Gravitational homophily clustering to curb the outbreak of Covid-19

Yanjun Liu and Zhikang Yu

Department of Computing, Georgia Institute of Technology

Abstract

It's been more than one year since the outbreak of COVID-19, and many nations in the world are still having trouble managing the spread of the virus and at the same time keeping the economy running in their countries. One of the most critical aspects of managing the spread is to define lockdown strategies that are both effective and efficient. In this project, we have explored the merit of using geographic clustering based on homophily and gravitational analysis to identify a small collection of the local postal areas for the authority to impose vital 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. In particular, we examined three innovative approaches of lockdown based on 1) gravity index; 2) homophily clustering; 3) hybrid approach. The result from these three approaches outperforms a baseline strategy of "naïve" neighborhood lockdown by a significant margin. We also performed sensitivity analysis around the lockdown period. The result suggests that the effectiveness of the lockdown increases shapely at first but flattens after 14 days. The study from this paper should be insightful for policymakers to achieve a higher effectiveness/cost ratio in their virus management process.

Keywords: Covid-19; Clustering; Gravitational Model; Emerging

Github: https://github.com/fredyuu/cda_proj

Table of figures

1 Introduction

Since the outbreak of Covid-19, governments globally have relied on contact tracing and extensive testing to curb local transmission of the disease. These two measures work well if 1) the origin of the outbreak can be identified; 2) the trajectory of the spread is trackable, and 3) the presence of asymptomatic carriers is rare. Unfortunately, it is improbable for all these three conditions to be met simultaneously in many parts of the world. 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.

1.1 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 the Appendix.

Figure 1 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.

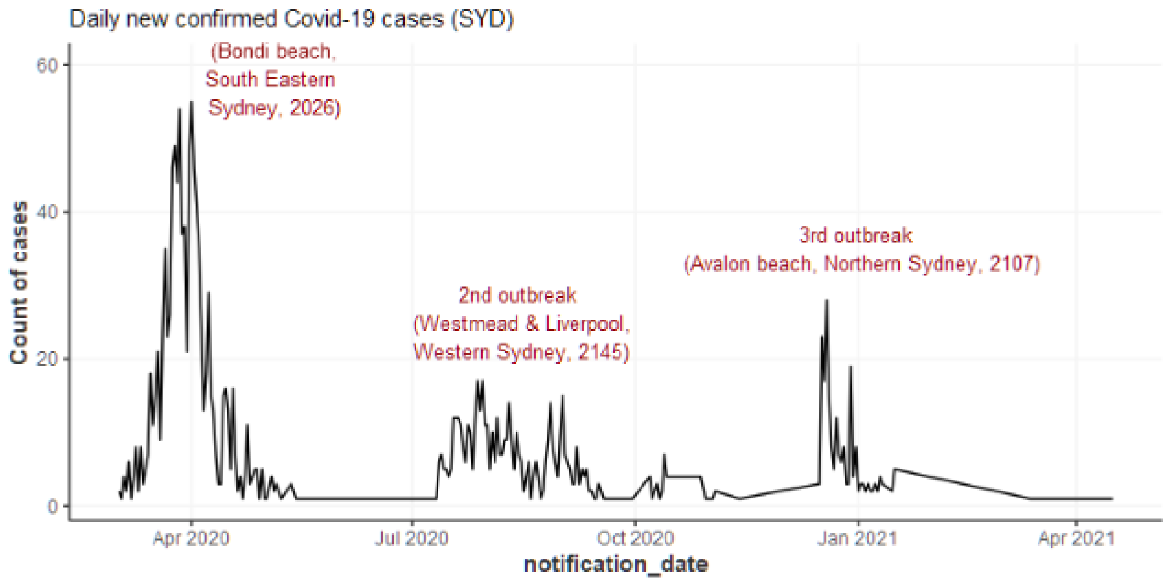


Figure 1: Daily new confirmed Covid-19 cases (SYD)

Figure 2 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.

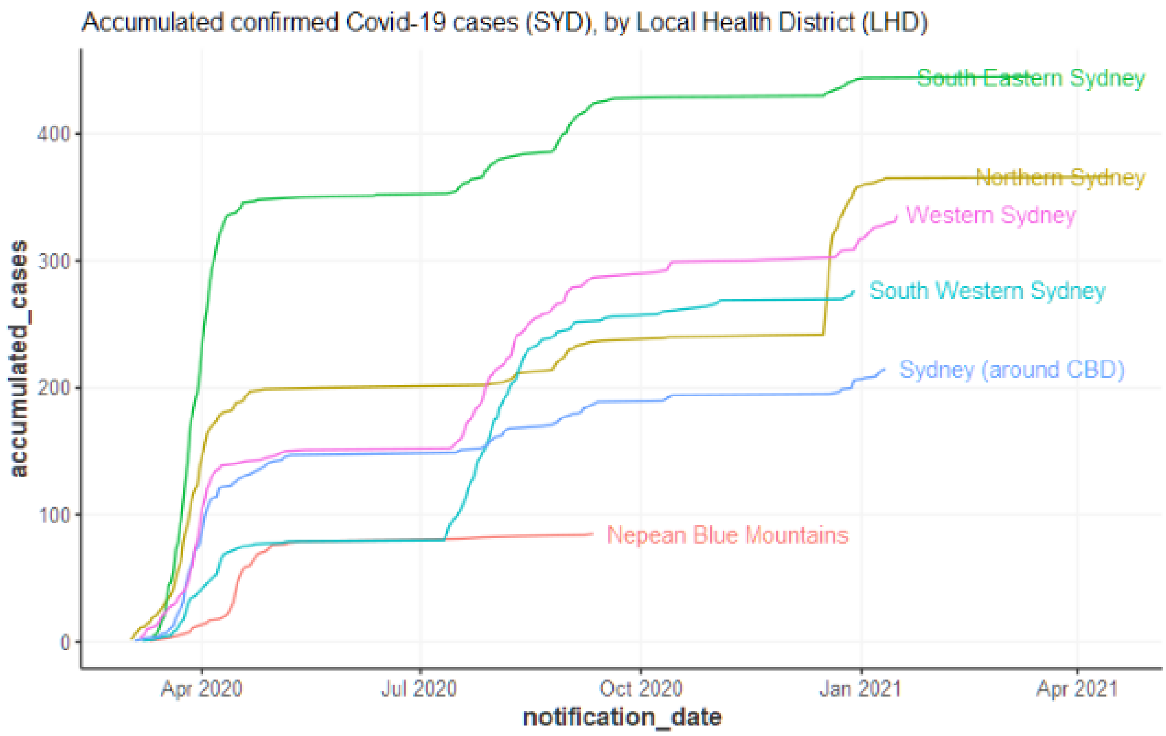


Figure 2: Accumulated confirmed Covid-19 cases (SYD) by Local Health District (LHD)

Figure 3 shows the more granular accumulated cases by postcode in each local health district from Figure 2. Each district appears to have a few “leading” postcodes with a much higher number of cases than the rest of the areas, except for the area around Sydney CBD.

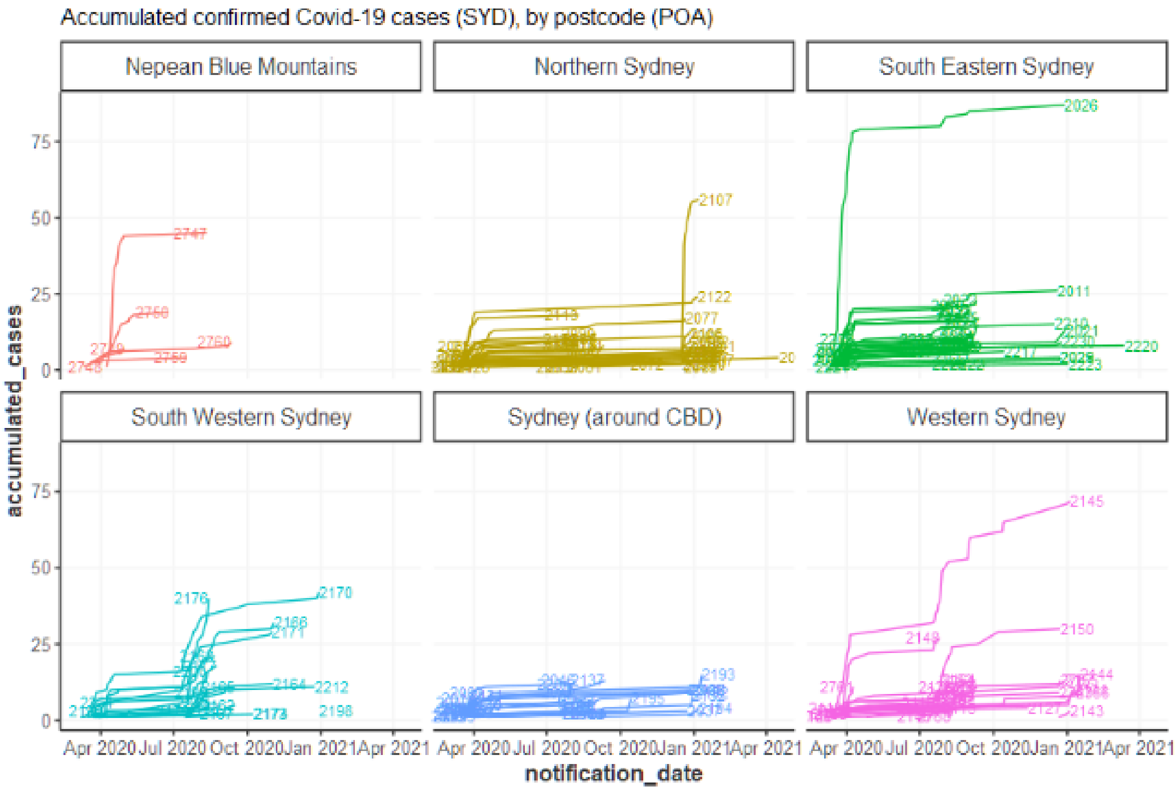


Figure 3: Accumulated confirmed Covid-19 cases (SYD) by postcode (POA)

Figure 4 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.

