# Introduction

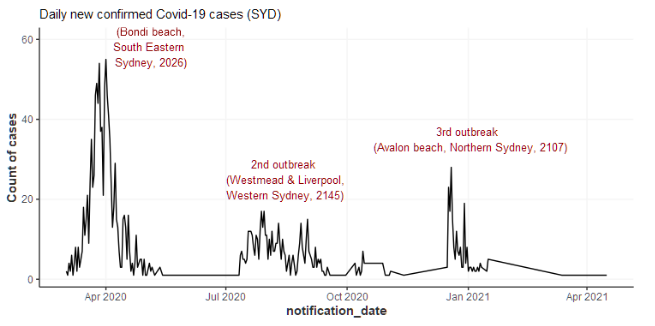
Since the outbreak of Covid-19, governments worldwide have relied on contact tracing and extensive testing to curb local transmission of the disease. These two measures work well if 1) the outbreak is discovered earlier so that the origin can be identified; 2) the number of confirmed cases is limited so that the spread is trackable; and 3) the presence of asymptomatic carriers is rare. However, in many parts of the world, it is improbable for all these three conditions to be met simultaneously. In most cases, local transmission cannot be contained without extreme measures such as extended lockdown periods. These lockdowns would result in a heavy toll on the economic performance. Considering this, effective methods to select only a small subset of the geography or population to enter lockdown or perform group testing can be constructive for the government to strike a balance between containing the virus and maintaining the economic output.

In this project, we explore the merit of using geographic clustering based on homophily and gravitational models to identify a small collection of the local areas for the authority to impose more substantial measure, such as lockdown and group testing, to effectively curb the local outbreak of an epidemic disease such as Covid-19, while minimising the impact on economic activities.

## The spread of Covid-19 in Sydney/Australia

The first positive case for Covid-19 in Australia was reported on 25th January 2020 from incoming travelers from China. One month later, on 27th February, Australian Prime minister Scott Morrison announced the Australian Health Sector Emergency Response Plan - 4 days before the first case of community transmission was reported in the state of NSW on 2nd March. Since then, Sydney, the capital city of NSW, has experienced three waves of the locally spread virus within its regional proximity – namely the “Bondi Beach Cluster”, the “Western Sydney Cluster” and the “Northern Beach Cluster”. A detailed Covid-19 related news timeline has been captured by Deborah (2020) – from which we have presented a visualisation in *Appendix A.*

{Chart index} shows the daily count of new cases (locally acquired) in the Sydney area. The three distinct waves can be identified in different periods. The magnitude of the outbreak is relatively small compared to what happened in the US and Europe. This creates an opportunity to investigate the spatial diffusion of the virus in a somewhat less “noisy” environment. Because the source of the virus can be narrowed down to the few suburbs that recorded the spike of new cases at the beginning of the spread – the four-digit postcode (a.k.a postal area or POA or POA\_NAME16) for each of the three clusters’ origins are shown in the red text annotation below.

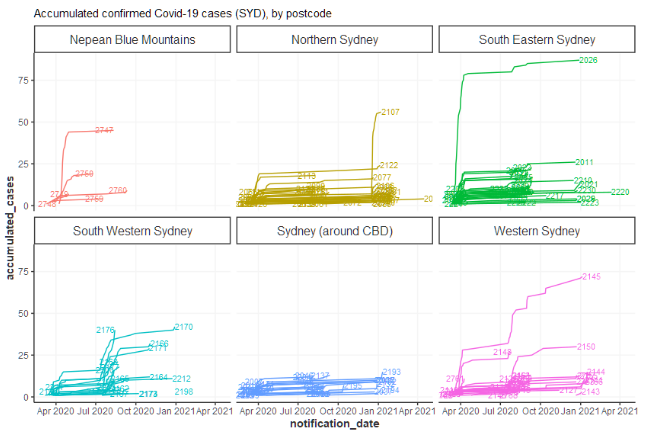


{Chart index} shows the accumulated cases by local health districts. The early spikes across all regions were led by the South Eastern Sydney cluster (Bondi beach). The following rise of Western Sydney and South Western Sydney cases is more gradual, accompanied by elevations in cases in other regions. The final cluster in Northern Sydney appears more abrupt and relatively isolated compared with the first two clusters.

