Viral vulnerability analysis documentation

1. Dataset description

a. Data sources

The ‘statisticalareas.csv’,‘neighbourhoods.csv’, ‘populationstats2016.csv’, ‘healthServices.csv’ and ‘NSW_Postcodes.csv’ datasets where obtained from canvas via the assignment brief and were all sourced from the ‘Australian Bureau of Statistics’ (ABS).

The level 2 statistical areas shapefile data(‘SA2_2016_AUST.shp’) was also downloaded and sourced from ABS with the current COVID-19 tests by date, location and result downloaded from data.gov.au and sourced by the ‘NSW ministry of health’ ("COVID-19 tests by date and location and result.csv", 2020) ("1270.0.55.001 - Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2016", 2016).

An additional NSW government school head counts dataset (‘schoolheadcount.csv’) was obtained from ‘NSW department of education’ with the GPS location of each school acquired through the open source ‘Open street maps’ API ("NSW government school enrolments by head count (2004-2018) ", 2019)("OpenStreetMap Nominatim: Search", 2020). These datasets were retrieved with permission under ‘Open street map’s’ Open Data Commons Open Database License (ODbL), with the school headcount, COVID tests and Sa2 shapefiles datasets being under the Creative Commons Attribution License (cc-by).

b. Pre-processing of data

All datasets were filtered for columns which suit the schema, had duplicates removed in ‘primary key’ columns and removed null and invalid values, i.e. postcodes ‘9999’ or ‘0’, and entries with longitude/latitude being ‘0’. All geometric points were turned into shapely objects and data was pushed into postgreSQL via Jupyter notebooks and psql as the copy command was used for ‘neighborhoods’, ‘schools’ and ‘covid_tests’ datasets.

Neighbourhoods and ABS shapefile

The shape objects of the appropriate statistical areas in the abs shapefile dataset were extracted and appended to the ‘neighbourhoods’ dataframe. All null values in neighbourhoods were replaced with ‘0’ to prevent loss of any neighbourhood areas for analysis.

Populationstats2016

Population statistics had no missing values, however populations with the ‘total persons’ being 0 were removed as each age category was converted into a proportion of the total population in the sa2 area before being pushed into the database.

Health services

The health services dataset was split into hospitals, which included ‘num_beds’, and Gps to normalise the dataset. Missing values were only present in ‘num beds’ but entries were kept to preserve the number of health services. The health service ‘id’ was kept as a primary key for both ‘Gps’ and ‘Hospitals’ as health service names were similar in both categories.

COVID-19 testing data

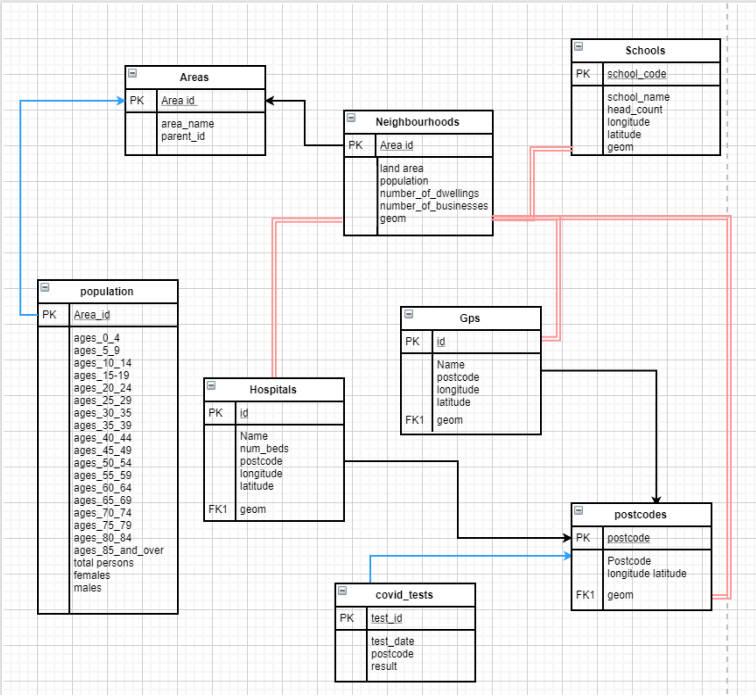
All recorded tests with no postcodes were removed as this would not be included in the spatial join and ‘test date’ was rearranged to ISO 8601 format to suit postgreSQL’s formatting requirements.

