Visualising the Spread of COVID-19 in NSW

Jesse Serina Narvasa
Student ID: 500525438
The University of Sydney
Sydney, Australia
jnar3156@uni.sydney.edu.au

I. Nature of the Data

The advent of the COVID-19 pandemic has brought in new measures and restrictions in different states and countries worldwide, on the movement and actions of people, in an effort to control and minimise the number of preventable hospitalizations on its citizens. Since the first arrival of the virus in Australia with the first confirmed case in January 2020 [1], testing along with contact tracing have been the key line of defence to minimize the number of community transmissions.

Data on the number of tests conducted, and more importantly the number of positive cases arising from each geographical area are therefore useful for government officials such as the state premiere and NSW Health, in ensuring control over the virus. The availability of this data would allow for the identification of areas of concern, where key indicators such as the percentage of positive cases over the total number of tests within the past week, might suggest a deteriorating situation where government intervention such as lockdowns might be required.

The dataset used for this project, will be a combination of the number of tests conducted by location [2], along with the number of confirmed cases by location [3], as provided by the government agency, Data NSW. The number of tests performed by location is provided in an aggregated format as a sum of the number of tests conducted by postcode for a given date of testing. On the contrary, the datafile for the confirmed cases is not in an aggregated format, with each row representing a confirmed case, with the date of notification to the concerned person and officials. Shared attributes amongst the two files includes the Local Health District (also referred to as LHD) code and name, Local Government Area (also referred to as LGA) code and name, along with the post code.

The attributes available within the tests by location datafile therefore represents the total number of tests performed on a given day, by postcode, which is a subset of LGAs, and is in-turn a subset of LHDs. On the other hand, the confirmed cases datafile represents each individual COVID-19 case, described by the date it was notified to health officials, along with the person's residential location as previously described in the postcode, LGA, LHD hierarchy.

The collection of data presented within the datasets would have been obtained through the network of testing laboratories in partnership with NSW Health, as they conduct the test on presence of COVID-19 on each person's swab from a testing location. The testing laboratory, however, will be dependent on the information provided by the testing clinic where personal details along with the nose & throat swabs are collected. Therefore, the accuracy of data, with respect to the number of tests conducted by postcode, along with the location of each positive case will be dependent on the accuracy of information provided by the person getting tested.

II. Consumers of the Data

A. State Government – Government of New South Wales

As briefly mentioned in Section I, the government of New South Wales would be amongst the primary users of this dataset, due to the nature of their role in ensuring safety amongst the people in New South Wales, and the effect and impact this pandemic has the potential to cause on the safety of the people along with the economy of the state.

Data provided such as this would allow the state government to make an assessment on the current outlook of the state in terms of containing the virus and identify any location of concern. Further consumption of the data would then allow the government to set key indicators that may trigger government intervention to prevent any further deterioration in circumstances an area might have. This application is evident in the introduction of stricter lockdown policies in particular LGAs of concern, that enforces more restrictive movement of people within those residential locations in which the decision enacted would have been supported by information such as the trend on the number of cases detected from certain LGAs being marked as a concern.

B. Health Organisations – NSW Health, Local Health Districts, Hospitals, Pathologists

In conjunction with the state government, officials within the NSW Health would have also relied on such data to make decisions on the recommendations it provides to the state government with regards to the level of restrictions required, along with other additional measures that needs to be put in-place. As such, health advice and recommendations provided to the government would be supported by information derived on cases and testing data.

Moreover, Local Health Districts can also use this data to prepare for expected surges in cases and allow reallocation of its staff within various hospitals to focus on areas that currently or is predicted to require more personnel. Capacities can be readjusted amongst different wards in hospitals in proximity to areas with increasing case numbers and other forward planning measures can be put in place.

Lastly, the number of tests being performed within each location can give officials a sense of idea as to whether there is enough encouragement within the community to get tested or if further campaigns are required to promote that. Likewise, testing clinics and pathologists can use this data to gauge utilization amongst its different testing clinics within the network, and if capacity needs to be increased in certain areas of the state.