Line chart

Description automatically generated

{Chart index} shows the more granular accumulated cases by postcode in each local health district from {Chart index}. Each district appears to have a few “leading” postcodes that had a much higher number of cases than the rest of the areas, except for the area around Sydney CBD.



{Chart index} shows the accumulated cases as choropleth by postal area (closest shape file related to postcode, which is what the Covid-19 case data is reported from) in the map of Sydney. The clustering pattern is evident. The more saturated red colour presents the locations of the three clusters. The Western Sydney cluster seems to have spread more widely than the other two beach clusters. It is conceivable that the flows of people to and from the beaches are more scattered, thus difficult to predict, whereas the spread of the virus in the Western Sydney area are more clustered/concentrated.

Chart

Description automatically generated

{Chart index} shows the accumulated cases over time in a constant scale of filling colour in the choropleth map. Interestingly, the Western Sydney cluster might originate from the further western area of Cambridge Park, before it spread out to the middle of the Western Sydney area, a local transportation and employment hub.

Diagram

Description automatically generated with medium confidence

## The motivation of the gravity model

Following the previous section, {chart index} overlays the public transportation network on the choropleth map. It is conceivable that the virus might have been transmitted along public transportation lines (Metro and Train in Sydney are like subways in the US).

Map

Description automatically generated

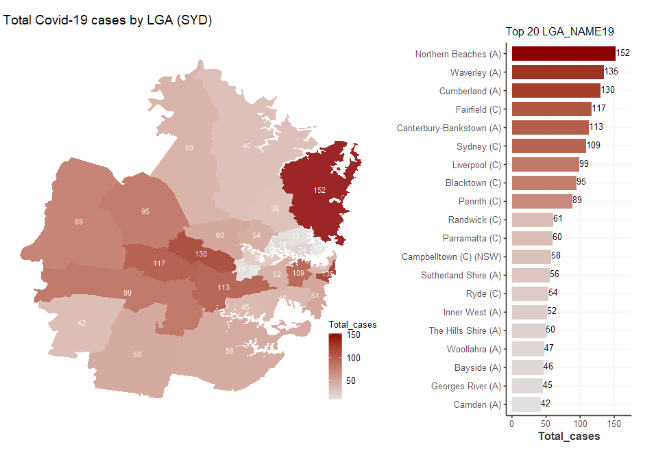
However, the public transport lines are also associated with dense populations and geographic hubs (schools, hospitals, supermarkets, etc.). {chart index} shows the number of supermarkets/groceries in each postal area. Compared with above, the Western Sydney area with more saturated green colour in {chart index} appears to coincide with the areas with the darker grey colour from {chart index},

**A picture containing diagram

Description automatically generated**

These observations inspire the use of the Gravity Models to measure the level of “connectivity” between geographical areas, using a series of “point of interest” (the number of schools, supermarkets, public transports, etc.) as the “**mass**”; and the physical distance or travel time as the “**distance”** to model/predict the spread of the virus (see details in section 2.1).

It is noted that postcodes (or postal areas) with larger sizes tend to appear more prominent in the choropleth map. This would lead to a bias in the visualisation, overlooking the more densely populated areas with smaller sizes, which could have collectively accumulated many cases but do not have a more saturated colour due to the split of total cases into smaller sub-areas. To verify this, {chart index} shows the choropleth map filled with the aggregated number of Covid-19 cases by Local Government Area (LGA). It is evident that the spread of the virus in the western areas is more severe and spread out, while the collective number of cases from the northern and the south-eastern regions are relatively lower and comparable.



## The motivation of the homophily model

*Need a map showing demographics, % of ethnic groups, etc., to demonstrate the presence of homophily in the geographic area.*

# Methodology

## The Gravity Model

<https://oxfordre.com/economics/view/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-327#acrefore-9780190625979-e-327-bibItem-0048>

The gravity model is an intuitive way to measure the volume of trade tween two countries in economics. It follows the analogy to Newton’s Law of Gravity, which states that the force of attraction between two bodies is proportional to the product of their masses and inversely proportional to their distance squared. In the context of international trade, the “mass” becomes the Gross Domestic Product (GDP) of the two countries, and the distance is the physical geographic distance.