School headcount

The school headcount dataset was filtered to only have the head counts for the latest year, 2018. Schools with no headcount or headcount with the placeholder ‘SP’ were dropped. The GPS locations of each school were acquired through ‘Open street maps’.

2. Database description

The image below depicts the schema used, where black arrows indicate referenced foreign keys, red lines indicate potential spatial joins and the blue arrows indicate potential joins without foreign keys.



'Area_name' is unique to each 'area_id' and hence was removed from both 'neighborhoods' and 'population' ("1270.0.55.001 - Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2016", 2016). No foreign keys were made between 'population ', 'areas' and 'neighbourhoods' as 'population' included sa2 areas classified as 'rest of NSW areas' whereas 'neighbourhoods' and 'areas' only had 'greater sydney region'. Inner joins were used instead for later analysis.

Although multiple postcodes contained multiple suburbs and multiple suburbs contained multiple postcodes, a primary key of only 'postcode' rather than 'postcode, suburb' was chosen. This was because upon looping through the dataset, postcodes with different geometries often only had 1 varying geometry with '2880' having 4. Further there would be no way to join 'covid tests' to a primary key of 'postcodes, suburb' without duplicates, as suburbs could only be implicated through local government areas, but these also covered multiple suburbs. Hence the most common geometry was chosen for postcodes with varying geometries.

'Covid tests' used a self-produced id as a primary key as multiple tests were conducted on the same day in each postcode. As the covid tests dataset had test cases outside the NSW region, no reference between 'postcodes' and ' covid_tests' were made in terms of foreign keys.

All Geometry columns were set as not null and transformed into GDA94 SRID, i.e. 'schoolheadcounts' data, as these were more precise coordinates for working with data in Australia ("FAQ - OpenStreetMap Wiki", 2020).

Indexes

R- trees using postgreSQL's PostGIST extension were created on all geometry objects. This was created to allow the system to process queries with spatial joins more efficiently as all datasets with geometry objects would be joined with neighbourhoods with a spatial join.

Additional hash indexes were created on covid_tests 'postcode' and 'result' columns to maximise efficiency in joining the large 'covid_tests' datasets to 'postcodes' and filtering by test results in the where condition.

3. Vulnerability score analysis

a. Vulnerability score construction

Each area's vulnerability score was computed from the following parameters:

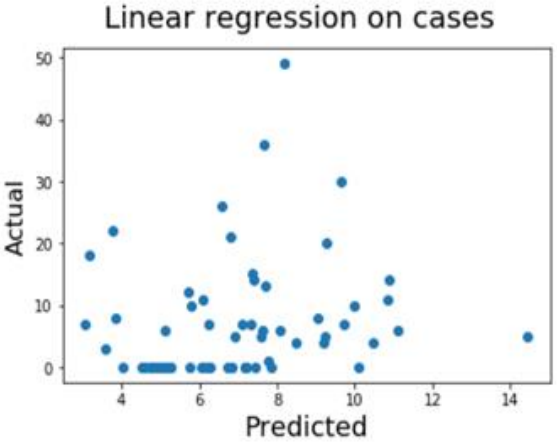
1) Population density
2) Number of dwellings
3) Number of businesses
4) Number of school children
5) Number of public hospitals
6) Average number of beds available in public health systems
7) Number of private hospitals
8) Number of GPs
9) Percentage of population over 70

In each area, the number of COVID tests was highly correlated to the number of positive cases. A linear regression model was used to create a line of best fit which showed the expected proportion of positive cases to a certain number of tests conducted in each area (see graph in Appendix 1). Subsequently, the residuals for each area compared to the line of best fit was used as an indicator of an area's vulnerability score. Machine learning was then used, inputting parameters above, to create a linear regression algorithm which predicted each area's expected 'residual' score. This residual score was then fed into a sigmoid function to obtain a vulnerability score between 0 and 1, (see Appendix 1 for further details).

b. Justification of vulnerability score construction

The main obstacle in constructing a vulnerability score was eliminating bias created by areas without any cases. Initial attempts included simply applying linear regression to predict the number of cases, however immediate problems arose where areas with no cases created bias as the algorithm was unable to distinguish between areas that are truly less vulnerable and areas yet to contract the virus.

The graph below depicts a semi solid line at 'Actual cases = 0', demonstrating how regressing on cases without tests may create bias in the training process.