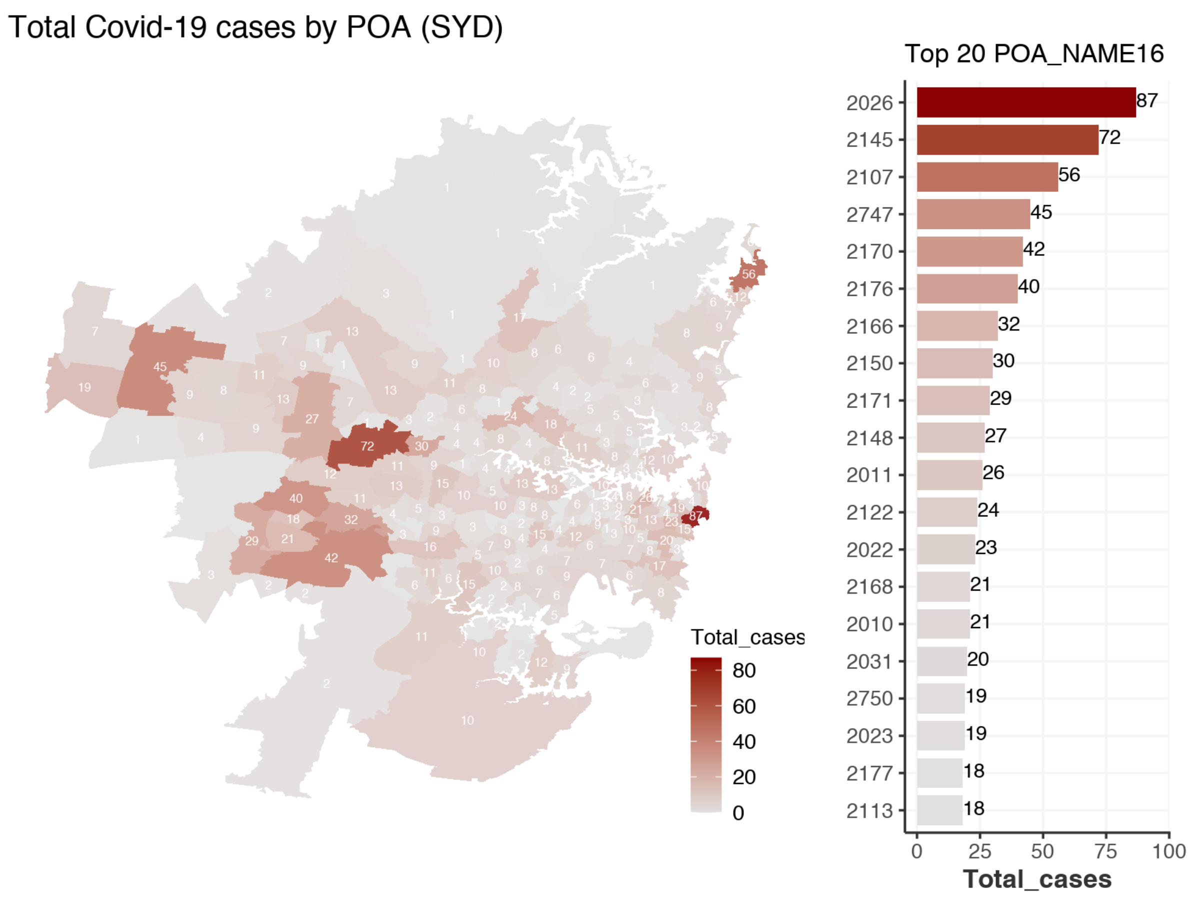


Figure 4: Total Covid cases map by POA (SYD)

Figure 5 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.

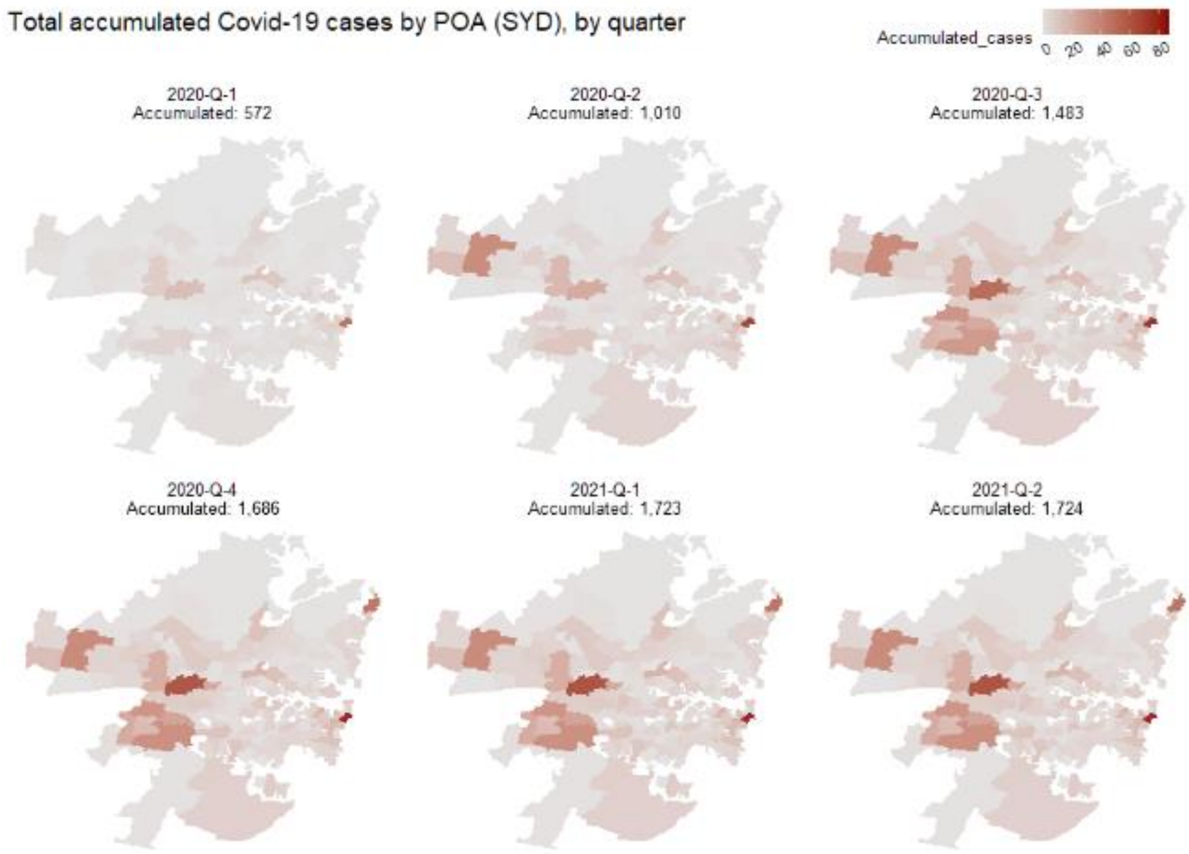


Figure 5: Total accumulated Covid-19 cases by POA (SYD) by quarter

1.2 The motivation of the gravity model

Following the previous section, Figure 6 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).

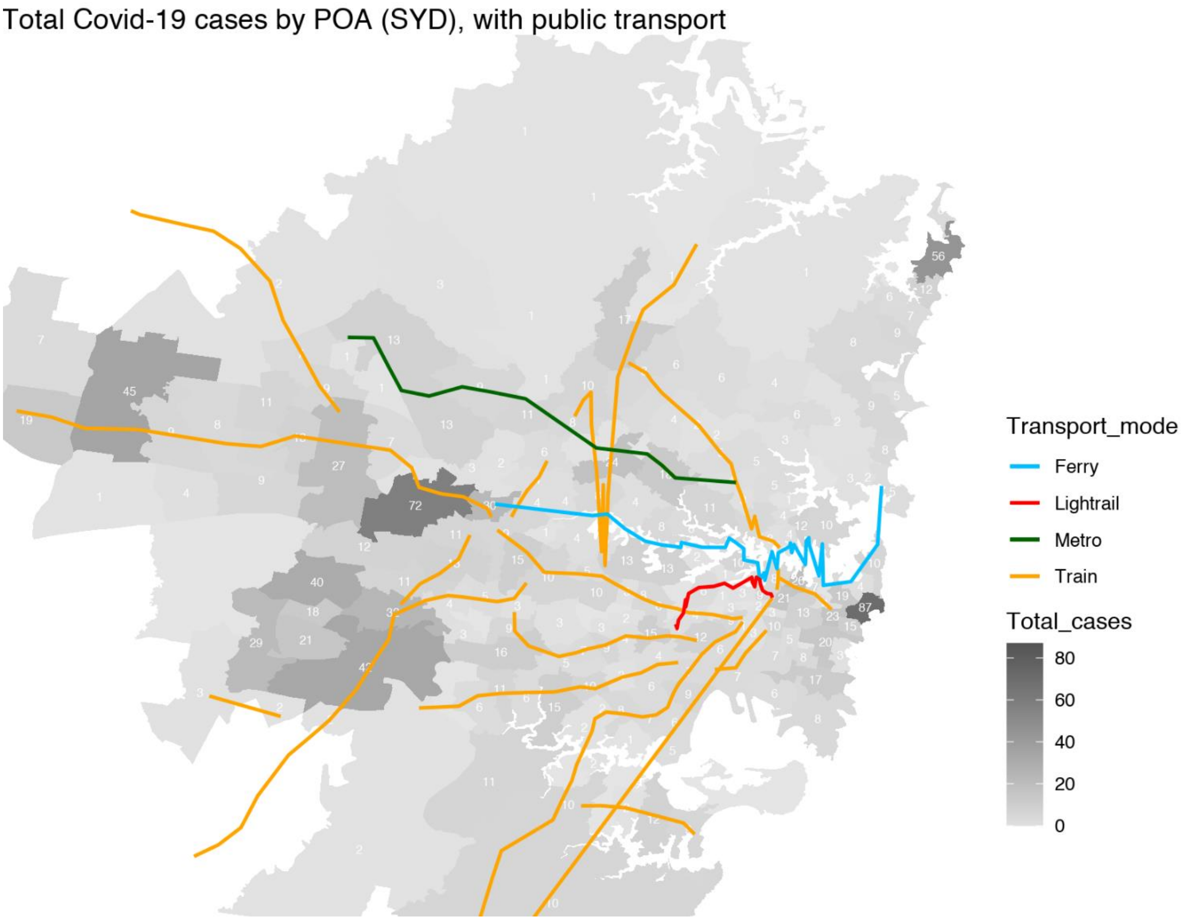


Figure 6: Total Covid-19 cases by POA (SYD) overlayed by public transport

Notably, the public transport lines are also associated with dense populations and geographic hubs (schools, hospitals, supermarkets, etc.). Figure 7 shows the number of supermarkets/groceries in each postal area. Compared with the above, the Western Sydney area with more saturated green colours coincides with the areas with the darker grey colour from Figure 6.

