7CCSMSDV Simulation and Data Visualisation CourseWork

Jinhong Jiang[*]
K21021687

1 Part 1: Analytics

1.1 Exploratory Question

- **Question 1**: *Analyze the development of virus mutations over time. Are there detectable seasonal trends?*

  Within this question, since the number of virus mutations is an aggregate function that operate on time dimensions, i.e. a *time series*. We will look into its stationarity, to check whether there are seasonal trends. By seasonal, here it means time series seasonal.

- **Question 2**: *How do new confirmed cases correlate with the mutations development?*

  In this question, we would look into the correlation between new confirmed case number and the mutations amount. It could be either negative correlation or positive correlation.

- **Question 3**: *Is Covid-19 starting off harming their hosts, but evolving toward a more benign coexistence?*

  One popular theory, endorsed by some experts, is that viruses cause high morality rate, but as evolving, morality rate drops. In this question, we would use our data to prove it.

1.2 Dataset Evaluation

To answer **Question 1**, we need the mutations data over time. The datasets provided by NextStrain could be used to explain this question, as there are the name of the new found mutations and the corresponding date. The first dataset is within the UK and the second one is global.

To answer **Question 2**, we need dataset related to the amount of confirmed cases over time, since we need to observe the relationship between the amount of confirmed cases and the mutations development. But it's easy to get data from the Internet using API.

To answer **Question 3**, since we already have the cases data and mutations development data. We will need the morality rate of Covid-19 over time to represent the harmfulness.

1.3 Correlation Analysis

Since the **Question one** is about seasonality, the provided dataset could be viewed as time series. Given the datasets, the time interval could be day, week or month. According to Wikipedia, seasonality is the presence of variations that occur at specific regular intervals less than a year. The time intervals in the dataset are less than a year, so they could be used. Graphs could be plotted showing the variation and its frequency to check whether seasonal.

In **Question 2**, new daily cases will be set as the independent variable and newly discovered covid-19 mutations as dependent variable. If the amount of newly discovered covid-19 rises as the cumulative cases rise, then they are positively correlated. If not, they are negatively correlated.

To prove the theory mentioned in **Question 3**, it is important to find the correlation between cumulative cases and morality rate, and the correlation between newly discovered mutants amount and morailty rate. If both of them are negative, the theory is true for Covid-19.

- nextstrain _ncov_open_global_metadata.tsv, this dataset has Covid-19 strain found in the world with corresponding date.
- `https://api.covid19api.com/total/country/` + ”slug”, this url returns new confirmed cases and deaths for one specific country. Data for question 2 and 3. The slug is the country name.
- global_cumulative_cases.csv, this dataset has global cumulative cases, deaths amount, new death, new cases and morality rate until 15th April. And the source is `https://covid.ourworldindata.org/data/id-covid-data.csv`

2 Part 2: Design and Prototyping

The whole design of this project is inspired the WHO Coronavirus (COVID-19) Dashboard which uses a world map to show each country's data. For mutations, a tree map will show the hierarchy. In this section, *Matplotlib* and *D3.js* would be used for making prototypes.

2.1 Question 1

Question 1: *Analyze the development of virus mutations over time. Are there detectable seasonal trends?*
To detect the seasonality, line charts will be used in this question where y-axis stands for the amount of newly discovered mutants and the x-axis stands for the date/week. The development of mutations could be analysed from 3 different aspects.

2.1.1 Mutations Development Summary

In this section, all clades will be viewed as one, and line graph of daily new discovered mutations will be shown to find the trend.

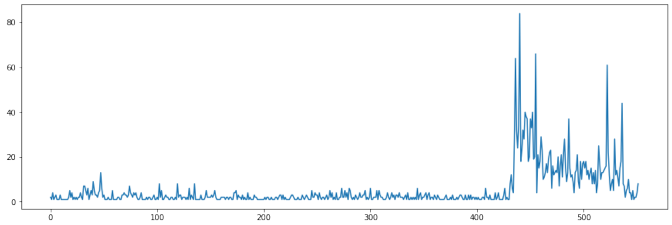


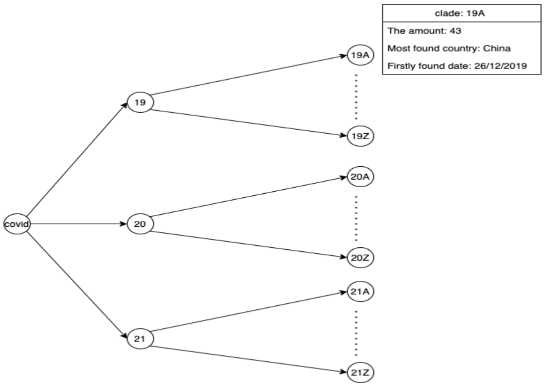
Figure 1: Global mutations development summary, interval: week

This line chart shows the overall development of mutations, but it ignores the difference between countries and the difference between clade. So the defects will be solved in the following 2 aspects.