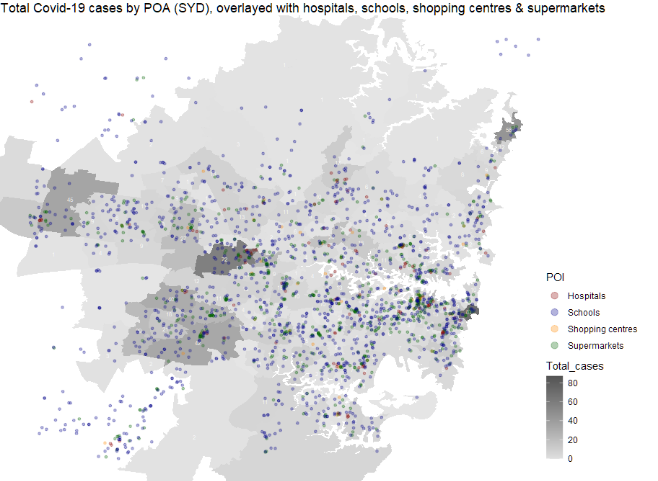
According to Baier and Standaert (2020), the empirical estimation of the gravity equation above consistent with the “naïve specification” above where the coefficient estimates of power term on GDP were close to unity, the elasticity of trade with respect to bilateral distance was negative. Estimation from the “naïve specification” accounts for a reasonable amount of the observed variation in international trade.

The empirical success of the gravity model has led to the development of a variety of theoretical models to underpin the gravity equation - – such as the structural gravity model based on the Multi-Country Ricardian Model (Eaton and Kortum, 2002), the structural gravity model with heterogeneous firms from Melitz (2003), Chaney (2008) and Redding (2001).

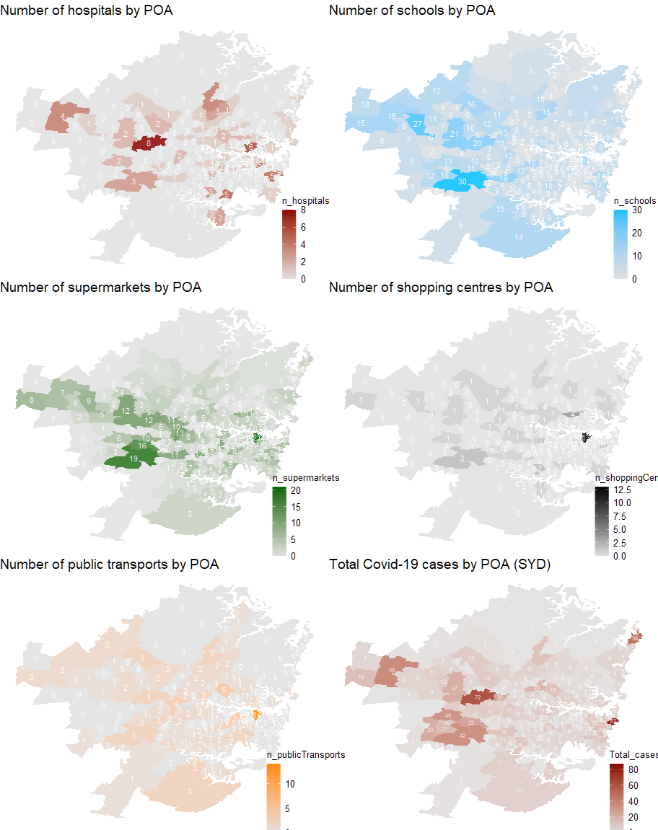
In the context of our geospatial analysis, the gravity between two Postal Areas is defined as:

Where the POIs can be hospitals, schools, supermarkets, shopping centres, and public transports. The calculated gravity scores are scaled or normalised for the purpose of visualisation on the map and the spatial clustering algorithm. The gravity score can be viewed as an improved measure over physical distance to represent the “connection” between two geographical areas.

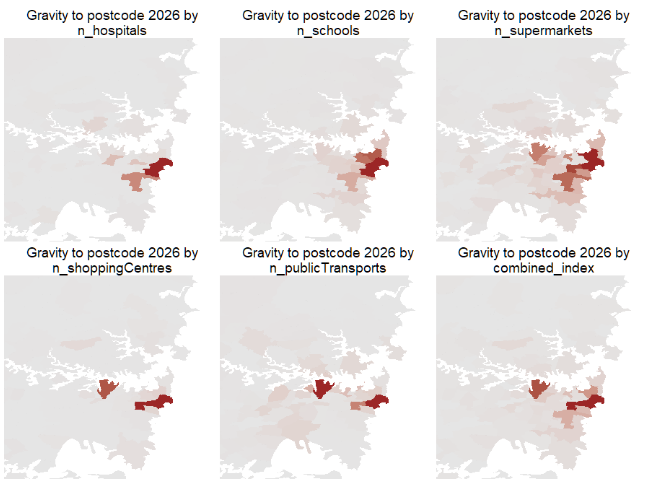
{Chart index} shows the geo-coordinates of POIs (hospitals, schools, shopping centers, and supermarkets) in the Sydney area, overlaying the choropleth map of accumulated Covid-19 cases.

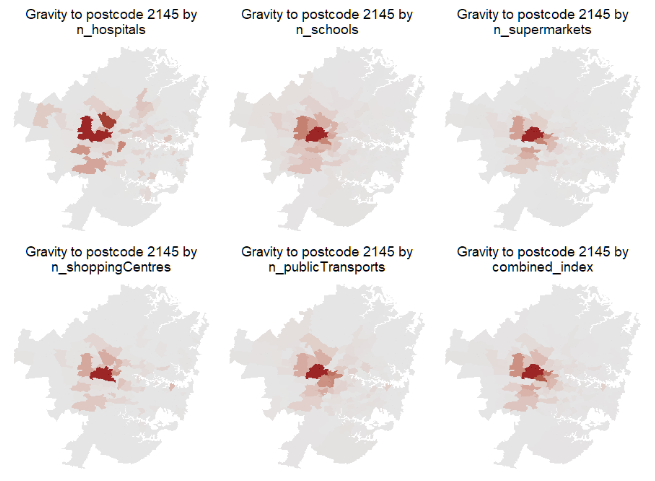


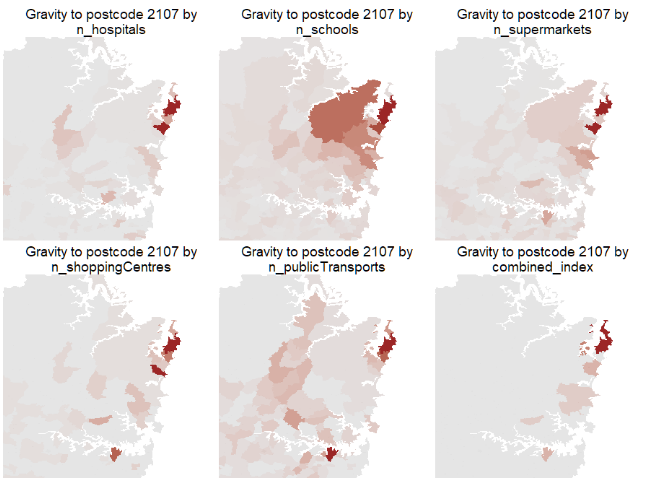
We used the [sp](https://github.com/cran/sp) package in R to convert the geocoordinates to the count of POIs in each Postal Area. This is shown in {Chart index} below.



As an illustration, we can calculate the gravity score from each Postal Area to the likely origin of the three Covid-19 clusters in Sydney, namely the “Bondi Beach Cluster” (2026), the “Western Sydney Cluster” (2145), and the “Northern Beach Cluster” (2107), based on the “naïve” gravity equation. The calculated result is visualised in {Chart index} below:

