time: 30 days (about 4 and a half weeks)

Pseudocode

- Initialized the different population groups: the infected, susceptible, and recovered groups. Initially 10 people are simulated as infected, the rest of the population are susceptible to the epidemic, and 0 recovered individuals.
- While there are infected patients, and the time is within the limit
- Calculate the infection constant a and b
- The infection constants are used to calculate whether a patient has recovered or whether there are newly infected individuals.

- If the randomly generated probability is less than a value, then simulate a new infection otherwise simulate a new recovery
- Update time through the formula: $t[j+1] = t[j] - \frac{ln(u_2)}{a+b}$
- Plot the times with the infected population



A simulation to generate prediction curve alongside the simulation data.

An additional logistic fitting function is applied to fit the simulated data using the formula:

$$result = \frac{a}{1 + b \times e^{-c \times t}}$$

The logistic fit function is applied on the infected data to help with the prediction of the future infection cases, the simulation data can be helpful at giving an insight for the prediction of the future real-life outbreaks.

A simulation to apply fitting to different dates.

The simulation considers different dates as stages in the pandemic. Using each date to simulate the projections after the date helps to give us better insights about the pandemic data at each stage in the pandemic. It is also easier to record the data in this way because it is a more realistic method of simulation: data is collected at specific time intervals – in real-life, it is not possible to continuously update the data. This data collection is more realistic at simulating the pandemic cases.

1. Last day of the simulation, projection:
2. Average it over 100 simulations.
3. Produce a figure for the simulation

A simulation to calculate the average error for measuring the infected patients.

1. A simulation generates the synthetic data to model the real-life data.
2. Logistic fit is performed every 3 days on the synthetic data to generate simulated data.

3. Compare the last day data of the simulated data with the synthetic data (real-life) to calculate the error in the measurement of the infected patients. The last day is considered because the data is cumulative, so the previous patients are considered.
4. The errors are initially stored in a list where the error is the absolute difference between the synthetic data and the predicted data.

Limitations

The SIR model suffers from several pitfalls. In general, the model can only be considered as a theoretical framework to investigate the spread of SARS-CoV-2 within communities in the starting point, since it over-simplified the real-life situation.

The current SIR model is not completely accurate at projecting the different population groups. The model is simple at identifying main population groups and encapsulating the processes down to 4 simple processes: infection, recovery, exposure, and removal. The simplification generalizes the spreading of COVID in 4 variables which meant that it will not be possible to consider each individual factor e.g., the differences between the rural and the urban environment, how the regulations vary by region, and differences in everyone's immune system. The model assumes a normal distribution to determine whether an individual is infected or recovered. In real-life scenarios, the probability of infection is determined by contact between individuals. The probability of recovery is dependent on a mixture of factors: age, immune system, genetic composition, and other factors. The current model assumes that the patient is infected and recovered one at a time. An infected individual can contact multiple others. The current model has a risk of underestimating the infection rate because the rate in which the model changes is different to the real-time update. However, the model is more accurate than the data collected because it collects the data at a more regular interval. To predict the simulated data, logistic fitting is applied to the synthetic data to generate the projection curve for comparison in the graph. However, the logistic fitting assumes a linear relationship between the distinct factors and is not suitable in the situation where there are multiple independent variables, the fitting only considers the factors in terms of 3 variables: a, b, and c which represent the factors to determine the infection and the recovery rate.

The additional measures should include strict lockdown regulations, medical treatment, and the isolation of susceptible individuals to avoid mixing them with non-symptomatic and self-quarantined individuals. Second, we have not taken the human network into account. Concerning a completed social structure and urban spaces' spreading patterns, which include individual biological variation and dynamic social contact networks, they exert impacts on the region of environmentally driven pattern and not only policies. Human lifestyles are inexorably connected to the risk of infectious disease contamination, such as with or without vaccination or social isolation. Those factors would influence our estimation on any assumption on S and R during the epidemic phase, especially in its initial curvature.

To have a more intuitive understanding of modelling SARS-CoV-2, we assume a homogeneous mixing of the infected and susceptible populations and that the total population is constant in time on our SIR model. These assumptions could not be valid in a real-life situation because new cases are springing up around the world, and non-infected people are vulnerable to changes in the neighboring communities at any moment.

In addition, the recovered individuals are no longer susceptible to our assumption, which means they are immunized. However, new research shows that the virus could be reactivated or that already infected people could get infected once again. Although the susceptible population steadily declines towards zero in the classic SIR model, we simplified the calculation due to the time. We will not only adjust the proposal and the variation in the total population but also take the surges in the increase of susceptible population into account at a later stage.