C. *Residents of New South Wales, the Country, and the World*

People within various communities in the state are likewise interested in this data. As evident by the daily broadcast and media reporting on the number of cases detected within the state, along with testing figures, there is demand from people to know how well the state and each LGA is containing the virus. In effect, everyone within New South Wales has an interest in this data because it affects the roadmap to freedom for the state, the restrictions enacted upon everyone's Local Government Areas, and the mobility and actions one is allowed to part-take within each of those communities.

Residents of Australia will also have an open interest in how New South Wales is containing the outbreak. This is because the outcome of the country's strategy in handling the pandemic can be greatly influenced by the situation in NSW, given it is the most populous state. Therefore, decisions made by the Federal Government, in response to any deterioration of situation can also have a negative consequence amongst residents of other states.

Finally, people around the world would be interested in how New South Wales progresses. Being a hallmark example throughout most of 2020 for its successful containment of locally acquired cases, there will be interest in the policies the government sought to enforce and how it corresponds to the increase or reduction in the number of cases.

III. TYPICAL VISUALISATIONS OF THE DATA

The datasets concerning the number of positive cases along with the number of tests conducted are generally visualized in a geographic manner, due to the location-based nature of the data, with the positive cases and test numbers being used as the quantity demonstrated within the map.

In particular, the attributes "postcode", "lhd_2010_code", "lhd_2010_name", "lga_code19", "lga_name19" are amongst those which are of nominal datatype. Specifically, these attributes provide the geographic information in both government and health district level. This is demonstrated by the thematic map visualisation provided by the NSW Government [4], where the data can be split on different granularities: postcode and LGA.

