Viral Vulnerability Analysis

Covid-19 in NSW

Kevin Marcelino, Sherry Fan, Suyeon Choi

Dataset Description

We preprocessed the data by removing any duplicates and unnecessary data as well as removing any area_id with Na values before importing them to PgAdmin. Here are the data sources used:

SA2 Geometry Dataset: The shp file contains information about polygon geometry for every area_id. We extracted the SA2_MAIN16 (later renamed area_id), the SA2_NAME16(later renamed area_name) and geometry(polygon geometry information) from this dataset. The cleaned dataset has 3 columns which are area_id, area_name and geometry, a set of polygon coordinates of each area.

Postcode Geometry Dataset: The dataset 'NSW_Postcodes' contains information about the postcodes within Nsw. After removing duplicates postcodes, we extracted the postcode from this dataset and then converted the latitude and longitude values into scalar geometry values with the geometry type set as POINT. The cleaned dataset has 2 columns which are postcode and geometry, a set of coordinates of where the area of the postcode is located.

Age Dataset: The dataset 'PopulationStats2016' contains statistics of the population in each area. We extracted the area_id, and calculated the number of elders(individuals ages 70 and above) per area_id. We did this by summing the number of elders aged 70-74, 75-79, 80-84 and 85_and_over from the original data set for every single area_id. The cleaned dataset has 2 columns, the area_id and elder which is the total number of people aged 70 and above per area_id.

Health Service Dataset: The dataset 'HealthServices' contains specific information about health institutions. We extracted name, category, num_beds, longitude and latitude from this dataset and then converted the latitude and longitude values into scalar geometry values with the geometry type set as POINT. The cleaned dataset has 4 columns which are name,category, num_beds and geometry, a set of coordinates of where the health service is located.

Employment and Transport Method to Work Dataset: An additional dataset, 2016Census_G59_NSW_SA2 is used in our project. The original dataset is under the package of *Statistical Area Level 2 (SA2) ASGS Edition 2016 in .csv Format* from ABS database. This dataset contains the counts of people with different methods of travelling to work by neighbourhood. We extracted area_id and number of employment from this dataset, and we calculated the number of people who travel to work only by public transportation. This calculation was done by summing the counts of people only travelling by bus, ferry, tram/train and taxi. The cleaned dataset has three

columns which are area_id, employment counts, and counts of people only traveling by public transportation.

Neighbourhood Profile Dataset: This dataset was created by extracting area_id, area_name,land_area,population and median_annual_household_income(renamed as income) from the neighbourhoods dataset and then merging it with the cleaned Age dataset and the cleaned employment and and transport method to work dataset using the area_id as the unique identifier to join them. The cleaned and merged dataset contains 8 columns, area_id, area_name, land_area, population, income, employment, trans_public and elder.

Test Dataset: This data set contains information about covid19 tests taken in NSW, with data sources taken from the NSW Government data, a reliable source as it is data provided by the government. We extracted result_x, result_y and the postcodes and grouped the postcodes from the dataset to obtain the total number of cases confirmed that occurred for each postcode. We also used a dictionary to obtain the area ids of the corresponding postcodes to merge and join with the postcodes form the dataset, also extracting the area_id. We then merged the grouped dataset with the original dataset using the postcode as the key identifier. The cleaned and merged dataset contains 4 columns: postcode, n_test(renamed from result_x), cases (renamed from result_y to cases) and area-id.

Database Description

Database Schema

We wrote the cleaned SA2 geometry dataset into an SQL schema called `sa2geom`, and the cleaned postcode dataset called `postcode`. After spatial join based on the geometry point from `postcode` and geometry polygon from `sa2geom`, we stored the result in a new schema called `sa2postcode`, which makes future work easier by linking postcode to SA2 area id without additional spatial join. And it was written to Jupyter notebook to create a postcode to sa2 area id python dictionary. We wrote the health service dataset into a schema called `health`, and spatial joined the location of each health service to sa2 area id. Number of health services and number of hospital beds are counted from grouping the spatial join result by sa2 area id, and the result was stored into a new schema called `healthcount`. Cleaned neighbourhood profile was written to SQL schema called `profile`.

`Profile` and `Healthcount` were joined on area id, the query result was written to Jupyter notebook for calculating vulnerability score. The result of vulnerability score, number of tests and number of confirmed cases by each sa2 area id was then written into SQL schema called `score`. The purpose of this step was to enable a quick query just for score, tests and cases in SQL.