[*] jinhong.jiang@kcl.ac.uk

2.1.2 One clade development

In this section, the line chart is about only 1 clade showing its development over time.
One hierarchical tree map will show all the clades and its hierarchy. Clicking on one specific clade, it will jump to the page showing the one's development over time. And when the mouse stops on the clade, details like the amount, the place where it is found the most and the date when it is first discovered will be shown. I used *draw.io* to draw this hierarchical figure.



clade: 19A
The amount: 43
Most found country: China
Firstly found date: 26/12/2019

Figure 2: Mutant hierarchy (when mouse over 19A)

After clicking on one specific clade, the page will jump to a page showing this clade's development over time. A line chart will show the development where x-axis is the date, y-axis is the newly found amount during that time.

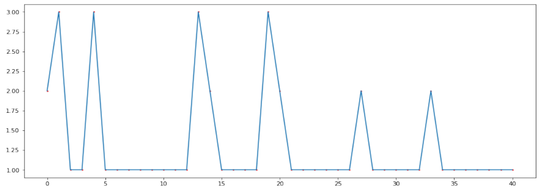


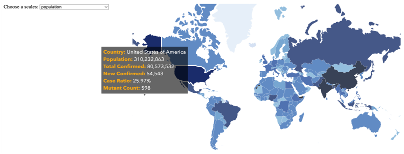
Figure 3: 19A development over time

With the hierarchical tree, it is easy to view one specific clade's development by just clicking on it.

2.1.3 Mutations development within a country

For interaction, a world map would be drawn where user could click on one specific country and jump to the page showing the mutations development in that country. And when the mouse is over one country, detailed data of this country will be shown.

The following figure will show the map part of this interaction system.



Choose a scales: population

Country: United States of America
Population: 310,232,863
Total Confirmed: 80,573,532
New Confirmed: 54,543
Case Ratio: 25.97%
Mutant Count: 598

Figure 4: World map, mouse over the states

After clicking on one specific country, it will jump to the page showing the mutation development over time in that country. The following figure will the show the line chart of one country. This prototype is made by *matplotlib*.

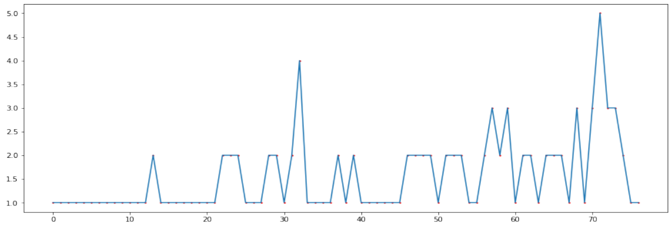


Figure 5: The mutation development in the UK over time

2.1.4 Conclusion

Inspired by the control variates method, I focused on different aspect respectively. Because there is one independent variable in each aspect, the error of an estimate of an unknown quantity could be reduced.

2.2 Question 2

Question 2: *How do new confirmed cases correlate with the mutation development?*

To see the correlation between 2 variable, 2 lines chart will be drawn, the x-axis of both of them are date, while the y-axis for the first one is the cumulative cases number and that of the second one is the newly discovered mutations amount.

2.2.1 Global

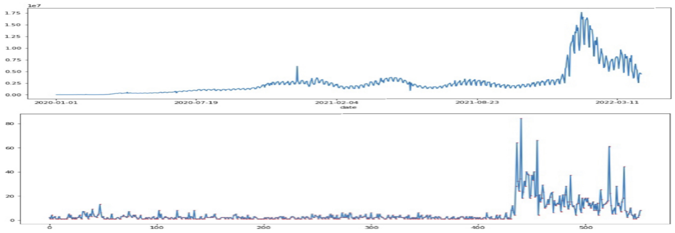


Figure 6: New cases and mutations development

Since we have the development of both new daily cases and global newly discovered mutations amount, the correlation could be detected.