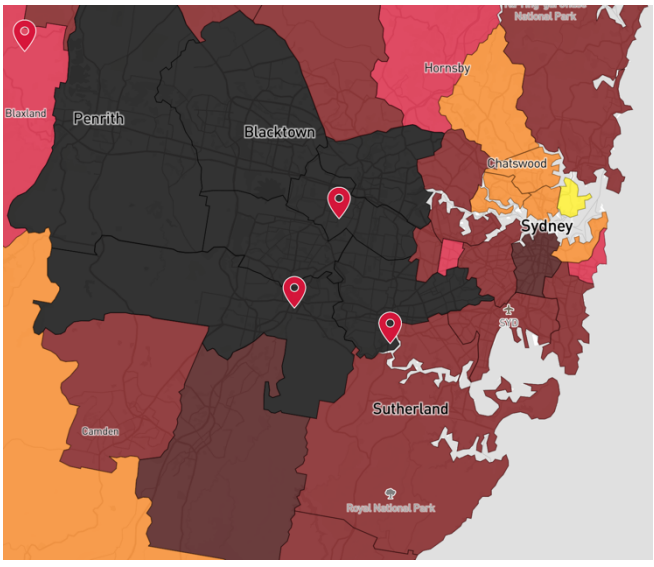

Figure 1 - Case Numbers split by LGA

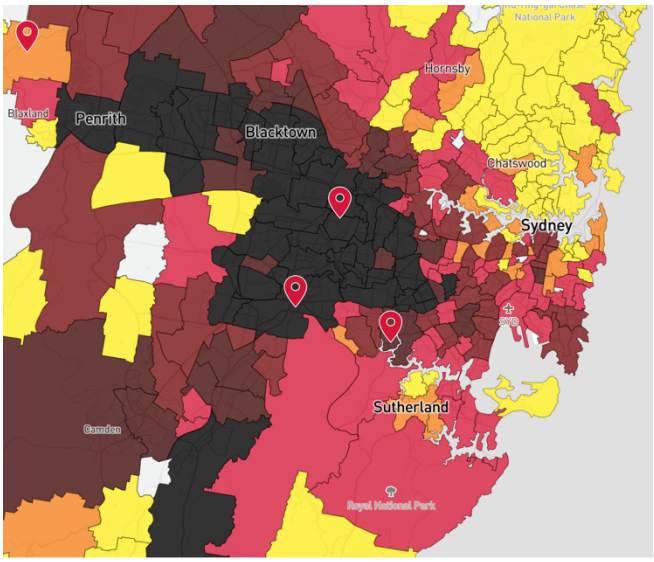

Figure 2 – Case Numbers split by Postcode

While there is a hierarchy between postcode and LGAs, with Local Government Areas encompassing numerous postcodes, the visualisations itself are independent of that. Instead, the categorization by postcode and by LGA, both attributes being treated as categorical nominal data, are separated into two thematic maps, illustrating the same quantity values, as demonstrated in Figures 1 and 2.

As previously mentioned, the aggregated number of positive cases or number of tests conducted are then used as the numerical measure to provide the colour fill within the boundaries of each category (represented as either a postcode or LGA) within the map. Due to these measures having a true zero, meaning that there is an absence of COVID-19 should there be 0 positive cases or no test conducted if testing number is 0, then we can conclude that these quantity measures are of ratio data type. These ratio data can then be represented within visualisations in a colour scale that would have at least a zero minimum i.e., the scale does not go negative, with the case numbers by postcode being demonstrated in Figure 2, and test numbers similarly shown in Figure 3 below:

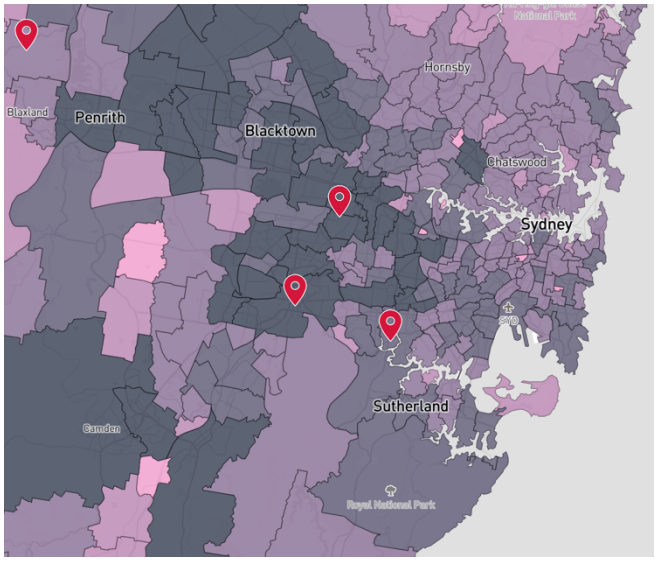

Figure 3 - Test Numbers split by Postcode

What thematic maps are able to achieve in its visualisations, is present the consumer with a geographic representation of how one location is doing in terms of positive cases or testing, in relation with another. A result of this is the ability to quickly compare different areas in its performance, identify patterns with high COVID-19 cases and potentially provide additional information such as logistical difficulty in setting up more testing clinics or limited hospital capacities in certain areas, which other visualisations may not be able to provide.

More high-level visualisations are also available, such as the number of COVID-19 cases announced daily since the start of the recent outbreak in NSW, as provided by The Guardian [5]. This visualisation introduces the temporal aspect of the "notification_date" attribute within the dataset, allowing case numbers to be presented as a function of time. The ability to present in an (ascending) order manner is one of the key advantages on the use of this attribute. Moreover, since date is represented as an interval datatype, viewers are able to measure the number of days taken to reach a certain threshold in the number of cases being reported, along with a sense of the rate in which case numbers are increasing or decreasing. In this sense, there are information available within this simpler visualisation that thematic maps cannot provide.

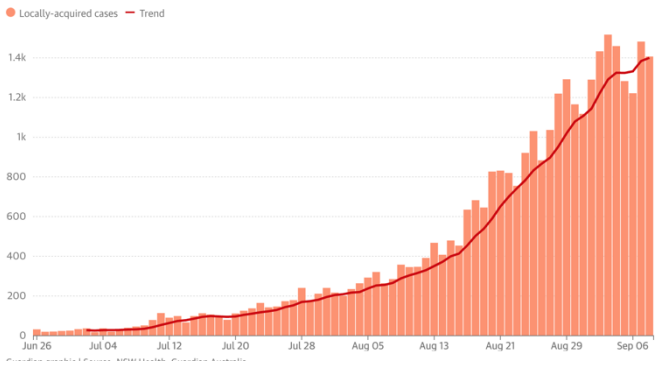

Figure 4 - Daily COVID-19 cases announced each day with trendline

IV. INTENTIONAL AND UNINTENTIONAL MISREPRESENTATIONS OF DATA

Reporting data on the number of positive cases, along with the number of tests conducted by location on a given date can be susceptible to misinterpretation by the consumer, whether intentional or unintentional, depending on its delivery.

For example, a steady number of positive cases being detected consistently for a few weeks without increasing beyond its deviation, may not signal a positive result. However, without knowing the virus's true R0 value (denoting contagiousness of a disease), the reproductive rate of the virus currently being encountered may be well below what it's truly capable of. This means that the policies put in place might well be effective in limiting the spread of the virus throughout the community. And because there is often a delay in the onset of symptoms, it is possible that the true reproductive rate of the virus now is under 1 (since the current positive cases are a result of infections occurring a number of days ago), meaning that for every one infected person, less than one other person is being infected, and hence the situation can actually be improving, contrary to what the data or visualisation may suggest. As a result, the announcement on the value of positive cases alone may not be a good indicator to assess progress of containment. This potential for misinterpretation is possible for all visualisations outlined in Section III, since the thematic maps simply uses the sum of cases or tests aggregated over a time to supply the range of its colour scale. Similarly, the bar chart in Figure 4 suffers from the same issue with using the sum of cases notified for each given day; although it does have the advantage of being able to at least visualise the rate of increase or decrease in the number of cases over time.

A bigger potential for misunderstanding of this data, particularly when presented in a thematic map, such as those visualized in Figures 1 to 3, is the notion of assessing a location's situation on the outbreak, based on the total number of tests conducted/positive cases. This is because locations with higher populations may be overrepresented in the number of positive COVID-19 cases within the state, when those localities might in-fact have a lower percentage of positive cases when measured against its population size. On the contrary, regional areas which may have an outbreak, will unintentionally be underrepresented in these visualisations mainly due to its expected lower number of COVID-19 cases. However, when compared against its total population size, these areas may in-fact have higher figures when cases are expressed as a percentage of its local population.

This is not to say that the visualisations outlined above are flawed nor is the data in its current state deemed unusable. However, there is potential in its current form that the data and its corresponding visualisations may cause misrepresentation on certain information being derived.

V. HOW THE DATA CAN BE VISUALLY REPRESENTED

Given how the data and some form of its visualisations can cause misunderstanding with its consumers, as discussed in Section IV, it's important to take those into consideration when designing other forms in which this can be visually represented.

One strategy, which can be considered as an important intermediate step for further visualisations, is through the creation of a derived attribute – "% Cases over Tests". This derived attribute can be calculated by dividing the aggregated number of cases against the total number of tests conducted, thereby creating a new ratio attribute, which takes into consideration the proportion of positive cases to the total number of tests performed for a given area. This addresses one of the criticisms mentioned in the previous section, regarding areas with high population numbers being overrepresented and low population areas being underrepresented, when it comes to assessing the severity of an outbreak in thematic maps shown in Figures 1 to 3.

It's also been discussed that maps have a unique advantage compared to other charts, in its ability to convey quantitative and/or qualitative data in-relation to geo-referenced information. In this regard, maps make for a great candidate for the visualisation of this dataset, given the geographic nature of attributes such as postcode, LGAs and LHDs. In particular, we are interested in using a symbol map to illustrate the number of cases, the percentage of positive cases in relation to the total tests conducted, and the Local Health

District (LHD) that each data point belongs to; where each data point represents a Local Government Area (LGA) placed in its location within the map.

As each data point represents an LGA, its size within the map will then be associated to the derived attribute we've created, "percentage of positive cases over tests conducted" for each given LGA, along with a text label of the LGA name written below it. On the other hand, the colour of the datapoint is then associated with the number of cases belonging to each given LGA. And while the number of cases may initially be thought to be better represented by the size of the data point and the percentage of positive cases to tests conducted better represented by colour, this would actually represent a problem in the visualisation. The problem is that LGAs with a low number of positive cases but actually have a high percentage of positive cases in-relation to the tests conducted, will be marked in the map with a small, albeit red, data point; creating the potential for the user to miss important information such as high community transmission in lower population areas, as demonstrated in Figure 5. This then circles back to our original problem, in which low population areas can be underrepresented, and higher population areas being overrepresented, due to a location's population density and size being linked to total case numbers.

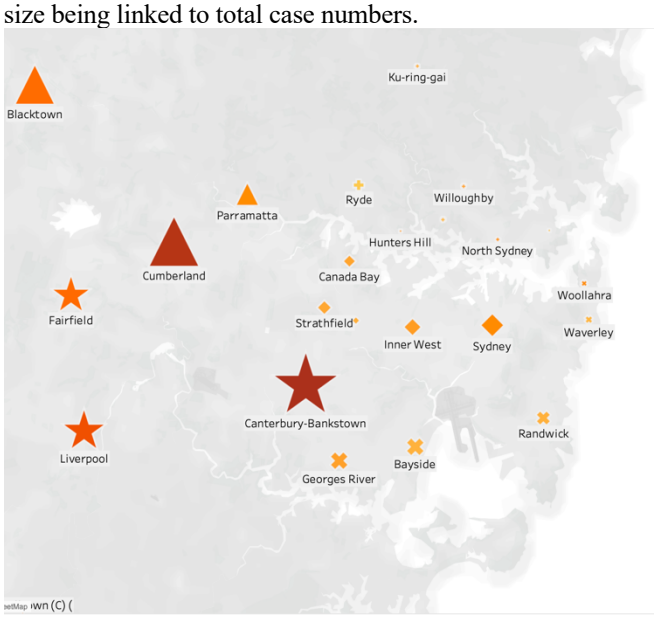

Figure 5 - Emerging areas of concern are less prominent when size is associated to number of cases

On the contrary, the decision to demonstrate the percentage of positive cases with respect to tests conducted with the retinal variable of size, allows the visualisation to demonstrate locations which may currently have low case numbers but are emerging to be areas of concern such as Woollahra and Waverley, demonstrated in Figure 6. While the number of cases is represented as colour, to allow users to see LGAs which still have high case numbers such as Cumberland and Canterbury-Bankstown. This decision can also be supported in that the retinal variable of size is best to show quantitative differences, such as the percentage of positive cases against number of tests conducted, and that the number of cases, while it's a quantity, can be argued to be better represented as a qualitative variable since the quantitative measure of interest is now the percentage, and this total case number is relegated to a qualitative measure of good to bad – which colour hue excels in.

Figure 6 - Emerging areas of concern are easier to identify through size

Finally, the retinal variable of shape will be used to distinguish collections of LGAs belonging to the same Local Health District (LHD). The inclusion of this attribute is to allow the consumer to identify LHDs which may require extra assistance, particularly when it encapsulates multiple LGAs which have a high percentage of positive cases, since this would signal that community transmission is quite prevalent. On the other hand, this would also allow for the identification of LHDs with potentially underutilised capacity in its hospitals, where LGAs in its scope have low community transmission and low number of cases, hence making it easier to see where patients from LHDs with no remaining capacity can be transferred to if needed.