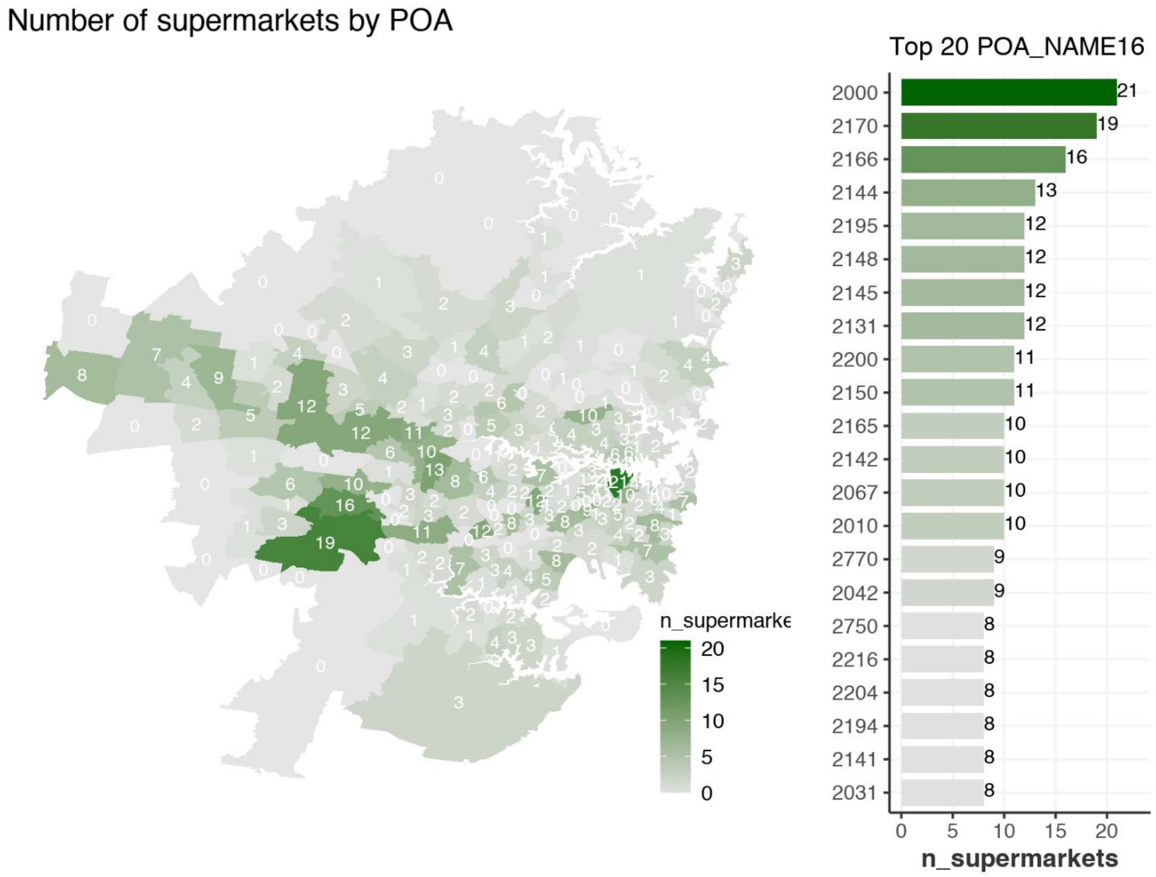


Figure 7: Number of supermarkets by POA

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).

1.3 The motivation of the homophily model

The gravity model measures geographical connections between different suburbs in a locality-driven method, which does not explain how some local transmission cases happened between suburbs that are not geographically close or connected.

Here is where homophily comes in. People tend to associate and bond with similar others (Lauw et al. 2020). It has been observed in many aspects of society, including race & ethnicity, sex & gender, age, religion, education, occupation, social class, and interests, etc. Suburb selection reflects one's social profiles and is a good choice for considering homophily in the social context.

The benefit of applying homophily clustering on suburb geographic data is that it no longer limits lockdown suburb selection to be geographics-oriented but rather considers local transmission from a social interaction point of view. This approach can be effectively used to explain why cases can transmit to suburbs that are not geographically close.

2 Methodology

2.1 The Gravity Model

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.

$$Trade\ volume = G * \frac{GDP_x * GDP_y}{Distance_{x-y}^2}$$

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:

$$Gravity = \frac{Number\_of\_POIs_x * Number\_of\_POIs_y}{Distance_{x-y}^2\ (or\ Travel\ time_{x-y}^2)}$$

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.

Figure 8 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.

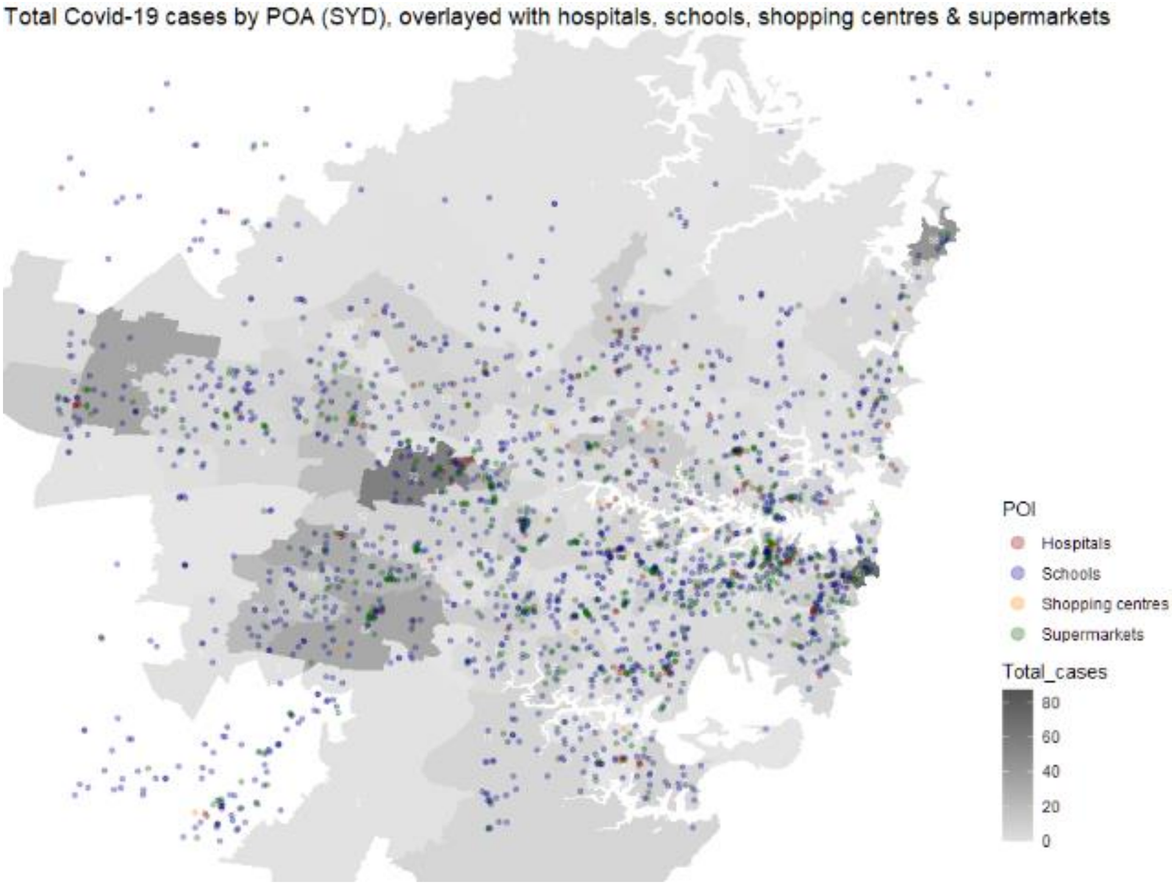


Figure 8: Total Covid-19 cases by POA (SYD) overlayed with amenities/POIs

We used the sp package in R to convert the geocoordinates to the count of POIs in each Postal Area. This is shown in Figure 9 below.