In addition, this simplistic model does not account for the fact that areas with more tests will naturally exhibit more positive cases.

Pros

Our method was able to account for both concerns raised above. The machine algorithm was trained a random sample of areas with COVID cases to avoid the bias caused by areas with zero cases. Additionally, by using the residuals as the metric for vulnerability, bias from the imbalance of test distribution across NSW is eliminated.

Drawbacks

The main limitation of this model is selection bias created from the exclusive use of areas with COVID cases in the training data as there may be patterns within the subset that may skew the training of the model. In comparison to the problem this method is intending to fix (that is bias caused from areas yet to exhibit COVID cases dampening the vulnerability algorithm), we deemed it as a beneficial trade-off. Further investigation such as tests for homogeneity between the subset of data (cased areas vs non-cased areas) will be beneficial to quantify selection bias.

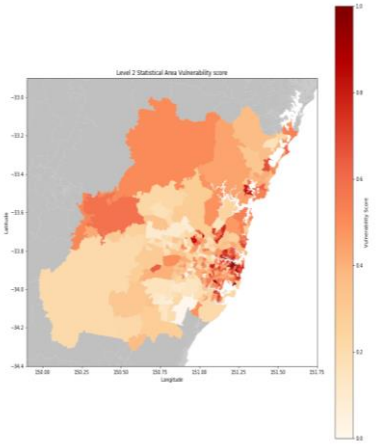
c. *Results*

The linear regression algorithm resulted in the 9 parameters weights:

names	coefs
number_of_dwellings	-0.000147
number_of_businesses	0.000368
head_count	-0.000379
pubhos	0.966245
pubbeds	-0.005892
privhos	2.652033
gps	-0.206606
senior	21.272270
density	0.056848

From the coefficients, we see the most significant contributor to vulnerability was senior population, followed by private and public hospitals. This is consistent with preliminary understanding of COVID as it is well documented that the senior population is more vulnerable to the virus. The positive coefficients of private and public hospitals initially seems surprising, as it would be expected that areas with health infrastructure in place would be less vulnerable. However, as most testing is done in hospitals, the high risk of transmission in health facilities accounts for the results, (see Appendix 3 for school and health facility distribution).

Overall, our findings are as follows with a histogram of the vulnerability scores and a heatmap of the scores for each statistical area.



The 5 areas with the highest vulnerability score were:

	area_id	number_of_dwellings	number_of_businesses	head_count	pubhos	pubbeds	privhos	gps	senior	density	vfinal	area_name
88	118011342	0	0	0	0	0.0	0	0	1.000000	0.015159	1.000	Centennial Park
70	117011324	0	135	0	0	0.0	0	0	0.857143	0.024964	1.000	Port Botany Industrial
83	117031337	8963	40285	220	1	61.1	0	43	0.041508	72.233869	0.977	Sydney - Haymarket - The Rocks
75	117031329	3914	2260	322	1	431.5	1	4	0.052287	144.854142	0.972	Darlinghurst
80	117031334	7599	3933	290	0	0.0	0	4	0.042590	161.522468	0.969	Pyrmont - Ultimo

The 5 areas with the lowest vulnerability score were:

	area_id	number_of_dwellings	number_of_businesses	head_count	pubhos	pubbeds	privhos	gps	senior	density	vfinal	area_name
290	127031598	7443	2589	3676	2	1593.8	0	30	0.079881	46.819117	0.001	Liverpool
276	127021511	5800	1872	3015	0	0.0	0	34	0.088609	34.345708	0.028	Cabramatta - Lansvale
112	119011572	3942	1991	1974	1	389.7	0	25	0.078079	38.335971	0.050	Bankstown - South
192	123021437	5360	1744	2226	2	469.1	0	20	0.099817	14.066773	0.056	Campbelltown - Woodbine
47	116011303	5044	2319	1894	1	428.6	0	16	0.072841	22.405016	0.067	Blacktown (East) - Kings Park

From above, it can be noted that areas with low vulnerability score tended to have a lower senior population and lower population density than those with a high vulnerability score. It is interesting to note that the number of school children in areas of higher risk were significantly lower than those areas with lower risk.