Overall, the use of a symbolic map allows the consumer to achieve numerous elements within the modality of interaction. This includes associative memory, whereby the user can easily digest the information that South-West and West Sydney currently has high number of cases and also have a high percentage of positive cases (citing community transmission), simply by remembering the large and red colour datapoints from briefly viewing the map. This visualisation also aids in problem solving, through the identification of possible emerging areas of concern as previously discussed along with how different Local Health Districts might be able to cope with the situation in its LGAs. Lastly, the use of a symbolic map allows for special awareness due to the geographic nature of the data, allowing users to see the situation of neighbouring LGAs in-relation to the person's location.

VI. Symbolic Representation of the Visualisation

The visualisation described in Section V can be represented using the symbolic representations from Semiology of Graphics in Figure 7, below.

Within the symbolic representation, the use of maps within our visualisation is denoted by the x and y axis symbol with the text GEO. This type of imposition allows us to place our data points within a geographically aware special coordinate system. Four retinal variables are then used, two of which are of quantitative nature and are assigned to colour and size. The

reason as to why colour and size have been chosen to represent the two quantitative variables are due to their ability to demonstrate the numerical differences within them. Although as previously discussed, while colour may not be the best to show quantitative differences, it still does a good job in representing numerical values in the context of an ordered qualitative value. In the example of our visualisation, lower case numbers would be provided with a yellow hue, with the scale gradually turning into a red hue to denote worsening conditions. The third retinal variable of type shape is assigned a qualitative component, and this is used within our visualisation to group together LGAs which belong to the same LHDs. Finally, the fourth retinal variable is the text label, which allows the user to easily know the LGA name, particularly useful when the consumer is not familiar with some of the areas within the map.

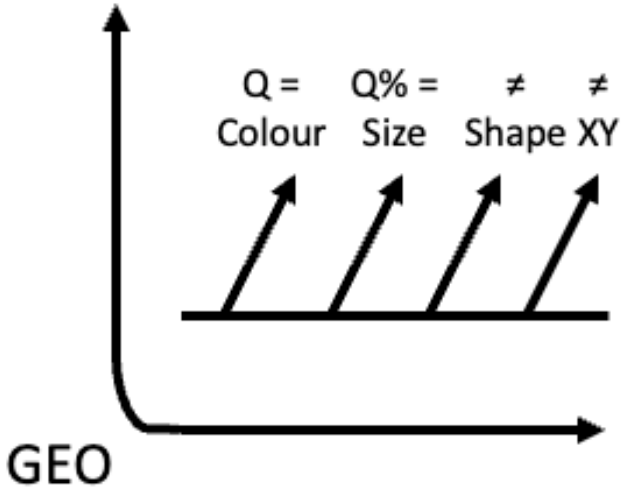

Figure 7 - Symbolic representation of the visualisation in Section V

VII. Deriving an Equivalent but Different Symbolic Representation

An equivalent, but different symbolic representation from the visualisation described in Section VI, can be derived in Figure 8, as below.