Our SIR model used here is only a simple approach to predict, which is not accurate enough compared with other epidemic models for SARS-CoV-2. To better reflect the epidemic of reality, we need to improve the current model to incorporate more detailed and accurate observed data. For example, the Susceptible-Exposed-Infected-Recovered (SEIR) model includes an additional group of individuals. Denoted by E, the people who are exposed means they are infected but cannot yet transmit the virus. Agent-based Modeling (ABM) method also considered more detailed information, providing useful guidelines for outbreak management and policy development. Rather than assuming a closed population without migration, births, or deaths from causes other than the epidemic, the ABM technique enables the simulation of systems with complicated nonlinear interactions, conditions, and constraints that are difficult to model theoretically.

Future Work

Although the SIR model simplified the infection process, it wasn’t accurate enough to model the real-life processes. We attempted to apply the data using a network format to improve the modelling of contact between individuals or include real-life datasets for future works.

An alternative ISIR model can be applied instead of the current model. The SIR model assumes that the population is homogeneous, which that the nodes share the same linkage in between and the probability of each connection between any two nodes is equal. However, further research shows that nodes have distinct linkages, and the nodes have fewer links between communities than within a cluster. It’s vital if the prediction model could take the social contact networks into account. The novel ISIR model addresses this limitation. The protection effect is existing in the community structure. This phenomenon implies that if a great number of infected individuals do exist, the affected contacts between susceptible and infected individuals would not increase quickly. Therefore, under some typical scenarios, the linear force of infection implemented in our standard SIR model is limited.1 The model divides the total population into seven compartments: susceptible individuals in the free environment, undiagnosed and non-isolated infectious individuals, recovered individuals, death individuals, free Exposed, Confirmed and isolated infectious individuals, and Patients with suspected,2 which allows the high similarity in between prediction values and the reported values.

In the process of simplifying both mathematical modelling and data fitting, the model assumed the invariably fixed transmission rates as constants. However, the transmission rates vary with the epidemiological and socioeconomic status. Demographics, habitat, environment, health, age, gender, and subpopulations are possible factors. Super-infection and cross-immunity exist in a real-life situation. Thus, concluding a general theory to take all the details into account is a challenge. But we must put this issue on the top priority of our research with the advent of new data.

The other encouraging guideline for improving mathematical modelling is to link models with data-driven techniques, especially machine learning. Agent-based Modeling (ABM) is one of the useful approaches to outbreak management and policy development. All objects in ABM can be characterized as computational agents, and an environment can be codified computationally, with agent behaviour, agent-agent, and agent-environment interactions forming the system dynamics and causing the phenomena under investigation to arise. The validating machine learning prediction method allows the simulation of systems with complex nonlinear relationships, conditions and restrictions that is difficult to describe mathematically3, which could potentially help substantial progress on the COVID-19 study and even beyond.

Research Outcomes

Result

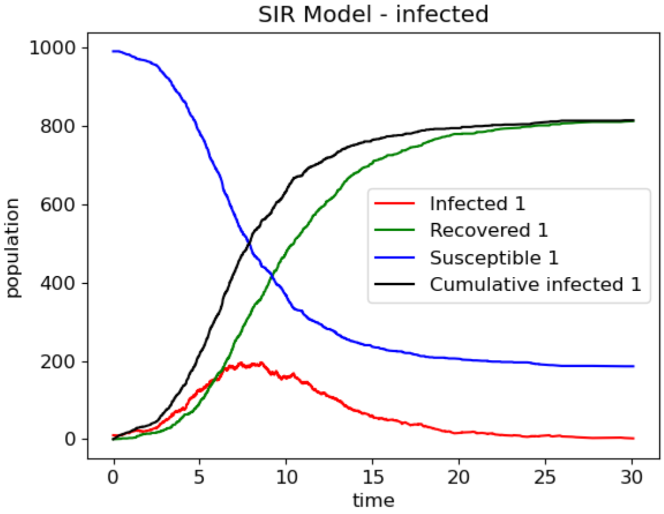


Figure 1. Epidemic Graph showing the infected, recovered, and susceptible populations over a specific period

Figure 1 shows the different population groups in an epidemic: infected, recovered, and susceptible groups. The population groups model the SIR model to help with gaining a better understanding of the epidemic model used in the simulations.

The figure shows that during an epidemic wave, the infected population increases and peaks before decreasing. The susceptible population decreases, and the recovered population increases throughout the epidemic.