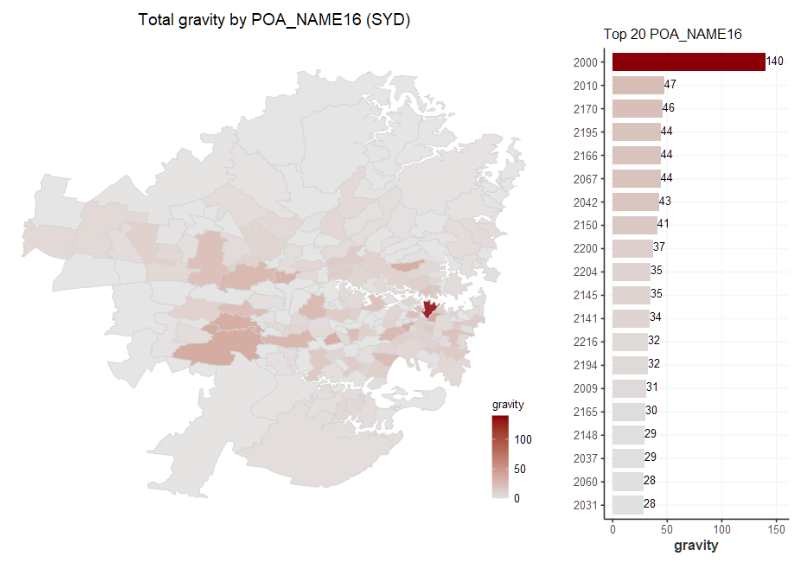






To combine the gravity scores from multiple specification of “mass”, we take the dot product of the normalised count of “point of interest”, so the formula becomes:

The combined gravity map is shown in {Table index} below. It can also be used as an additional feature in any clustering methods combined with the census/homophily features.



## The Homophily Principle

<https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.27.1.415?journalCode=soc>

## Geospatial Clustering

<https://towardsdatascience.com/geospatial-clustering-kinds-and-uses-9aef7601f386>

* Partition clustering (K-means)
* Hierarchical clustering (Agglomerative clustering)
* Density-based clustering (DBSCAN)

# Data

## Covid-19 confirmed cases SYD/AU

The number of Covid-19 cases by notification date and postcode, local health district, local government area, and likely source of infection is sourced from the official state government website:

<https://data.nsw.gov.au/data/dataset/nsw-covid-19-cases-by-location-and-likely-source-of-infection/resource/2776dbb8-f807-4fb2-b1ed-184a6fc2c8aa>

For this report, we have filtered out Covid-19 cases acquired overseas to focus on community transmission of the virus.

## Geographic location data SYD/AU

{Table index} below summarise the source of addresses and geographic coordinates of various “points of interest” in Sydney.

|  |  |  |
| --- | --- | --- |
| Point of interest (POI) | Source of address | Geographic coordinates |
| Hospitals | [Australian Institute of Health and Welfare](https://www.aihw.gov.au/) | [Australian Institute of Health and Welfare](https://www.aihw.gov.au/) |
| Schools (Primary & Secondary) | [Australian Schools List](https://asl.acara.edu.au/home) | [Australian Schools List](https://asl.acara.edu.au/home) |
| Public transport (train, metro, light-rail, ferry stations) | [Transport NSW official website](https://transportnsw.info/routes/train) | Google Geocode API |
| Shopping centres | [Wikipedia – List of shopping centres in Australia](Wikipedia%20–%20List%20of%20shopping%20centres%20in%20Australia) | Google Geocode API |
| Supermarket/Groceries | [Australian yellow pages](https://www.yellowpages.com.au/find/supermarkets-grocery-stores/sydney-nsw-2000) | Google Geocode API |

Additionally, the travel time between each Postal Areas is gathered from the Google Distance Matrix API (through the interface of the `[ggmap](https://github.com/dkahle/ggmap)` package in R).

## Suburb demographic data SYD/AU

{Table index}

## SSC to POA mapping

The Australian Bureau of Statistics (ABS) has an official webpage that explains the linkage of various definitions of geographic areas such as the Postal Area (POA) and State Suburbs (SSC) in Australia:

<https://www.abs.gov.au/websitedbs/censushome.nsf/home/factsheetsnas?opendocument&navpos=450>

From the following link, we obtain the mapping of each Postal Area (POA) and State Suburb (SSC) to the most granular geographic unit called “mesh blocks”: <https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.003July%202016?OpenDocument>

From there, we aggregate the SSC level data to the less granular POA level for the analysis.