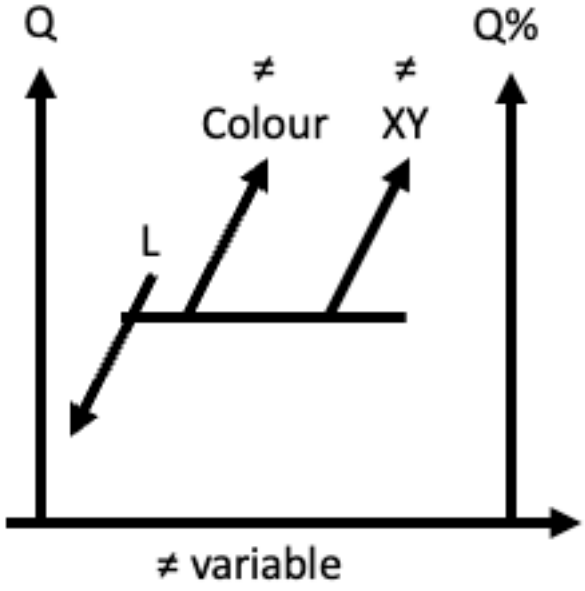

Figure 8 - Different symbolic representation of the one described in Section VI

In this alternative symbolic representation, we can see that the two quantitative variables, previously assigned with the retinal variable colour and size, are now represented by its own y-axis, thereby resembling a slope chart. This is also reinforced by the x-axis being marked with the inequality sign, to mark that the values alongst its direction are the y-axis of the two quantitative variables. The qualitative component previously shown through the retinal variable of shape, is now changed to be represented through the different colours of the line within the slope chart. As a result, lines along the slope chart can be classified as being in the same category based on its colour. Likewise, the lines within the slope chart have the symbolic representation of the downward arrow marked with "L". The qualitative component shown by the retinal variable of type text from the original representation, however, is retained in order to provide the same function of labelling each line to allow the user to determine the corresponding LGA.

It's worth noting that the two symbolic representations, Figure 7 in Section VI and Figure 8 in this section, have equivalency, although conveys it through different means. The symbolic representation of the original visualisation utilized the geo-spatial feature of maps allowing it to place the different data points to its corresponding geo-related information – in this case LGAs. Consequently, it doesn't have an axis to convey further information as this is replaced with latitude and longitudinal co-ordinate system, and as a result, has to rely on retinal variables for both quantitative and qualitative data. On the contrary, our alternative symbolic representation uses a completely different way to illustrate the same data. By not using a map, it is able to place the quantitative variables, previously restricted to the retinal variables of colour and size, to its own y-axis that a slope chart is capable of. Since the geo-spatial mapping is now lost, this has been replaced by illustrating each data point (being the LGA) as a line along the chart. The lines are then colour-coded based on the category the data point belongs to, where it was previously marked based on shape. The text labels, then serves the purpose of allowing the user to know which LGA is represented by which line.

Overall, the two symbolic representations are equivalent, with the change from a geo-spatial co-ordinate system to a slope chart allowing for direct input of the quantity variables within the two y-axes.

VIII. Visualisation of the Alternative Symbolic Representation

The illustration shown in Figure 9, is an example of an alternative visualisation derived from the symbolic representation in Section VII. In this alternative visualisation, a slope chart is used to provide comparison between different Local Government Areas, providing a picture on how each performs with respect to the two quantitative measures of number of cases and the percentage of positive cases against tests conducted.

While the visualisation is quite different, it still makes use of the same data. Similar information that can be derived with both the symbol map and this slope chart are the identification of LGAs with severe outbreaks, but also finding the LGAs which may have low case numbers but suggests that there is high local transmission through the percentage of positive cases over tests conducted. The main difference, however, is that with a symbol map, the user is required to scan through

the entire map to ensure that no LGA is missed, and that comparing the quantitative values through shape and size is harder relative to using a chart with quantitative measures on its y-axis such as the slope chart. As such, while the same information can be derived between the symbol map and the slope chart, the slope chart does have the advantage of being able to easily compare and rank the different LGAs against each other with respect to either the number of cases or the percentage of cases against tests conducted.