2.2.2 Country-wise

Like the previous question, a global map will be used as the ”Dashboard” where user could click on one specific country and it will jump to a page showing the country’s new daily cases and daily discovered mutations counts. The global map will be exactly the same as the one in question 1. And the line charts will be same as global ones except the data difference.

2.3 Question 3

Question 3: *Is Covid-19 starting off harming their hosts, but evolving toward a more benign coexistence?*
In this question, by harming hosts, it means the high morality rate. And by coexistence, it means the virus becomes more infectious but less fatal.

In this question, one more line charts will be drawn representing new mutations’ morality rate, since there is already chart about mutation development.

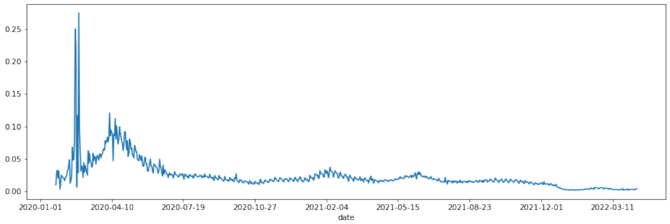


Figure 7: New mutations’ morality rate

3 Part 3: Implementation

I decide to implement the world map in Figure 4 and the tree map in Figure 2 as the main gate of interaction. My implementation could answer all the 3 questions.

3.1 Data Processing

It’s not easy to process data using Javascript, so I mainly used python in this project. And the following section will show how I did it.
For the mutation development, I used Pandas.groupy(date) to see the how the virus mutated. The result of the groupby will be like a key-value set where key is the date and value is the amount. For a specific country or clade, I used Pandas.Dataframe to filter.
Since there is no morality rate in the dataset, I used the quotient of new deaths divided by new cases.
I also download more data from the Internet to build the map. And the following table will show the attribute I used in this project.

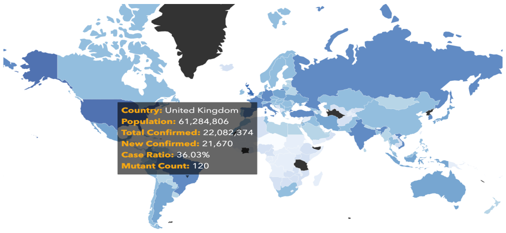
3.2 Data in Interaction

In Question 1, since we used the world map and tree map as the filter. When the mouse over the world map, details such as cumulative cases, new cases and deaths of the country will be queried. And the data will be shown in a pop out window like in the Figure 4.

When the mouse clicks on the world map, it will jump to the page showing the country’s data. The country’s name will be used as the parameter to request for the country data. The data will include the cumulative cases, mutation amount over time and morality rate over time.

When the mouse clicks on the mutation tree map, it will jump the page showing the clade’s data. The clade will be used as a parameter to query the specific clade’s mutation amount over time data.

The process will be shown in the following figures.



Country: United Kingdom
Population: 61,284,806
Total Confirmed: 22,082,374
New Confirmed: 21,670
Case Ratio: 36.03%
Mutant Count: 120

Figure 8: Mouse over the map

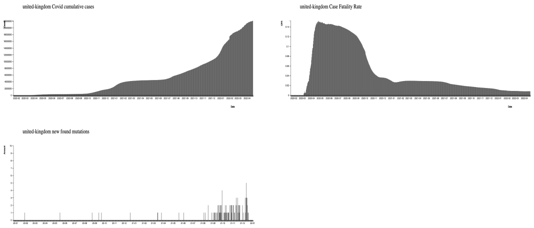


Figure 9: Jumping to one specific country

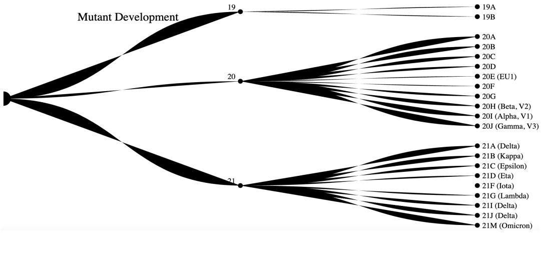


Figure 10: Mutations Hierarchy