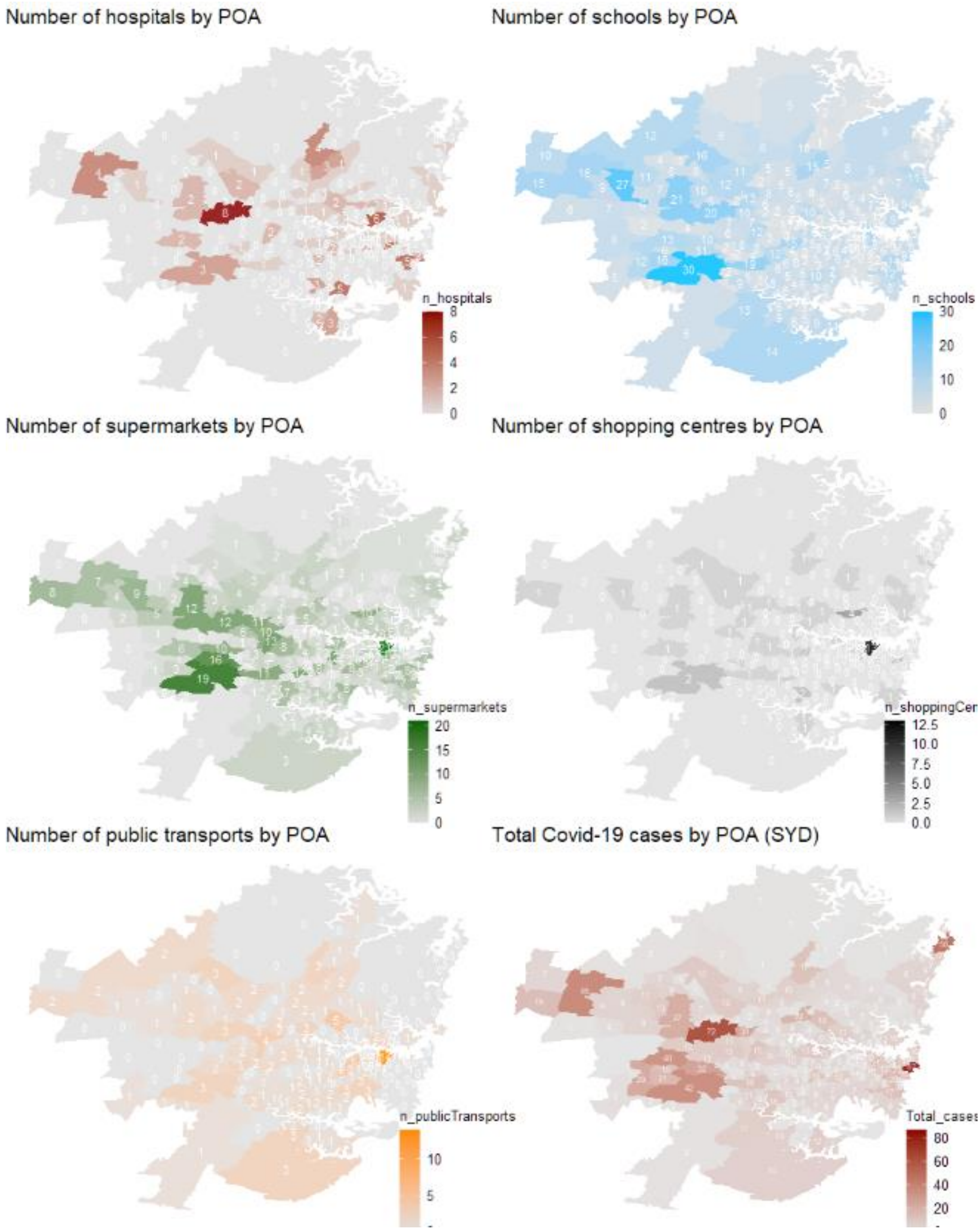


Figure 9: Converted count of amenities/POIs in each POA area in Sydney

As an illustration, we can calculate the gravity score from each Postal Area to the likely origin of the Western Sydney Cluster (2145) outbreak of the disease, based on the "naïve" gravity equation. The calculated result is visualised in Figure 10 below.

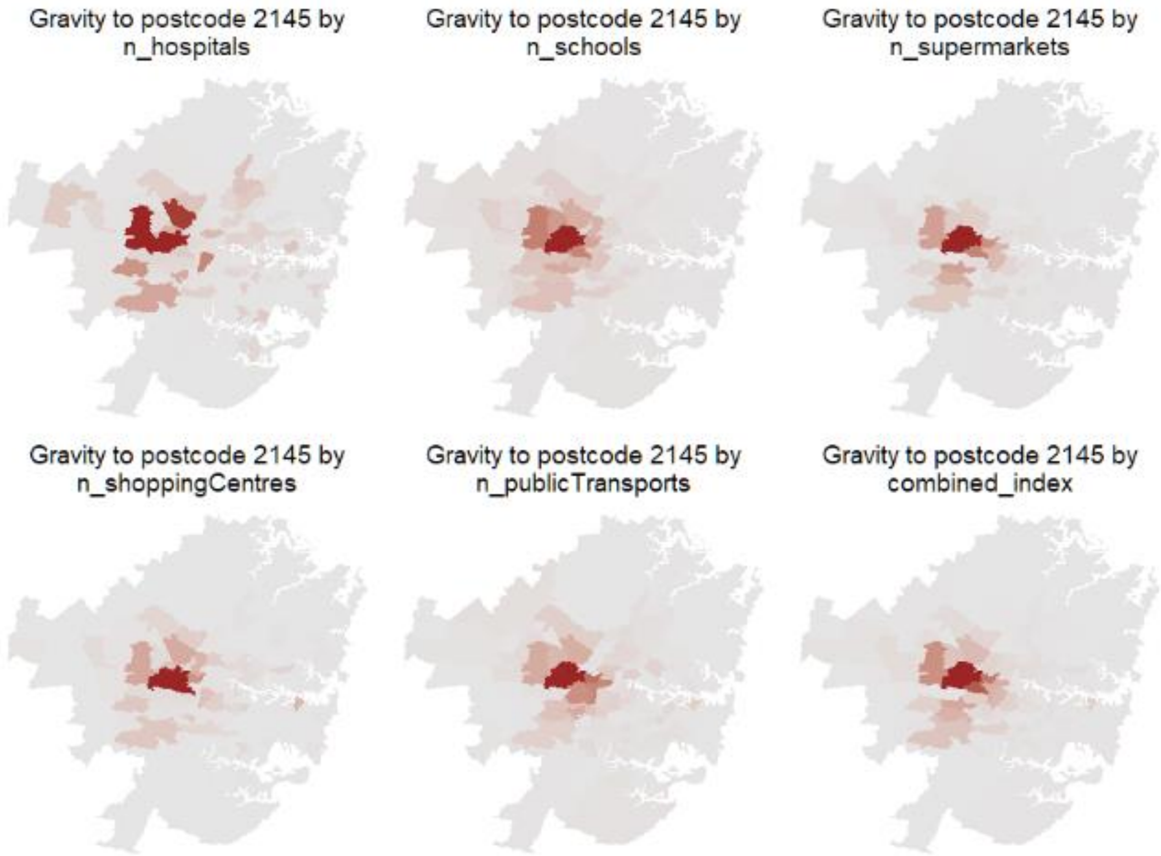


Figure 10: Illustration of the calculated gravity score from postcode 2145

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:

$$Gravity = \frac{\sum(Normalised\_number\_of\_POIs_x * Normalised\_number\_of\_POIs_y)}{Distance_{x-y}^2\ (or\ Travel\ time_{x-y}^2)}$$

For each POA area, we can sum up the gravity scores associated with each other POA area in SYD. The result is an aggregated gravity measure, which can be used as an additional feature in any clustering methods combined with the census/homophily features. The result is shown in Figure 11 below. It is not surprising that the CBD area of Sydney is associated with the highest aggregated gravity score.

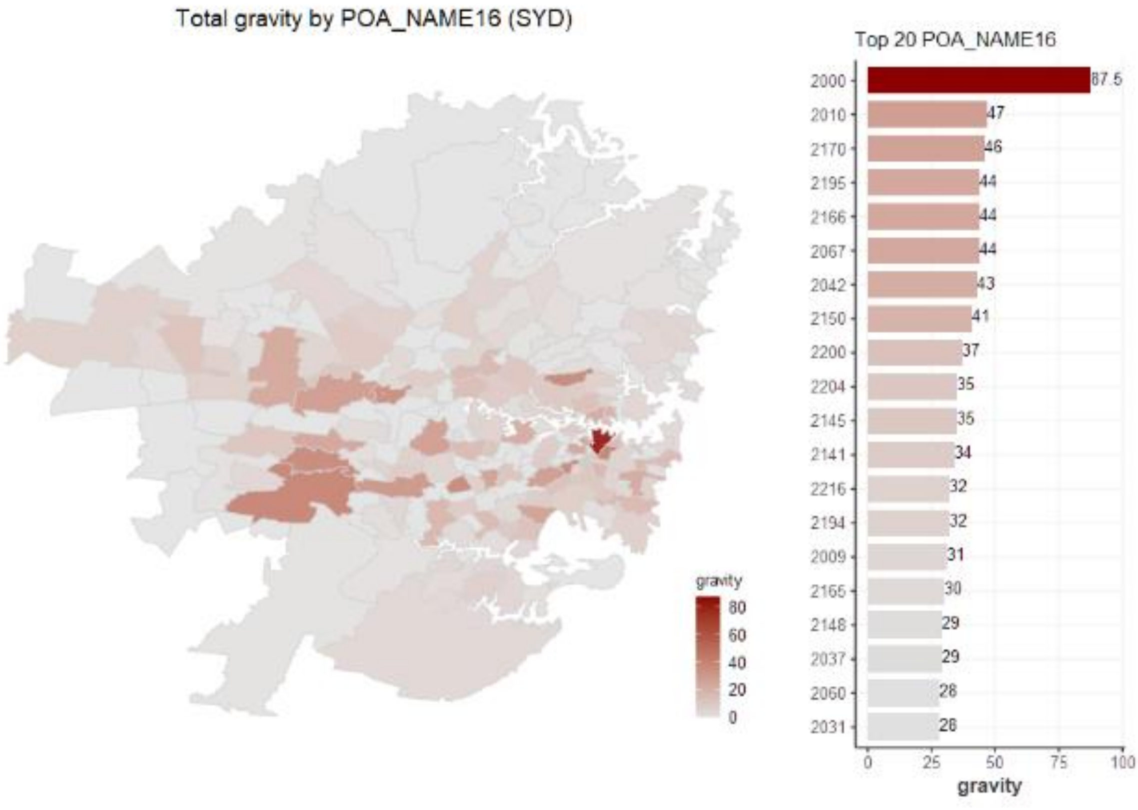


Figure 11: Total aggregated gravity score by POA area in Sydney

2.2 The Homophily Principle

The fundamental principle behind homophily is that "similarity breeds connection." In the context of curbing the outbreak of an epidemic disease, we can explore the extent to which the level of "connection" estimated from "similarity" between two areas can be informative regarding the spatial diffusion of a local outbreak.

In the context of this study, homophily refers to clustering based on a wide range of suburb features, including income, language, property type, rental price, repayment level, and property composition in the suburbs. Suburbs within the same cluster can be seen as those which share similar social profiles and are likely to draw similar people together.

As an illustration, Figure 12 shows the choropleth map for Sydney based on a subset of the homophily feature we use in this report to perform the analysis.

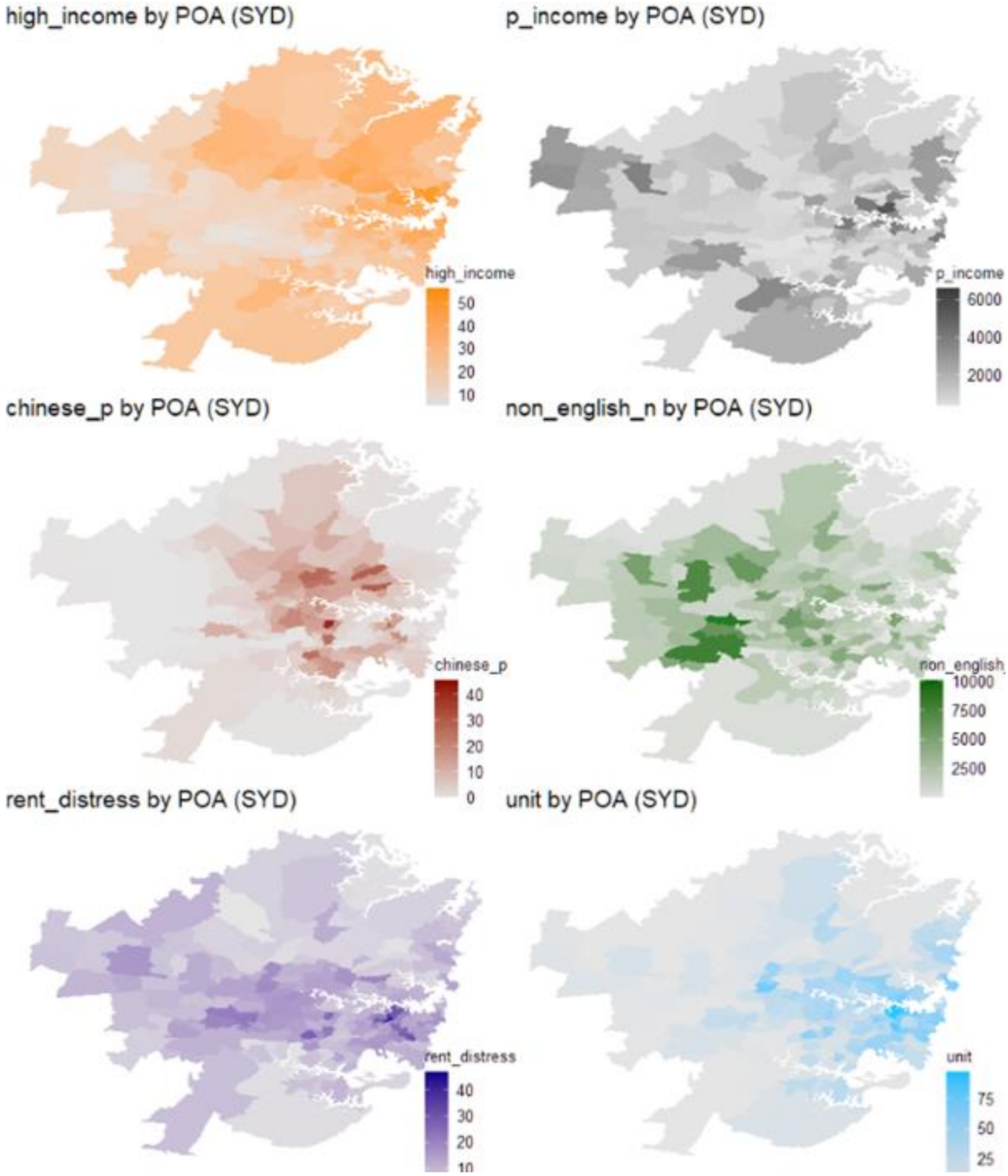


Figure 12: Illustration of homophily features in SYD area from the 2016 Australia census data

3 Data

3.1 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.

3.2 Geographic location data SYD/AU

Table 1 below summarise the source of addresses and geographic coordinates of various "points of interest" in Sydney.

Table 1: Data source for geocoding/address of amenities/POIs in Sydney

Point of interest (POI)	Source of address	Geographic coordinates
Hospitals	Australian Institute of Health and Welfare	Australian Institute of Health and Welfare
Schools (Primary & Secondary)	Australian Schools List	Australian Schools List
Public transport (train, metro, light-rail, ferry stations)	Transport NSW official website	Google Geocode API
Shopping centres	Wikipedia – List of shopping centres in Australia	Google Geocode API
Supermarket/Groceries	Australian yellow pages	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` package in R).

3.3 Suburb demographic data SYD/AU

In Australia, ABS conducts a census every five years, which provides one of the most accurate demographics data sources locally. As part of the census collection and summary, a significant amount of suburb-level data is made available. In general, there are two main ways to access suburb-level data from ABS. The first option is to access suburb-level data from TableBuilder, an ABS in-house product designed for scalable analysis. The second option is to summarise results from QuickStats, a frontend page that presents readily available suburb data across multiple dimensions, including demographic education, cultural language diversity, employment, family composition, employment status of couple families, dwelling information, and people characteristics.

QuickStats option was chosen as the preferred option for suburb-level data for two main reasons. One is easy reconciliation with the QuickStats page, and the second reason is simplicity due to all data being collated at suburb level already.

The main downside of obtaining suburb stats from QuickStats is around data access. As QuickStats has been designed for the public to view suburbs of interest instead of scalable analysis on the suburb level, API access is not available. This challenge led to the team using the web scrapping approach by dynamically traverse all suburb links behind every SSC code. Python has been deployed in this application with a heavy application of web scrapping and regular expression packages, including beautifulsoup, JSON, re, etc.

3.3.1 Missing data imputation

Since suburb data were collected from different sources, when the final feature set was prepared at the postal area (POA) level, suburbs were missing some or all required features. A weighted average imputation was performed. A weighted average of its neighbours with non-missing values was taken with weight determined by mesh block size for suburbs with missing attributes. This imputation results in 30% of suburb data to be retained compared to if no imputation is performed.

3.4 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.

4 Model Evaluation

A backtest model evaluation approach is proposed to compare model performance across multiple models. A fundamental hypothesis is that once lockdown has been imposed on cluster(s) related to POAs with cases over the last n days, then future cases over the next n days occurring in the same cluster(s) are assumed to be preventable. This assumption is optimistic but not going to bias our result since all lockdown strategies in our comparison share this assumption. The rationale is to compare the percentage of preventable cases overall cases by implementing a cluster-based lockdown. Suppose there is a case in suburb A on day 1, and there is a case in suburb B, which is in the same cluster as suburb A on day 4. The case in suburb B is considered preventable if suburb A and all suburbs in the same cluster enter a 3-days and above lockdown. This is illustrated in Table 2 below.

Table 2: Illustration of preventable cases in our assumption for the hypothetical 3-day lockdown strategies

Notification Date	Case Number	Suburb/Postcode	Geographic cluster	Preventable?
2020-03-01	1	A	Cluster - alpha	No
2020-03-02	2	B	Cluster - alpha	Yes
2020-03-03	3	C	Cluster - beta	No
2020-03-10	4	D	Cluster - alpha	No

Different lengths of lockdown also need to be considered for model performance evaluation, for both model optimisation perspective and candidate model selection perspective.

One significant difference between this testing approach and the traditional machine learning testing is that the training of models does not rely on any data related to the actual COVID-19 case. The purpose is to assess the predictive power of our model/strategy using only geographic and demographic information. Since the test dataset is independent of all model training, no separate training, cross-validation, or testing dataset is required.

5 Model outcome

5.1 Benchmark Models: Neighbourhood lockdown

All models implemented as part of this study are all by nature clustering-based models. The base scenario is a naïve neighbourhood lockdown strategy, which is our benchmark scenario. Real live implementation of this strategy might have a significant level of human overrides. Still, in this paper, we assume neighbourhood lockdown strategy refers to locking down neighbouring/adjacent suburbs/LGA/LHD/City when new COVID-19 cases emerge. This is illustrated in Figure 13 below.

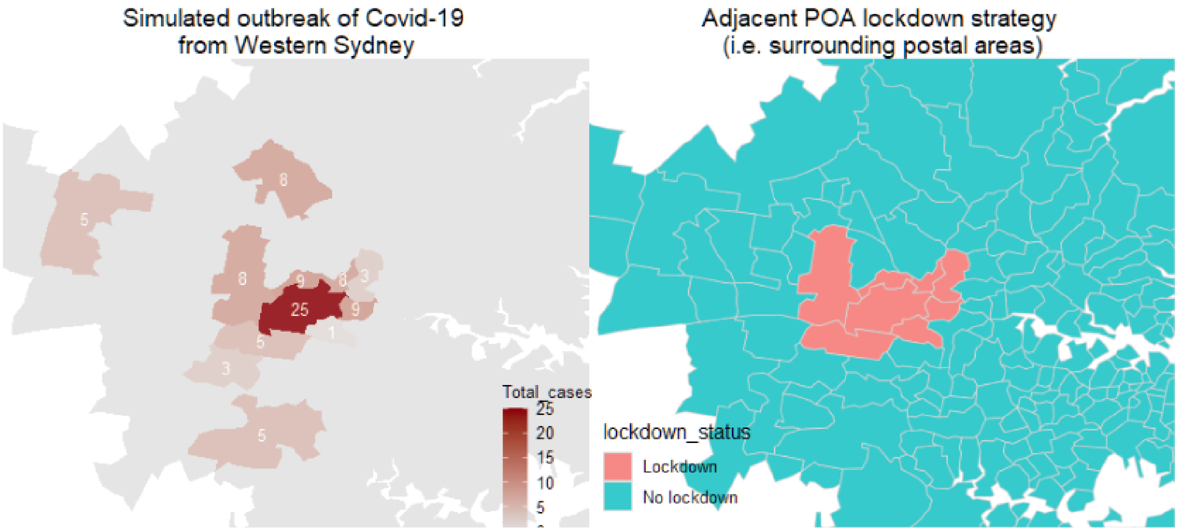


Figure 13: Illustration of the naive neighbourhood lockdown stragegy

The four variants of the neighbourhood lockdown strategy have different scopes. Neighbouring suburb lockdown has the most negligible impact, with an average of 5 neighbouring suburbs going into lockdown when a suburb has identified a new case. LGA lockdown would result in all suburbs falling under the same local government area going into lockdown, impacting an average number of 32 suburbs going into lockdown. LHD is one level higher than LGA where Sydney is comprised of just 6 LHDs. The extreme form of neighbourhood lockdown is to lock down the whole city. To illustrate, Figure 14 and Figure 15 demonstrate the size of LGAs and LHDs in the Sydney geography.

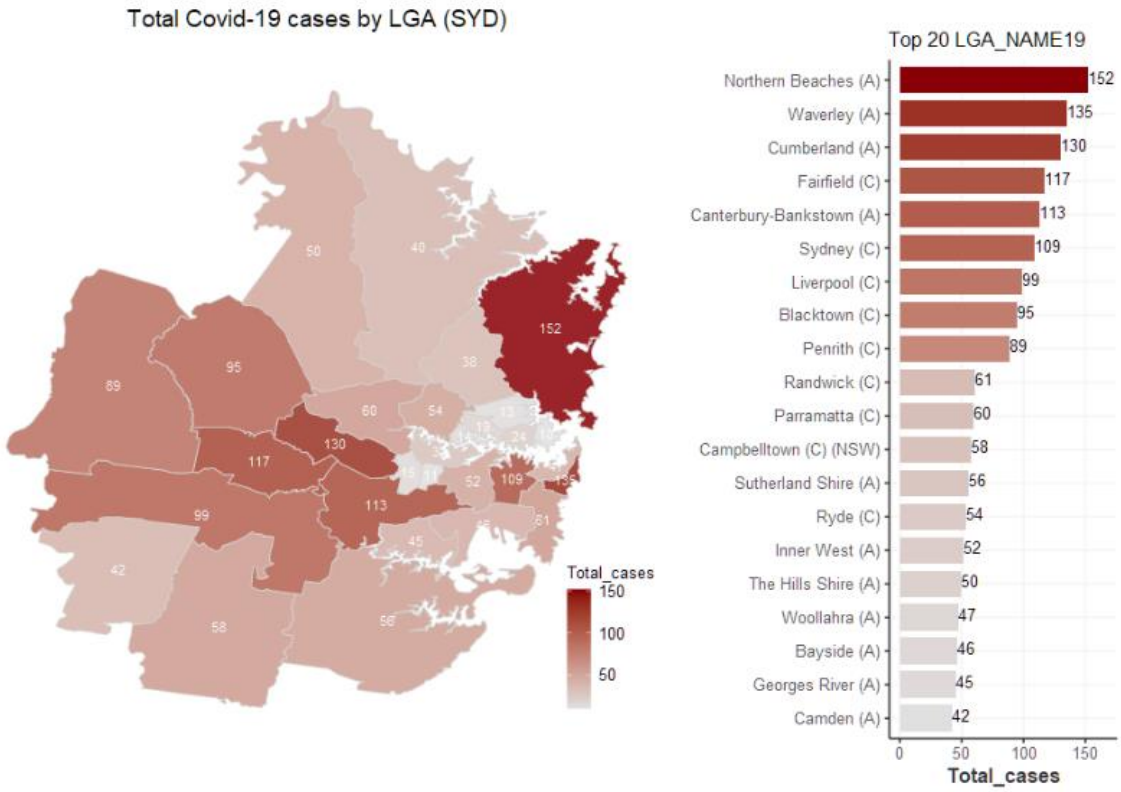


Figure 14: Size of Local Government Area (LGA) in Sydney coloured by total Covid-19 cases

Figure 15: Size of Local Health District (LDH) in Sydney

Lockdown period has been chosen as 14 days- because it is one of the most commonly used lockdown periods during COVID-19 by governments around the world.

Figure 16 shows the effectiveness of the neighbourhood lockdown approach as we increases the scope of the "neighbourhood" - with neighbouring suburb lockdown achieving an effectiveness of

70.6%, LGA and LHD lockdown having 85.5% and 94.3% effectiveness, respectively, and city-level lockdown having a staggering 99.8% of preventative effectiveness.

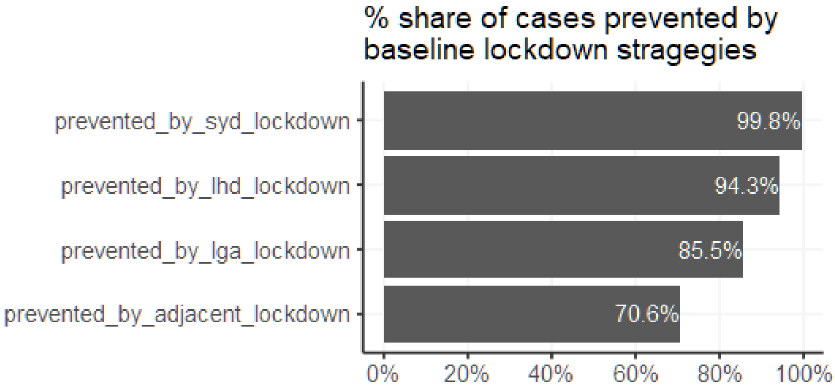


Figure 16: Lockdown effectiveness from the baseline neighbourhood lockdown strategeis

A blind effectiveness-following strategy cannot be followed as locking down the city may lead to unnecessary wastes and inefficiencies. A more balanced view is to look at both the number of suburbs impacted and the effectiveness.

5.2 Challenger

The three challenger models we developed adopt different approaches to form clusters. The gravity model is an improvement to the benchmark scenario by considering POIs and travel time for a better “connection” representation between geographical areas. Clusters are formed recursively for all suburbs by considering the top 5 suburbs with the highest gravity index to each suburb of interest. The number “5” is chosen because it would make the gravity model lock down the same median number of suburbs as neighbouring suburb lockdown model in the benchmark, allowing for a fair comparison of model performance. A standard gravity model leads to a lockdown effectiveness of 72.7%, which is marginally higher than the benchmark performance of 70.6% from neighbouring suburb lockdown.

The homophily principle considers the human nature of staying in communities with people alike. Implementation of the homophily principle involves data collection of demographics data like income, language, property type, rental price, repayment level and property composition in the suburbs. Principal component analysis has been introduced for dimensionality reduction given most features are correlated to each other, principles would be retained once cumulative variability has reached 90%. Following PCA, K-means and hierarchical clustering are applied to feature set of principal components to generate clusters. The number of clusters is chosen so that the median number of suburbs in each cluster would be 6, equivalent to locking down 5 other suburbs, which is in line with the benchmark having a median of 5 adjacent suburbs for each suburb.

Both K-means clustering and hierarchical clustering achieve significantly better effectiveness than the benchmark neighbouring suburb lockdown strategy, with effectiveness sitting at 75.9% and 77.1%, respectively. This out-performance is likely to come from the fact that people travel between culturally similar suburbs that may not be bound by geographical distance.

The hybrid approach can take two forms: 1) To introduce additional gravity index features into the original demographics feature set before K-means or hierarchical clustering, 2) To combine the results from the gravity model and homophily model to enforce lockdown. The expectation is that with the enrichment of geographical views, the hybrid approach will likely have improved superior 'preventative' power.

The result proves the prior hypothesis, with both hybrid approaches performing among the best two challenger lockdown methods. The integration approach of incorporating gravity features in clustering still guarantees a median of 6 suburbs per cluster. One practical implication of considering both demographics and geographic features is to acknowledge that people travel based on proximity and similarity, which closely resembles the real world.

The additive version of the hybrid method that locks down suburbs with high gravity and suburbs similar to the suburb of new COVID cases has the best performance among all challenger models, with an effectiveness of 92.9%. However, this method is not on the same playground as other methods because it is likely to lock down twice as many suburbs compared to baseline neighbouring suburb approach, homophily clustering approach and the gravity approach. So out-performance to a certain extent comes from this aspect. Nevertheless, this hybrid approach still outperforms LGA lockdown even if it impacts fewer suburbs, and it is at a similar performance level as LHD lockdown, which would lock down one-sixth of the entire Sydney. The hybrid approach still has impressive effectiveness after comparing with higher-level benchmark lockdown strategies.

The effectiveness from all lockdown strageteis, including both the baseline benchmark and challengers are shown in Figure 17.

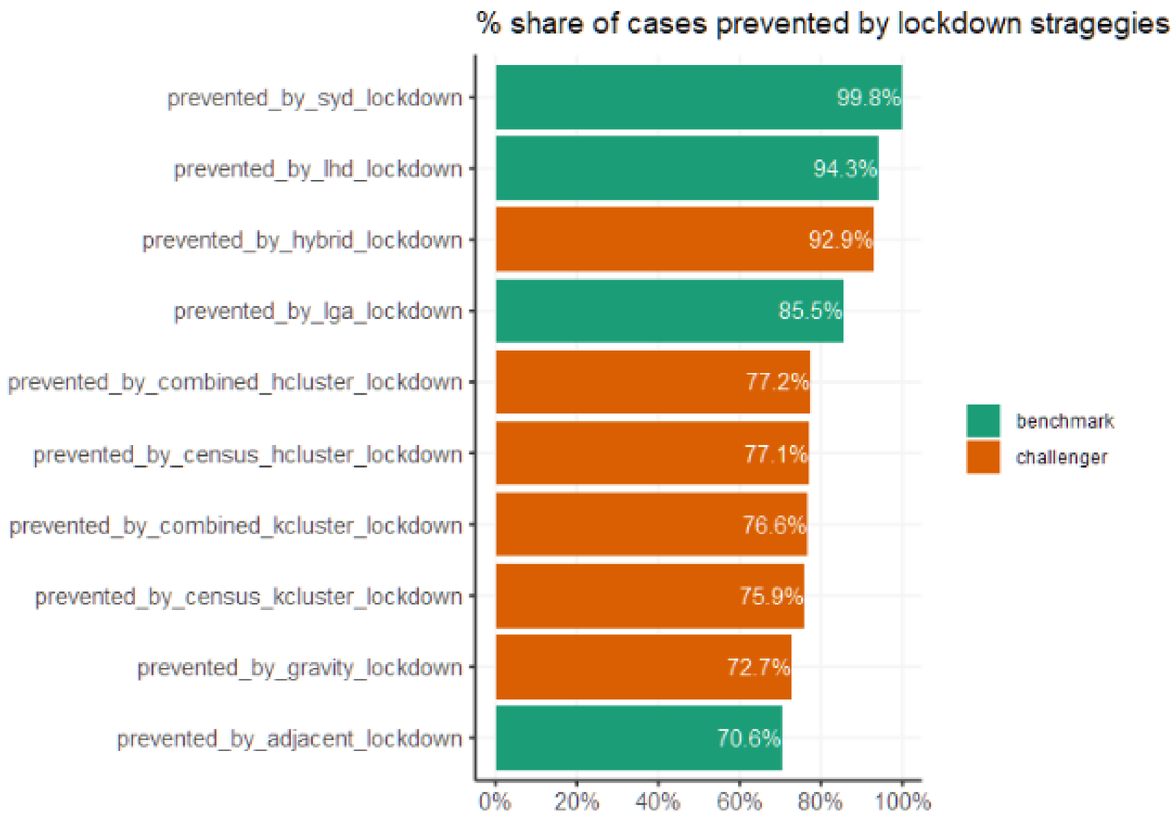


Figure 17: Lockdown effectiveness of the benchmark and challenger lockdown stragegies

Furthermore, another aspect of lockdown is the duration of the lockdown period. While model selection has a sizable impact on lockdown effectiveness. Duration also plays a significant part in the process. As shown in Figure 18, it holds for all models that the effectiveness of lockdown increases with lockdown duration. However, it is worth pointing out that there is a sizeable flattening happening towards the end of the lockdown curve, indicating that locking down for an extended period achieves little while costing too much. Depending on the government's tolerance level, different durations can be chosen. The common 14-day lockdown selection seems like a balanced option for effectiveness and costs.

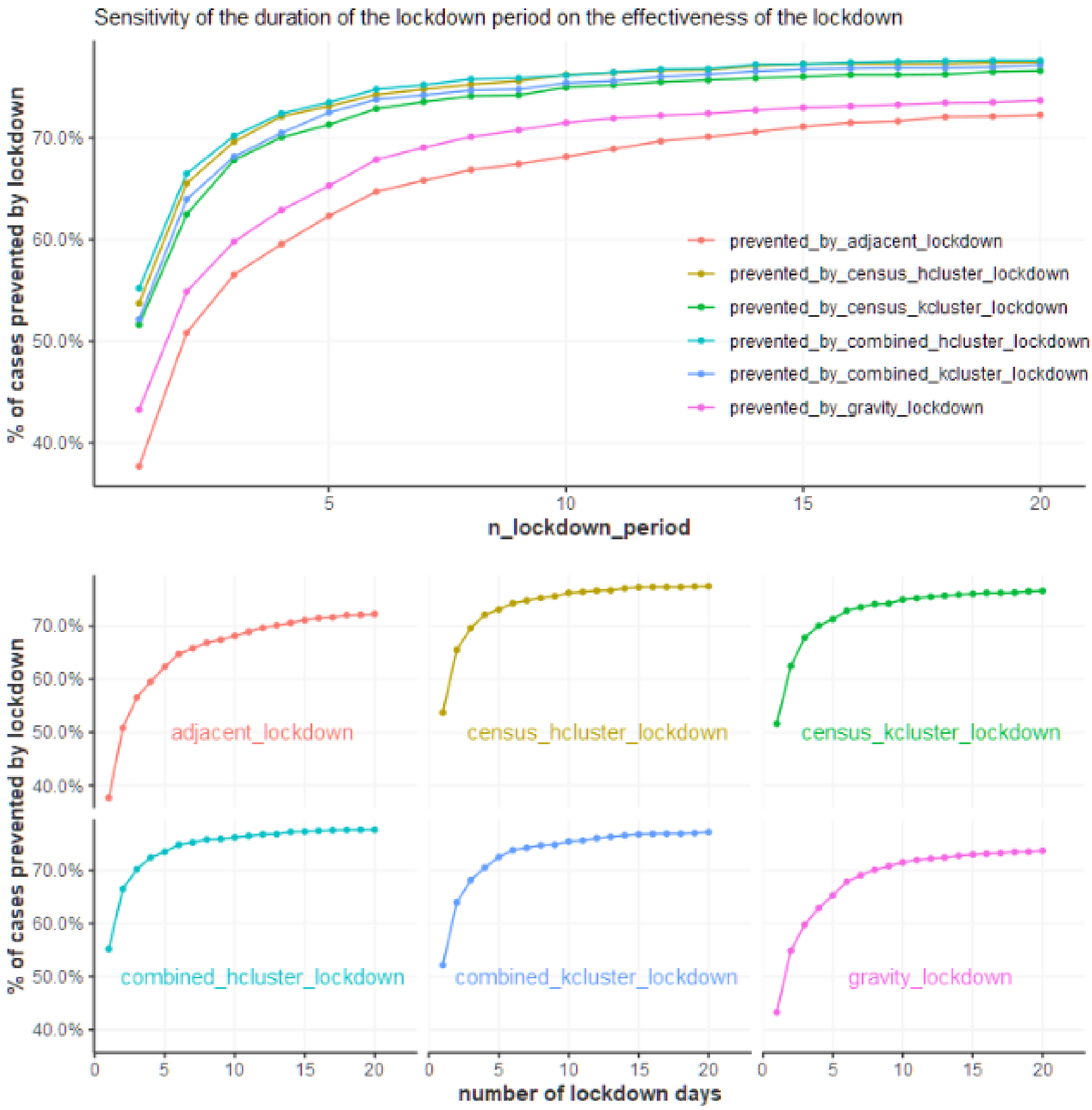


Figure 18: Sensitivity of the duration of the lockdown period on the effectiveness

6 Conclusion

This paper examined three innovative approaches of lockdown: gravity index-based lockdown, homophily clustering-based lockdown, and hybrid approach of lockdown. The idea behind those proposal is that 'similarity' and 'connectivity' of suburbs might play a significant part of virus local transmission, as opposed to a naive approach of purely considering adjacent suburbs or areas. The results turned out to back this idea with all models outperforming their respectively benchmark models with similar number of suburbs subject to lockdown. The other aspect of lockdown is lockdown duration. Based on analysis performed, every lockdown strategy including the benchmark ones would have better effectiveness if lockdown duration were higher. However, the other aspect of lockdown is the economic costs of doing so - with the lockdown effectiveness curve almost flat after 14 days, the most common lockdown period of 14 days is not without any virtue. Insights around modelling and lockdown period selection in this paper can hopefully inform policy makers to achieve a higher effectiveness/cost ratio in their virus management process.

7 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 Figure 19.

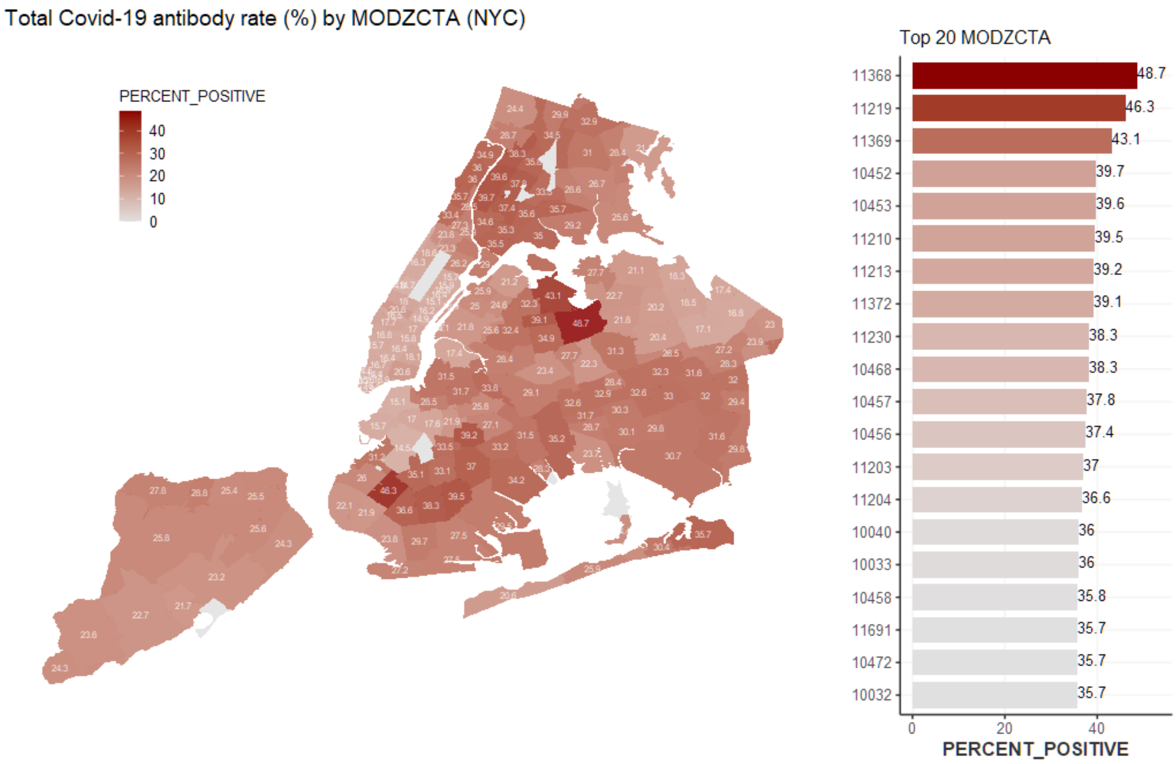


Figure 19: Total Covid-19 antibody rate (%) by MODZCTA (NYC)

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.