# Model Evaluation

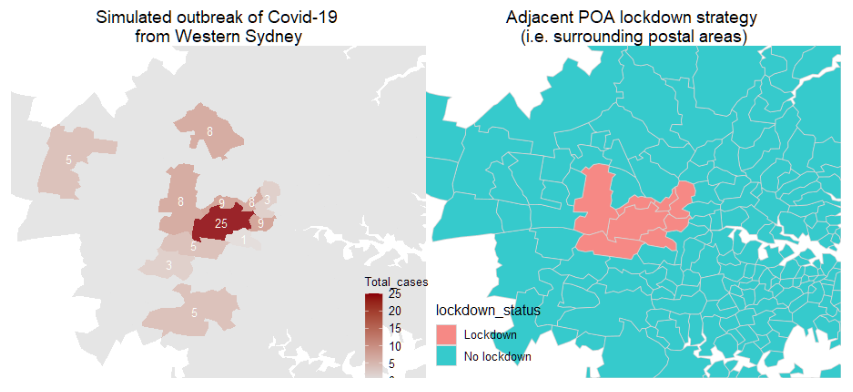
*Evaluation metric time series, e.g. conditional probability of positive case*

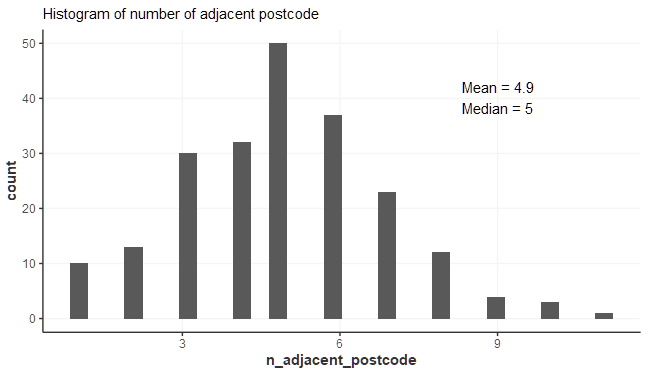
# Model outcome

*Show cluster maps by different specification, a matrix of clustering method and feature selected (and maybe # of PC selected)*

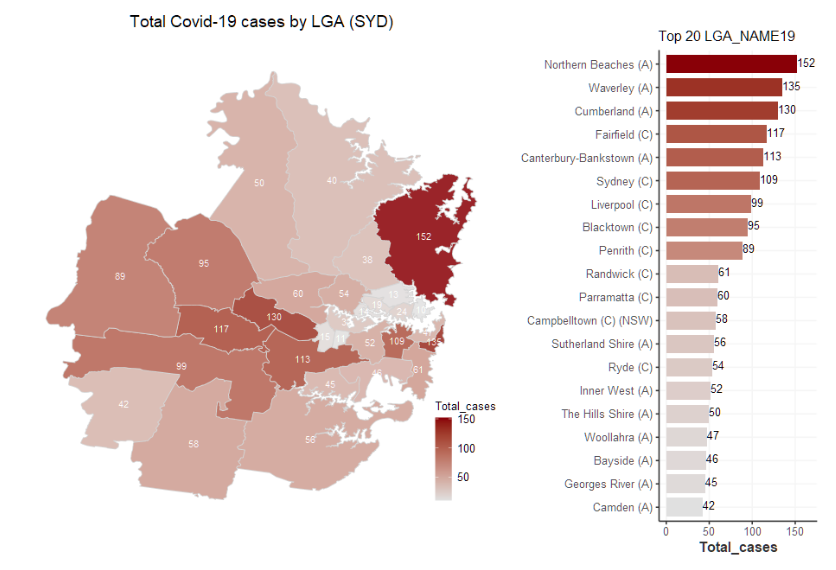
## Benchmark:

**Adjacent POA lockdown**





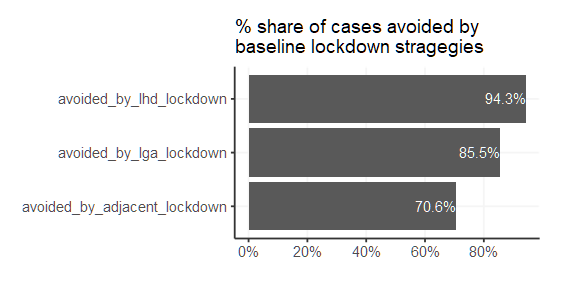
**Entire LGA lockdown**



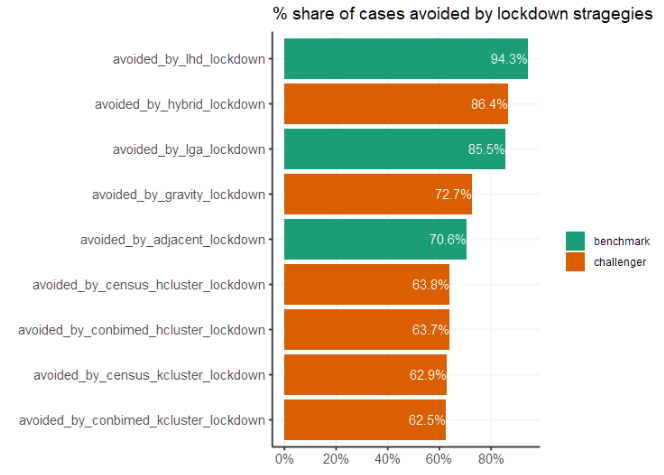
**Entire LHD lockdown**



**Benchmark**



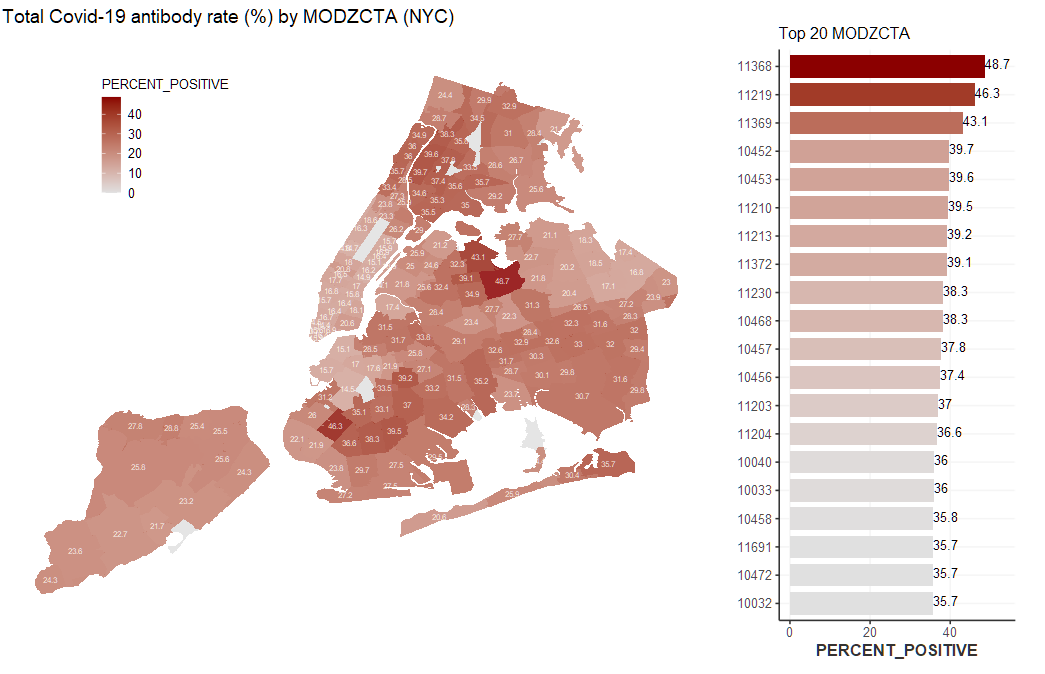
## Challenger



# Conclusion

# Further research

The methodology applied in the report can be extended to other geographic areas suffering from the Covid-19, especially with some data from the United States where the number of cases is more significant. The US data is also available in more detail, such as the percentage of antibody rate for New York City by Modified Zip Code Tabulation Areas (MODZCTA) as shown in {Chart index}.



Further analysis based on the US data is out of the scope of this report due to the submission deadline. We hope to further the research to broader geography after receiving feedback from the TAs.

# Table of reference

Bindiya Varghese and K. Poulose Jacob (2014), Spatial Clustering Algorithms – An Overview, Asia Journal of Computer Science and Information Technology, January 2014, sourced from: <https://www.researchgate.net/publication/235605835_Spatial_Clustering_Algorithms-_An_Overview>

Scott Baier and Samuel Standaert (2020), Gravity Models and Empirical Trade, Oxford Research Encyclopedias, source from: <https://oxfordre.com/economics/view/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-327#acrefore-9780190625979-e-327-bibItem-0048>

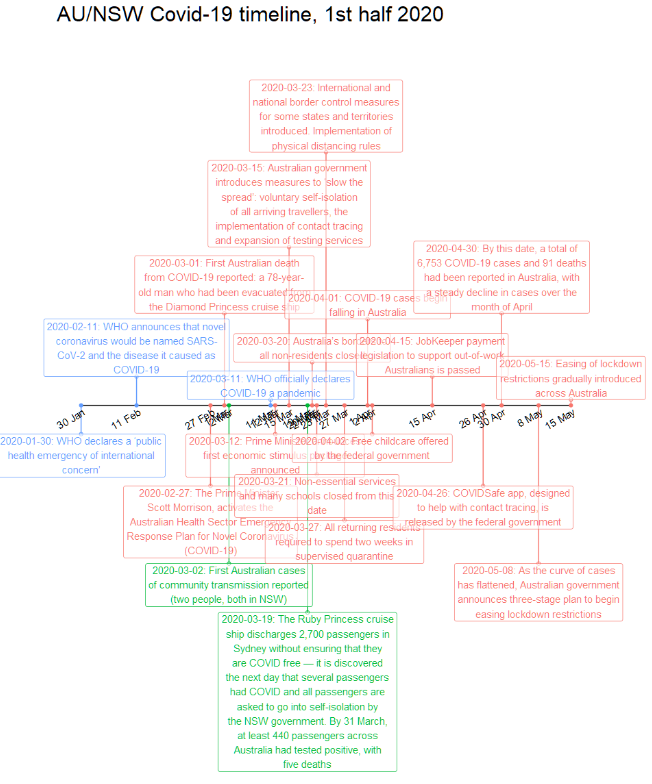
Vadim A. Karatayev, Madhur Anand, and Chris T. Bauch (2020), Local lockdowns outperform global lockdown on the far side of the COVID-19 epidemic curve, Proceedings of the National Academy of Science of the United States of America (PNAS), September 2019, sourced from: <https://www.pnas.org/content/117/39/24575>

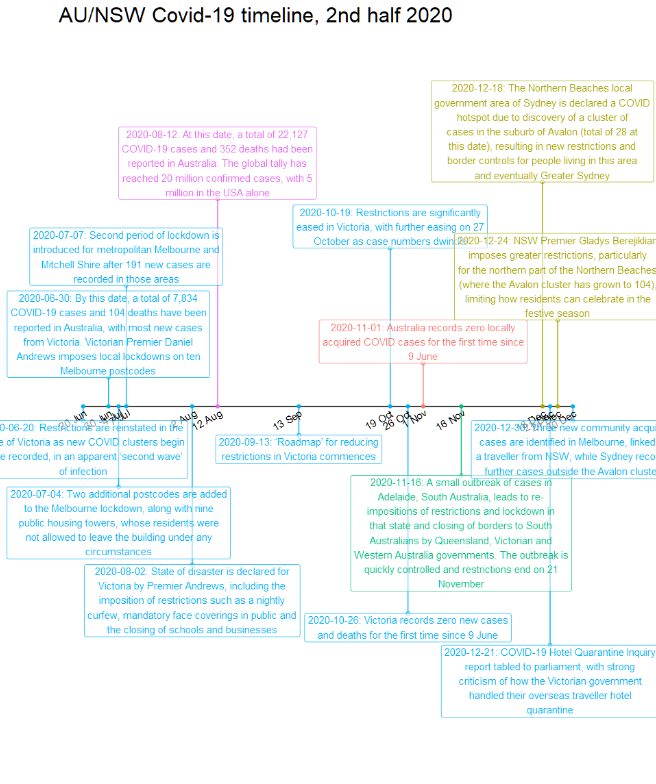
# Appendix

## Appendix A – Covid-19 related news timeline in Australia, 2020

Source: <https://deborahalupton.medium.com/timeline-of-covid-19-in-australia-1f7df6ca5f23>

Visualisation:





## Appendix B

## Appendix C