4. Correlation analysis

Using the Pearson correlation coefficient (r), the vulnerability score was correlated to both the number of cases and tests.

	cases	vfinal
cases	1.000000	0.128374
vfinal	0.128374	1.000000

With 'vfinal' being the vulnerability score, we see that between the cases and the score, the correlation coefficient is 0.129, a low correlation which indicates that the plot does not fit well along a line. The low correlation is further exemplified in the scatter plot of cases vs vulnerability score as there is no identifiable pattern in the plot (see Appendix 2 for map between vulnerability score, tests and positive cases).

	tests	vfinal
tests	1.000000	-0.034667
vfinal	-0.034667	1.000000

Furthermore, correlating the tests with the vulnerability score we see a negative correlation coefficient, indicating that there is a slight negative correlation between the number of tests and the vulnerability. However the score is so close to 0 that the correlation is insignificant to draw any meaning

The extremely weak correlation result between vulnerability score and number of tests and cases is expected due to the design of the vulnerability score and the nature of the data. Firstly, because our vulnerability score is a measure of an area's residual to the line of best fit above, it would be expected that there is a low correlation between it and the explicit number of cases or tests. Additionally, the test for vulnerability may not be well correlated to cases and tests as some areas, whilst vulnerable, are yet to have any COVID testing or cases.
A better alternative to quantify correlation is the RMS error and R-squared score from the machine learning algorithm. By only using 30% of the dataset to train the regression algorithm, we were able to use the remaining 70% to test the accuracy of the algorithm and the results are as follows.

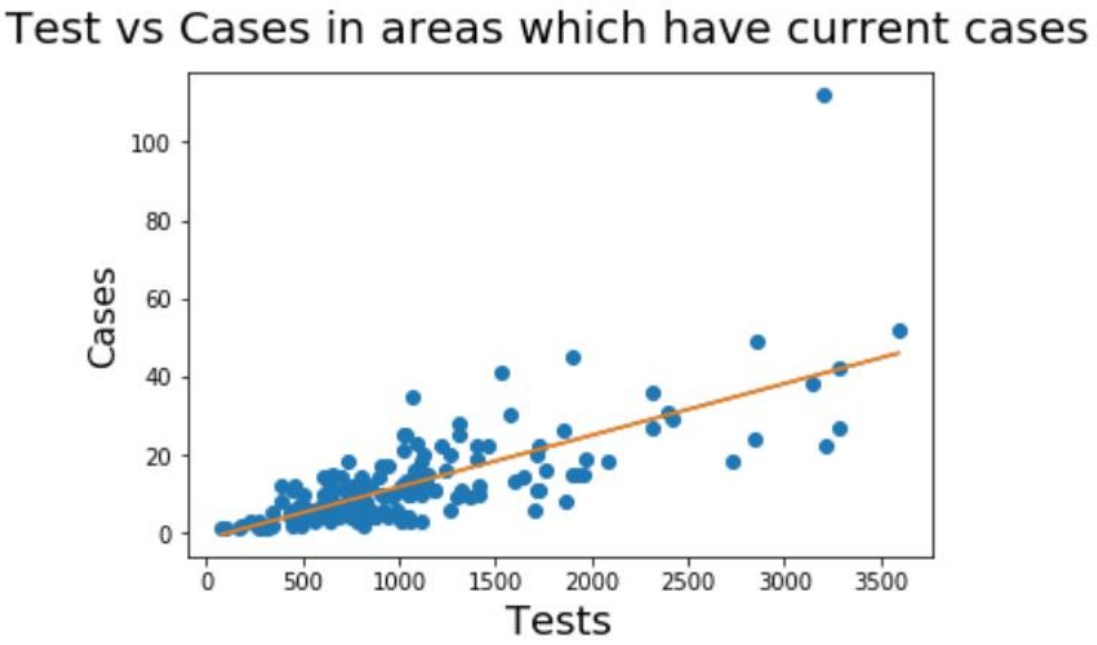
```
Root mean squared error (RMSE): 6.382955756131551
R-squared score: 0.1077926589460203
```

The RMSE shows that the standard deviation of the residuals is 6.4 and the R-squared score is 0.1077 which again are two very poor results showing that there is little correlation between the vulnerability score and an area's quantified vulnerability.