8 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

Lauw, H., Shafer, J. C., Agrawal, R., & Ntoulas, A. (2010). Homophily in the digital world: A LiveJournal case study. *IEEE Internet Computing*, 14(2), 15-23. sourced from: https://ieeexplore.ieee.org/abstract/document/5396305

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

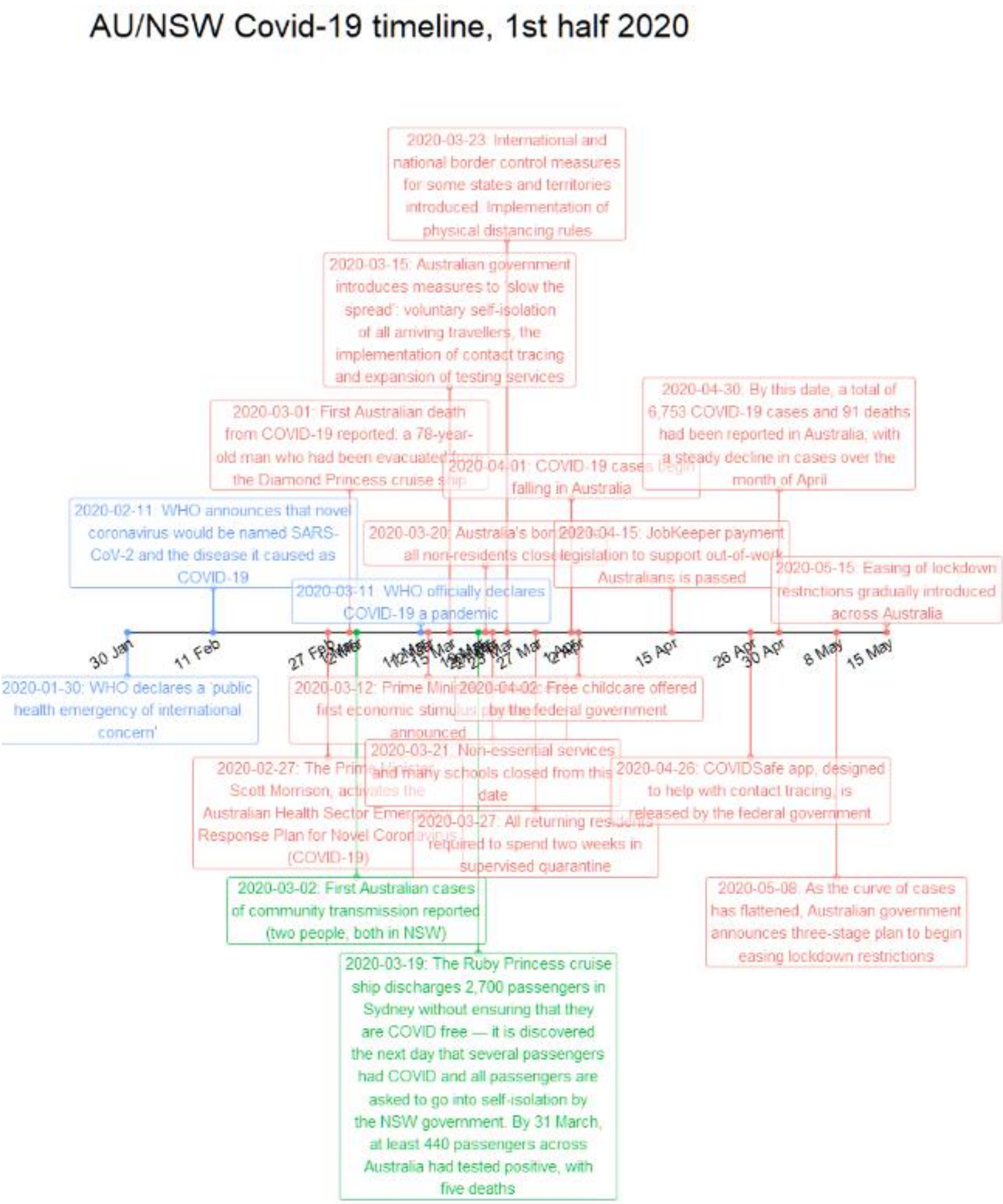
Vadim A. Karatayev, Madhur Anand, and Chris T. Bauch (2020), Local lockdowns outperform global lockdown on the far side of the COVID-19 epidemic curve, Proceedings of the National Academy of Science of the United States of America (PNAS), September 2019, sourced from: https://www.pnas.org/content/117/39/24575

9 Appendix

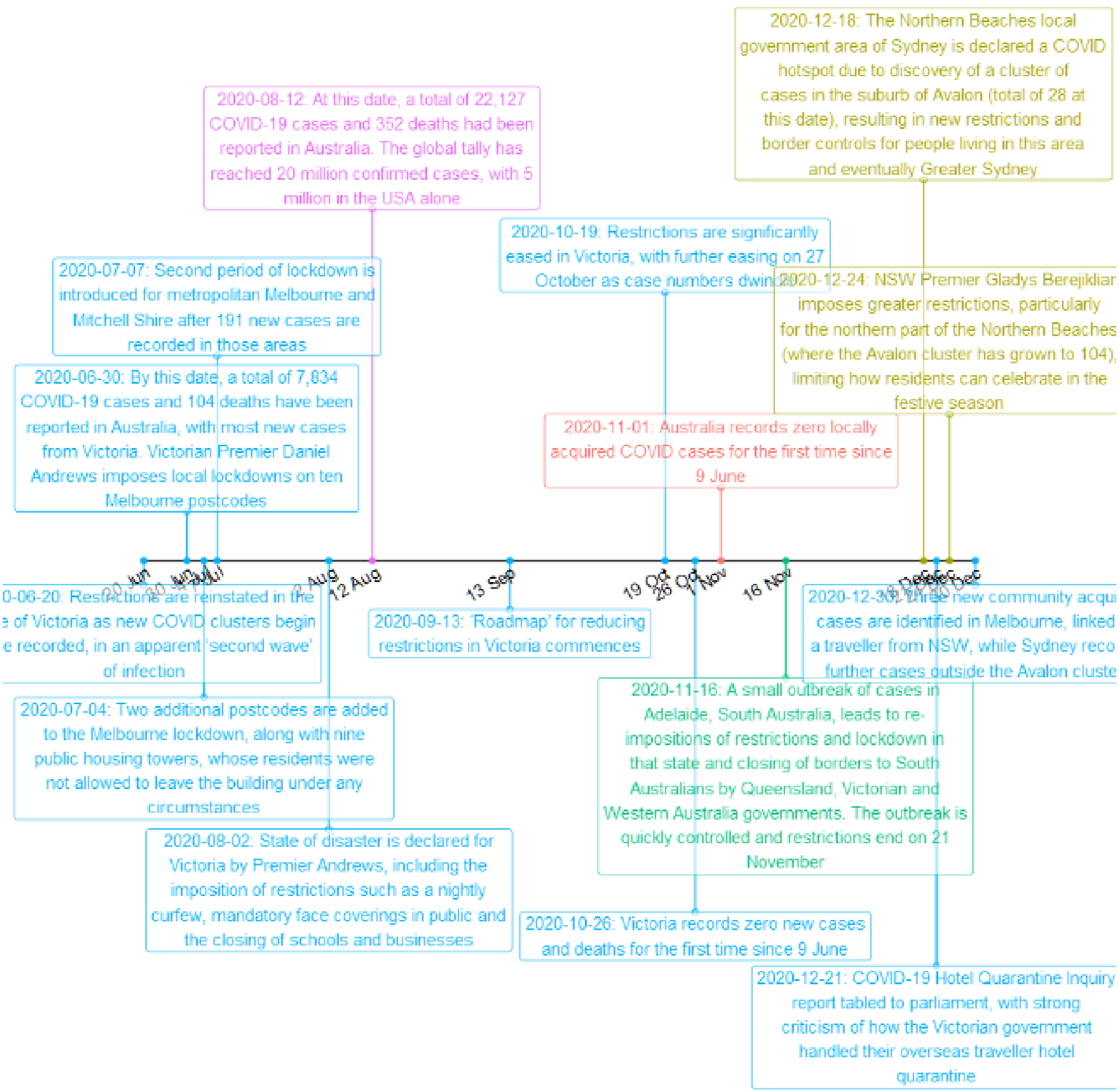
Covid-19 related news timeline in Australia, 2020

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

Visualisation:



AU/NSW Covid-19 timeline, 2nd half 2020



10 Breakdown of contributions

Stages	Steps	Contributor
Data Collection	shapefiles	Yanjun Liu
	Suburb quickstats	Zhikang Yu
	Covid-19 NSW/AU	Yanjun Liu
	Geographics data	Yanjun Liu
Methodology	The Gravity Model	Yanjun Liu
	The Homophily Principle	Zhikang Yu
	The hybrid approaches	Joint
Evaluation	Backtest strategy	Zhikang Yu
	Evalution results	Joint
Writing	Essay	Joint