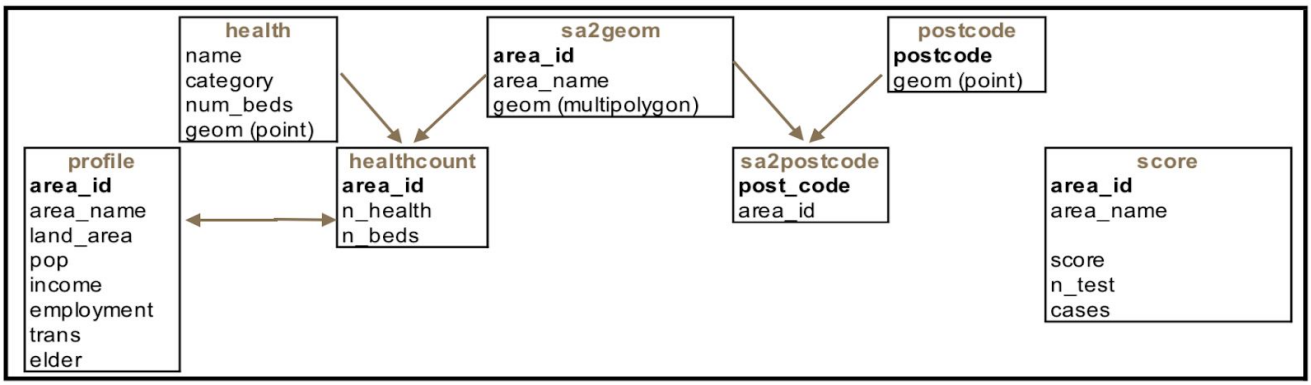

Figure 1. Schematic diagram of database schemas

Indexing

We created three indices in our SQL schemas. The first one is sa2 area id in `sa2geom`, `healthcount`, `profile` and `score`, which allows fast join and neighbourhood matches in SQL. The second one is postcode in `postcode` and `sa2postcode`. The third one is multipolygon geometry information in `sa2geom`. This is a spatial indexing, which indexes the bounding boxes of the geometry polygons to fasten spatial joins and spatial queries.

Vulnerability Score Analysis

Feature Selection

Query result of joining `profile` and `healthcount` was stored in pandas data frame `z_df` for easier standardisation and sigmoid function calculation. The following variables are considered for vulnerability score:

- Population density of each neighbourhood (pop_density): crowded areas are more likely to have a higher risk of human-to-human transmission, so there could be a positive correlation to the vulnerability.
- Percentage of elder people (elder_pct): older people (70+) are more vulnerable than other age groups. The vulnerability could positively relate to the proportion of elder group.
- Number of health services per 1000 people (n_health_1000): More health services can reduce the vulnerability. There could be a negative correlation.
- Number of hospital beds per 1000 people (n_beds_1000): More hospital beds can reduce the vulnerability. The correlation should be negative.
- Median income (income): Higher income means people have more money to travel around, which increases the risk of virus transmission. There could be a positive correlation.

- Percentage of employment over labour force (employment_rate): Higher employment means more people traveling everyday. The correlation should be positive.
- Percentage of people who travel to work only by public transportation over all traveling methods to work (trans_pct): The vulnerability level should positively relate to the proportion of people who go to work only by public transportation.

Scoring Function

All features are standardised before calculating the vulnerability score. This procedure could ensure all variables are in the same scale.

Vulnerability score is calculated using sigmoid function, which takes the sum of all standardised features' values as its independent variable.

$$vulnerability\ score = \frac{1}{1+e^{Z_{pop\ density}+Z_{elder\ pct}-Z_{n\ health\ 1000}-Z_{n\ beds\ 1000}+Z_{income}+Z_{employment\ rate}+Z_{trans\ pct}}}$$

Vulnerability Results

The top three vulnerable neighbourhoods are "Neutral Bay - Kirribilli" (score = 0.9999), "Crows Nest - Waverton" (0.9998), "Potts Point - Woolloomooloo" (0.9997). Figure 2 shows the vulnerability score map of NSW.