5. Appendix

1. Further explanation on the method used to compute vulnerability score

[1] Graph showing the line of best fit



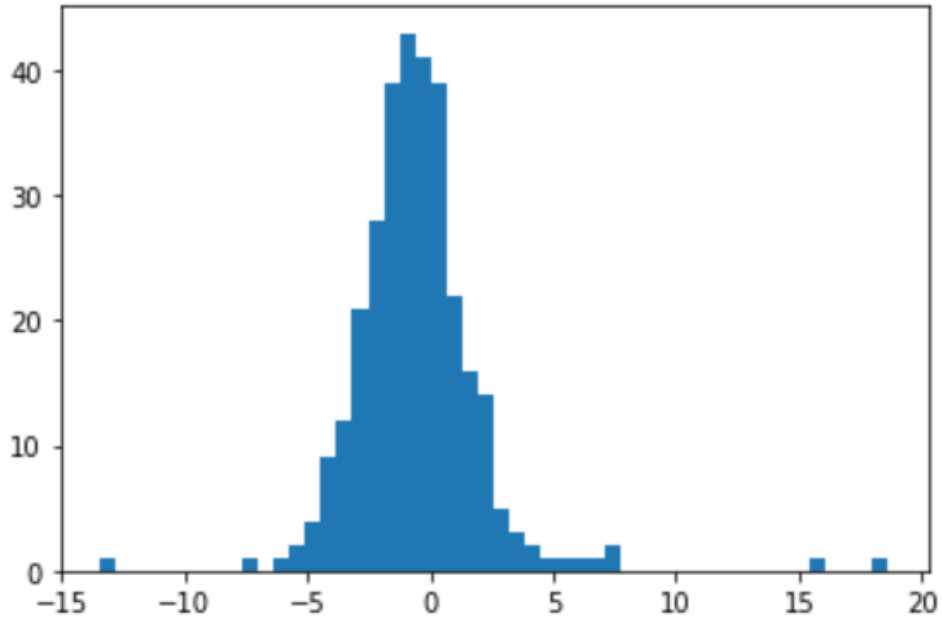
By plotting tests vs cases as done above, there is a clear linear correlation between the number of cases and the number of tests. With the assumption that tests were conducted randomly within each area, this line of best fit represents the expected number of positive cases given x number of tests, and the gradient represents an average infection rate for COVID 19 across NSW. The gradient and the intercept was computed using the python module stats.models.

If an arbitrary area falls above this line, then it suggests that the area has more cases than we expect, given the number of tests conducted, and thus a higher infection rate. The inverse is true if an area falls below the line. From this, specifically the residuals from the line of best fit, we were able to extract a metric for an area's vulnerability.

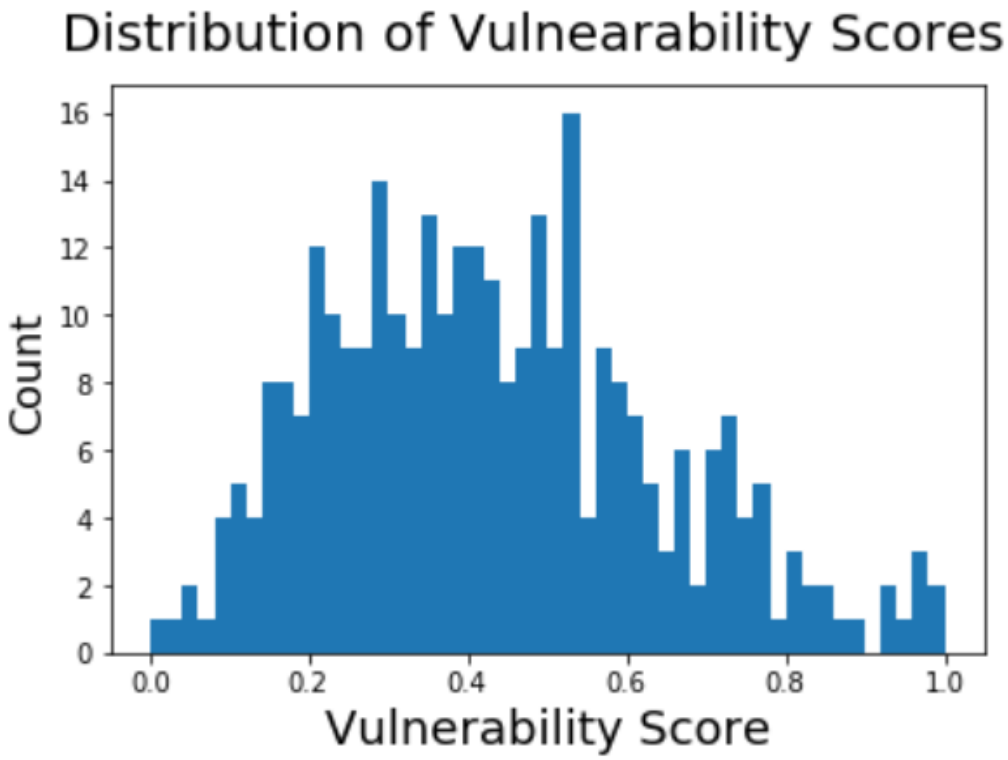
The vulnerability score was then computed using linear regression and supervised machine learning to predict the residuals with a positive residual equating to higher vulnerability and a negative residual meaning low vulnerability.

The python module SKlearn was used to create the linear regression model. The dataset was subset by areas with and without existing cases, and the regression model was trained using a random sample from the subset with cases. This was done so that the machine learning model could remain supervised without areas yet to have COVID cases, which created large biases.

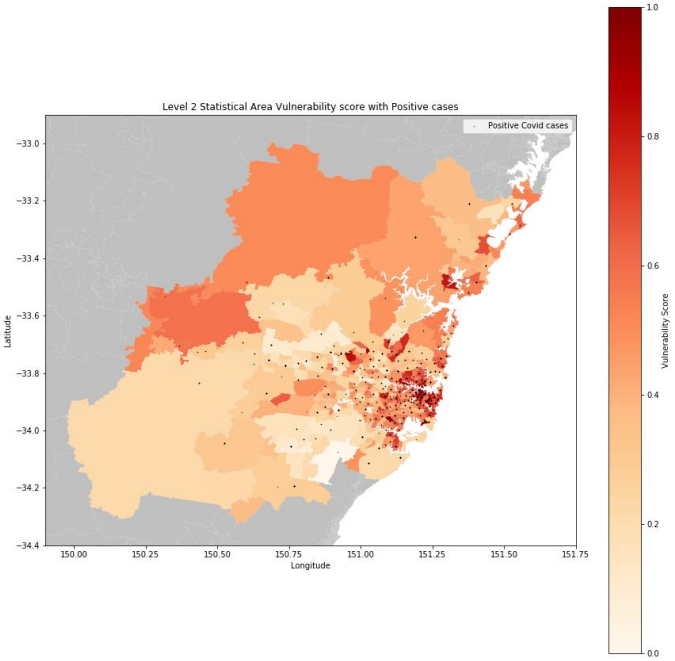
The regression model was then applied to the whole dataset to predict an expected residual for all areas, producing a distribution of scores as follows.

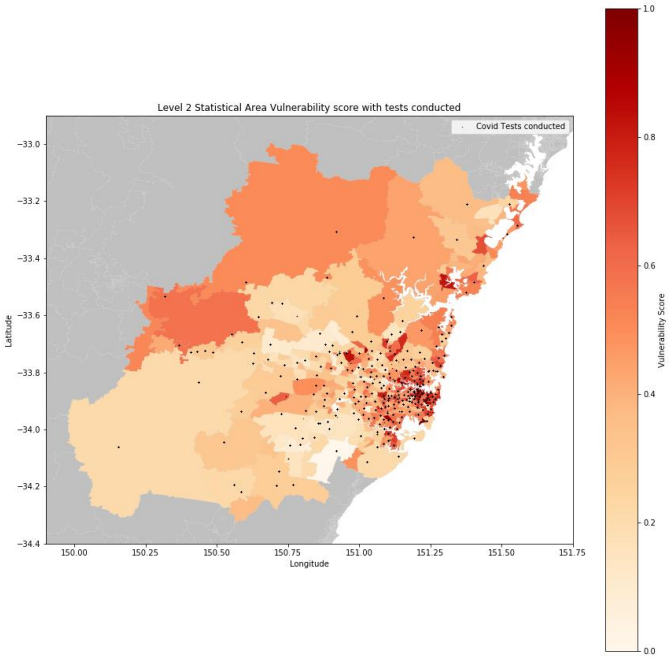


Finally, the predicted residuals were ran through a sigmoid function to produce a final score between 0 and 1, 0 representing low vulnerability and 1 representing high vulnerability. This sigmoid function was scaled to ensure most points remained within the range.