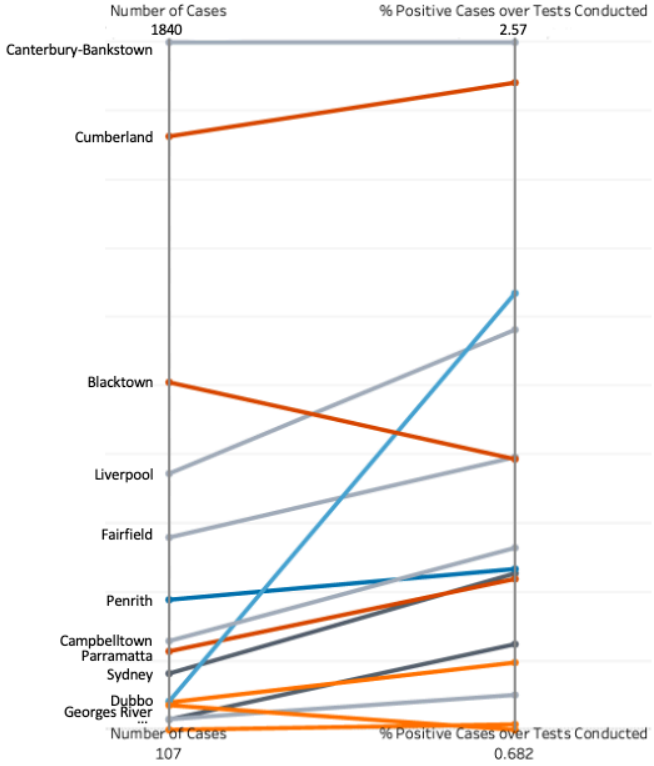

Figure 9 - Alternative visualisation derived from the symbolic representation in Section VII

Another advantage of this chart is through the extra dimension that the slope of each line provides. The line represented by each LGA allows the user to see the relationship between the two variables, such as the Canterbury-Bankstown area consistently topping both the number of cases along with the percentage of positive cases over tests conducted. Based on the illustration above, the consumer can also easily see that while Dubbo has one of the lowest case numbers, it is actually above average in the state with regards to the positive rate from tests, potentially signalling high local transmissions. Blacktown on the other hand, fairs a lot better despite having the third highest number of cases.

While it does have its advantages, the consumer does lose the geographical sense of where each LGA are in-relation to each other, whether neighbouring LGAs tend to present the same problems, or patterns with the spread of the virus that might reflected on LGAs consisting of similar demographic. This type of geo-related information that is tied to the LGA attribute is something the slope chart cannot demonstrate.

Overall, the slope chart represents a solid alternative to the original symbolic representation of the symbol map. While both visualisations are able to convey most of the same information, each style does tend to accentuate some information that the other is unable to convey. For symbol maps, it's the spatial awareness of each LGA and LHD from the geographical point-of-view, whereas slope charts have the ease of being able to rank each LGA with respect to the different quantitative measure. Nonetheless, the symbolic representations between the two have equivalency.

REFERENCES

[1] Hunt, G., 2020. *First confirmed case of novel coronavirus in Australia*. [online] Department of Health. Available at: <https://www.health.gov.au/ministers/the-hon-greg-hunt-mp/media/first-confirmed-case-of-novel-coronavirus-in-australia> [Accessed 9 September 2021].

[2] Data.NSW. 2021. *NSW COVID-19 tests by location*. [online] Available at: <https://data.nsw.gov.au/search/dataset/ds-nsw-ckan-60616720-3c60-4c52-b499-751f31e3b132/details?q=> [Accessed 7 September 2021].

[3] Data.NSW. 2021. *NSW COVID-19 cases by location*. [online] Available at: <https://data.nsw.gov.au/search/dataset/ds-nsw-ckan-aefcde60-3b0c-4bc0-9af1-6fe652944ec2/details?q=> [Accessed 7 September 2021].

[4] NSW Government. n.d. *Find the facts about COVID-19*. [online] Available at: <https://www.nsw.gov.au/covid-19/find-the-facts-about-covid-19> [Accessed 9 September 2021].

[5] Evershed, N., Nicholas, J. and Ball, A., 2021. *Covid-19 Australia data tracker: coronavirus cases today, trend map, hospitalisations and deaths*. [online] The Guardian. Available at: <https://www.theguardian.com/australia-news/datablog/ng-interactive/2021/sep/09/covid-19-australia-tracker-map-data-cases-today-coronavirus-tracking-stats-live-data-update-by-state-melbourne-regional-victoria-vic-sydney-nsw-how-many-new-active-case-numbers-statistics-deaths-death-toll> [Accessed 9 September 2021].