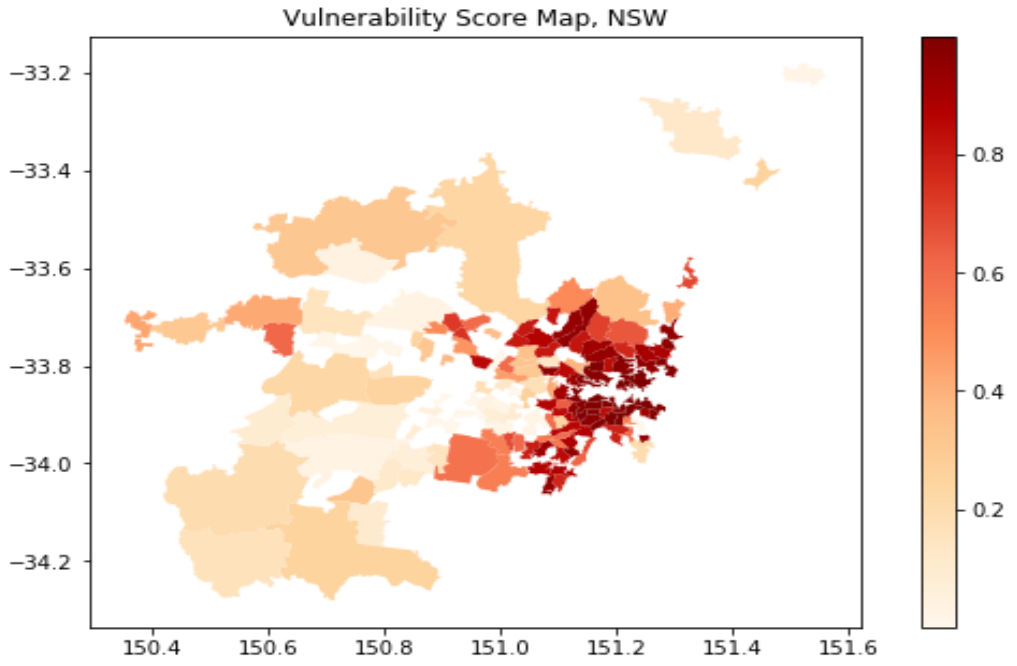

Figure 2. Vulnerability Map

Correlation Analysis

Using our scoring method with 7 identified features, the correlation coefficient between vulnerability score and number of confirmed cases is to 0.24, and the correlation coefficient between vulnerability score and number of tests is to 0.06. We found a weak linear relationship between confirmed cases and vulnerability score, but not for number of tests and vulnerability score.

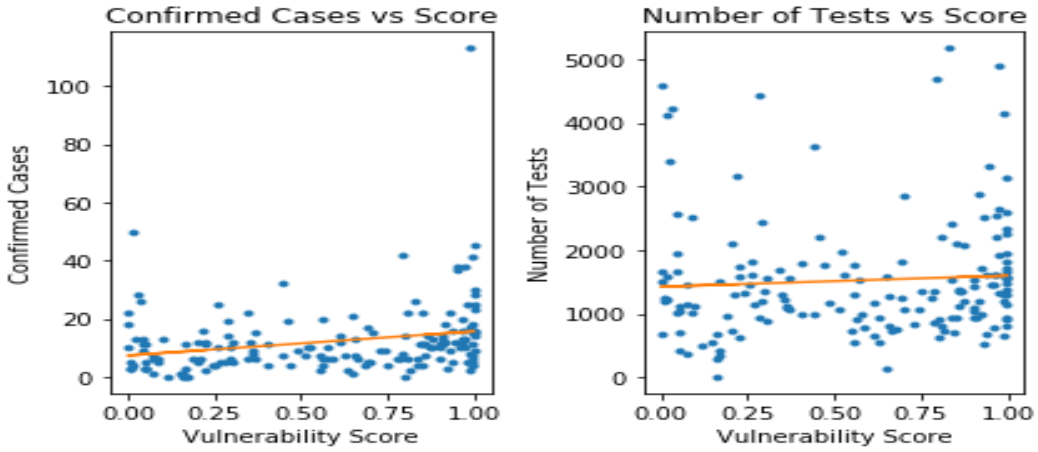

Figure 3. Result from correlation analysis

As Figure 3 shows, the diagram of Confirmed Cases vs Score shows there is a relatively small link between the vulnerability score vs confirmed cases, showing that an increase in the number of cases may increase the score. However, for the diagram of Number of Tests vs score, the graph shows there is very little relationship between Vulnerability score and the number of tests conducted.

Reference

Australian Bureau of Statistics (2016). *Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas,* retrieved on 22 May 2020, https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument