Using Demographics and Location-Based Social Network Data to Understand the Spread of COVID-19 in London

Thomas Hall

Monday, 04 May 2020

1. Introduction

Since it was first discovered in December 2019, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has had a massive impact as it has spread across the globe. The virus causes the disease known as COVID-19 and due to its novel appearance, there is great uncertainty surrounding its epidemiology, such as risk factors, transmission characteristics, and possible control measures. With such uncertainty and rapid onset, governments the world over have had no other option but to shut down their entire economies, confine their citizens under lockdown, and bide their time until an exit strategy can be found.

As of May 4th, 2020, there have been 24,988 total cases of COVID-19 in London, representing a significant outbreak of the disease [1]. This report seeks to use demographic profile data, venue location data and coronavirus case data to examine outbreak trends across London's 32 boroughs. Example questions to be answered are:

- Are deprived areas at more risk of developing COVID-19?
- Are areas with an older population or those with higher BAME populations at greater risk of having more COVID-19 cases?
- Is prevalence increased by the number of certain venue types, e.g. pubs, restaurants, or parks?

Findings from the investigation could be useful in informing how the Greater London Authority defines and manages the exit from the current lockdown.

2. Data

This section covers the collection, understanding and preparation (wrangling) of the data required for the project.

2.1. Data Sources

Data for this investigation is acquired from the following sources:

- **Demographics**: data is sourced from the Greater London Authority [2], which has comprehensive profiles of each borough.
    - Example data for each borough includes multiple statistics to describe; population density and composition, age, employment rates, household incomes, home ownership, carbon emissions, transport usage, life expectancy, and life satisfaction amongst others.
    - The data is available in both CSV and XLSX format.
- **Venue Location Data** is sourced from the Foursquare API [3] versioned to March 15th – just before the main onset of the coronavirus where ordinary trends are assumed to be unaffected.
    - The top 200 locations are acquired for each borough based on its latitude and longitude.

    - The data is read in from the Foursquare API in JSON format.
- **Coronavirus** data is sourced from Public Health England [1].
    - Daily updates are published for each upper tier local authority in England – in London this corresponds to the boroughs.
    - The data is available in CSV and JSON format.

2.2. Data Wrangling

The demographics dataset initially contains 80 statistics to characterise each area. A subset of 26 features is instead chosen to reduce the size of the profile, whilst still summarising the ethnic, economic, geographic, and medical make-up of the community (Table 1). The data is already pre-cleaned with the only missing value for these features being the gross annual pay in the Gross Annual Pay in the Royal Borough of Kensington and Chelsea. This missing value is handled using the ratio of median household income in the borough to the citywide average to calculate gross annual pay from the citywide average. The only non-numeric feature out of those selected is "Inner/Outer City", which is handled by using one-hot encoding. All variables are then standardised by the mean and standard deviation of the data.

Table 1: Demographic Features

Feature	Reason for Inclusion
Population Density Inner/Outer City	High density could indicate high probability of interaction.
Average Age Proportion of population aged 0-15 Proportion of population of working-age Proportion of population aged 65 and over	Age has been reported to increase risk of severe COVID-19.
% of resident population born abroad % of population from BAME groups	Being of BAME has been reported to increase risk of serious COVID-19.
Employment rate (%) Unemployment rate (%) Youth unemployment rate (18-24) Proportion of the working-age population who claim out-of-work benefits (%) (May-2016)	Employment rates may affect COVID-19 spread.
% working-age with a disability	Disability may increase risk of COVID-19.
Gross Annual Pay Modelled Household median income estimates	Economic comfort may affect exposure to SARS-CoV-2.
% of area that is Greenspace % of adults who cycle at least once per month	Proxy for activity levels, which may affect risk of serious COVID-19.
Male life expectancy Female life expectancy Childhood obesity prevalence People afed 17+ with diabetes	General health is likely to affect risk of serious COVID-19.
Life satisfaction score Worthwhileness score Happiness score Anxiety score	Mental health may affect physical health and risk of serious COVID-19.

The coronavirus dataset is published daily. It consists of lab-confirmed cases at the levels of upper tier local authority, region, and nation. Deaths data is also available but only at the levels of region and nation, hence it is too coarse for the purpose required. This is to be the target dataset and so only the cumulative cases for each borough up until May 2nd, 2020 are acquired from the dataset, the rest are disregarded. The names of

the boroughs are formatted consistently with the demographics dataset, making joining easy. So that prevalence can be compared between boroughs, the metric of cumulative cases is converted into cumulative cases per 1000 population using the population data from the demographics dataset.

Data for the top 200 venues in each borough is acquired through the Foursquare Places API. Searching for the venues requires coordinates for each borough, acquired through an open source geolocator. The returned venue data consists of the venue name, type, latitude, and longitude. The venues are grouped by venue type in each borough to create a venue profile, the venue types are then standardised across the boroughs.

3. Exploratory Data Analysis

First, the locations of the thirty-two boroughs within London were visualised (Figure 1). London consists of a mix of inner and outer city boroughs, which increase in geographical area towards the outside of the city.