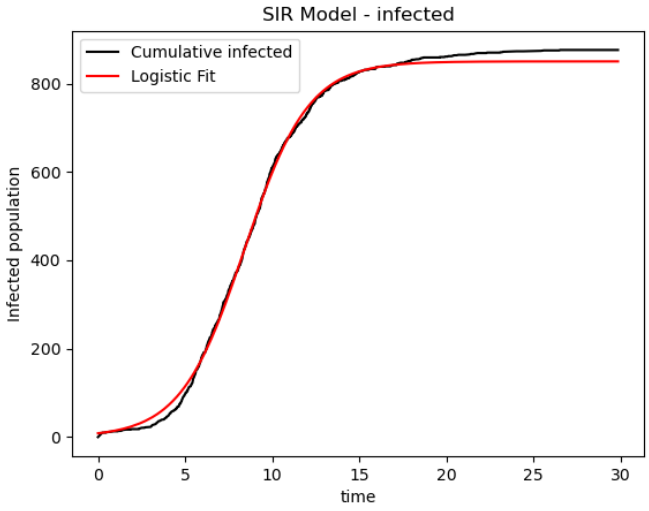


Figure 2. Comparison between the cumulative infected data with the logistic fit data

Figure 2 applies a logistic fit on the infected data to help to model the infected population during an epidemic. The figure applies logistic fitting on the synthetic data using the formula:

$$\frac{a}{1 + b \times e^{-c \times t}} \tag{1}$$

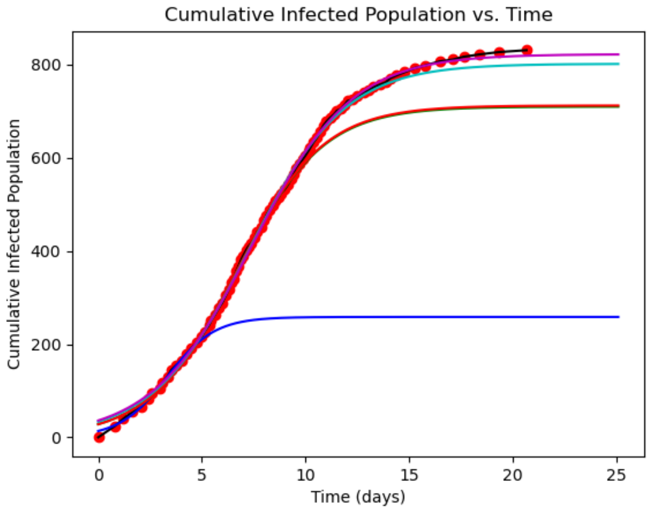


Figure 3. Projection of the number of cases at regular date interval

Figure 3 aims to simulate the real-life data recording to model the multiple stages of logistic fitting. The different logistic fitting happens at different points in time because the projection will be different at each stage in the epidemic. For example, in the first stage, the number of cases is few therefore the first few projections are less accurate than the projections made in the later stages.

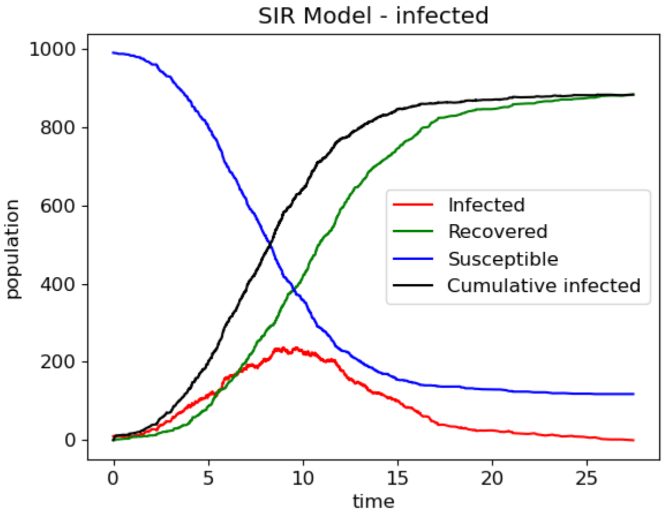


Figure 4. Epidemic simulation including the noise data

Data quality

Synthetic data is applied in the current program. The data is generated by normal distribution randomly. The data quality is not as accurate as the real-life data because encapsulation is applied on the data generation: hiding the complexity of the different real-life factors that may affect whether an individual is infected or recovering. The data quality assumes a homogeneous population and ignores the differences between the diverse groups e.g., age, pre-existing health conditions, and living environment.

Significance

The project goal is to develop an SIR model and a synthetic data structure for comparison to determine the error produced by the model. The project focuses on the infected population group.