**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. , and constants: infection and recovery constant
2. Initial sample population: 1000
3. Simulation time: 30 days

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:
* Plot the times with the infected population

Chart

Description automatically generated

**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:

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’s 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 current SIR model isn’t 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, genetics 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 are 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.

**Research Outcomes**

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.

Problems related to the data collection ✅

Scenario: perfect data and applying without noise ✅

Complicate the model ✅

Description: pseudocode for the program.

Graph 1: SIR graph ✅ 🗎

Graph 2 : cumulative infection data and fit ✅ 🗎

Logistic fit uses all data ✅ 🗎

Apply fittings to different dates✅ 🗎

Graph 3 : fitting from different dates. ✅

Add noise to the data points. ✅ - 3

Consider data points in the networks. ❎ - 4

Produce an average error plot. ❎ - 1

Compare the error with the epidemic trends. ❎ - 2

🗎 : written documentation for the technical solutions.

**Result**

Chart

Description automatically generated

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.

Chart

Description automatically generated

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.

Chart, line chart

Description automatically generated

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

Chart

Description automatically generated

Figure 4. Epidemic simulation including the noise data

Individual report: 8 pages

Group report: 9 pages

Another approach to plot the curve involves using the fitting curve from the start to the end of simulation. The simulation splits the data into different stages, each stage acts as a check point to simulate the data after. At the point, the future projection is applied to simulate the current unknown data and projection is compared with the simulated data to determine the error in the curves.