1. **Introduction**

Since the outbreak of Covid-19, governments around the world 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 extremely unlikely for all these three conditions to be met simultaneously. In most cases, local transmission cannot be contained without extreme measures such as extended periods of lockdown. 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 immensely helpful for the government to strike a balance between containing the virus and maintaining the economic output.

In this project, we will 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 stronger 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. **Background**

The first positive case for Covid-19 in Australia was reported on 25 January 2020 from incoming travellers from China. One month later, on 27th Feb, the 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”.

{Chart index} shows the daily count of new cases (locally acquired) in the Sydney area. The three distinct waves can be identified in different time periods. The magnitude of outbreak is relatively small compared to what happened in the rest of the world. This creates an opportunity to investigate the spatial diffusion of the virus in a relatively “controlled” environment, where the source of the virus is likely originated from the few suburbs that first recorded the spike of new cases.

Diagram

Description automatically generated

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

Line chart

Description automatically generated

Diagram

Description automatically generated

{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 locations of the three clusters can be identified here. The Western Sydney cluster seems to have spread more widely than the other two beach clusters. It is conceivable that the flow of people from the beaches can be more scattered thus hard to predict. Consequently, there is a higher chance for a clustering algorithm to “predict” the spread of the virus in the Western Sydney area.

Chart

Description automatically generated

{Chart index} shows the accumulated cases over time. It appears that the Wester Sydney cluster could be originated from the further west area of Cambridge Park, before it spread out to the middle of the Western Sydney area which is close to a local transportation hub.

Diagram

Description automatically generated with medium confidence

* 1. **Motivation of the gravity model**

{chart index} overlays the public transportation network on top of the choropleth (heatmap). It is visible and conceivable that the virus might have been transmitted along the train lines. Unsurprising, the train lines are also associated with areas with dense population.

Map

Description automatically generated

Similar pattern can be witnessed if we look at the number of supermarkets/groceries in each postal area. These observations inspire the use of the Gravity Models to measure the level of “connectivity” between areas using a series of “point of interest”, such as the number of schools, supermarkets, public transports, etc., as the “mass” and the physical distance or travel time as the distance measure to predict the spread of the virus (see details in section 2.1).

**A picture containing diagram

Description automatically generated**

* 1. **Motivation of the homophily model**

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

* 1. **User story**

*Example of lockdown around the world, how the lockdown areas were chosen, are they effective and why our clustering method can help.*

* 1. **Literature review**

*One paragraph on gravity model.*

*One paragraph on the homophily principle.*

1. **Methodology**
   1. **The Gravity Model**

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

* 1. **The Homophily Principle**

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

* 1. **Geospatial Clustering**

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

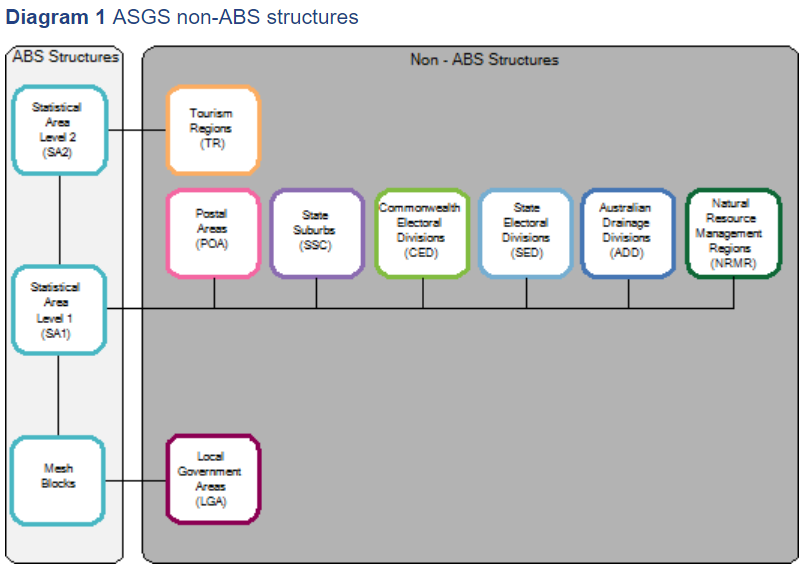
1. **Data**
   1. **Covid-19 confirmed cases SYD/AU**

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

* 1. **Geographic location data SYD/AU**
  2. **Suburb demographic data SYD/AU**
  3. **SSC to POA mapping**

ABS data explanation

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



Data source (mapping to meshblock) - <https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.003July%202016?OpenDocument>

1. **Model outcome**

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

1. **Model Evaluation**

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

1. **Conclusion**