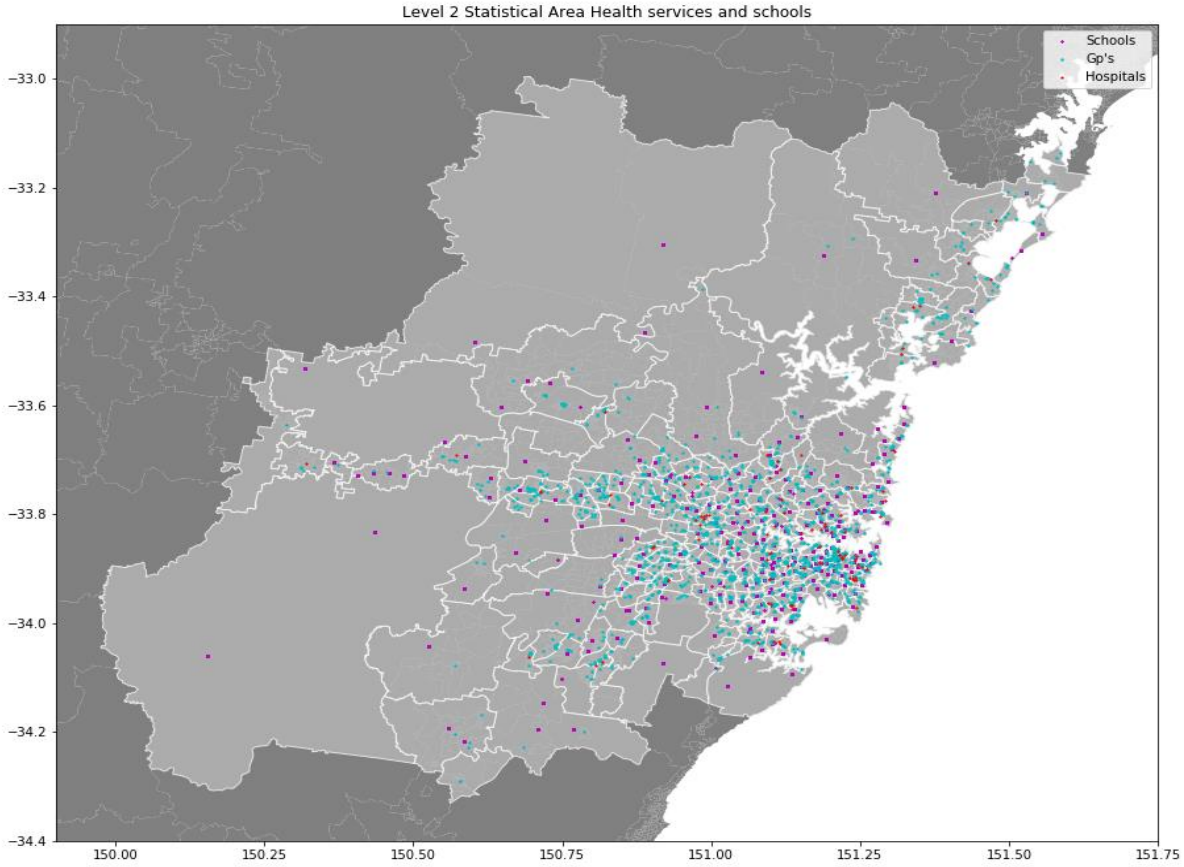


2. Positive cases and number of tests conducted against the computed vulnerability score





3. Map depicting the distribution of schools, GPs and Hospitals across NSW SA2 areas



References

1270.0.55.001 - Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2016. (2016). Retrieved 9 May 2020, from https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument

COVID-19 tests by date and location and result.csv. (2020). Retrieved 9 May 2020, from https://data.gov.au/dataset/ds-nsw-5424aa3b-550d-4637-ae50-7f458ce327f4/distribution/dist-nsw-227f6b65-025c-482c-9f22-a25cf1b8594f/details?q=

FAQ - OpenStreetMap Wiki. (2020). Retrieved 13 May 2020, from https://wiki.openstreetmap.org/wiki/FAQ#What_geographic_datums_are_used_in_OpenStreetMap.3F

NSW government school enrolments by head count (2004-2018). (2019). Retrieved 11 May 2020, from https://data.cese.nsw.gov.au/data/dataset/nsw-government-school-enrolments-by-head-count

OpenStreetMap Nominatim: Search. (2020). Retrieved 13 May 2020, from https://nominatim.openstreetmap.org/search