Figure 1: Locations of London boroughs.

The case count per borough was also plotted to determine the variation in caseload for the different boroughs (Figure 2). Croydon (1420), Brent (1387), Barnet (1224), Southwark (1218), and Lambeth (1156) were the five areas with the highest caseloads, whilst the lowest five caseloads were in Kingston upon Thames (472), Kensington and Chelsea (472), Barking and Dagenham (462), Islington (423), and Richmond upon Thames (384).

Also visualised was the cumulative number of cases per 1000 people by borough (Figure 3). Standardised by population, the five boroughs with the most cases are Tower Hamlets (4.2), Camden (3.9), Kensington and Chelsea (3.8), Richmond upon Thames (3.7), and Hammersmith and Fulham (3.5), whilst those with the least are Westminster (2.2), Brent (2.0), Kingston upon Thames (2.0), Croydon (1.9) and Havering (1.8). The mean number of cases per 1000 is 2.82 with a standard deviation of 0.61.

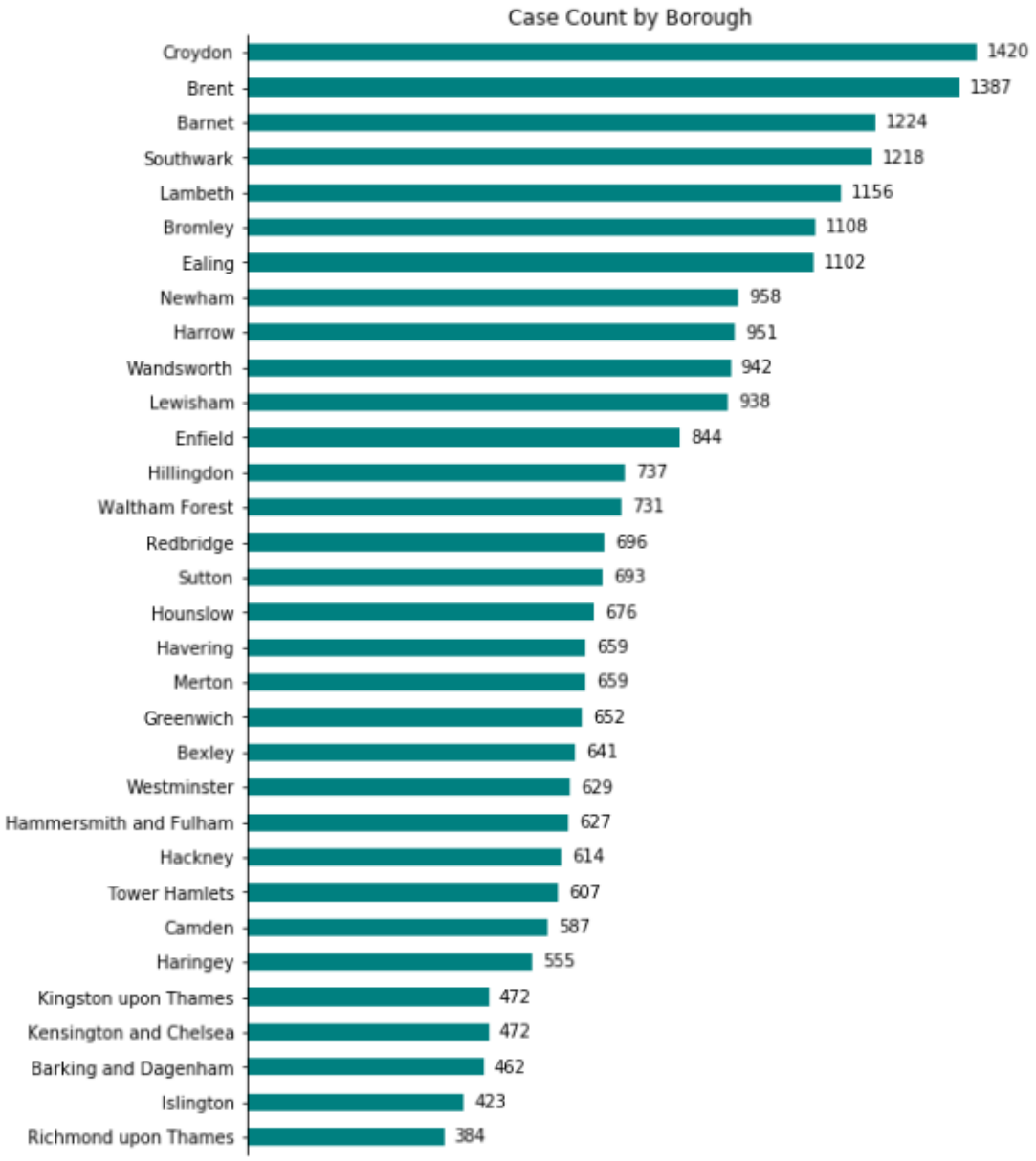

Figure 2: Cumulative Cases by Borough

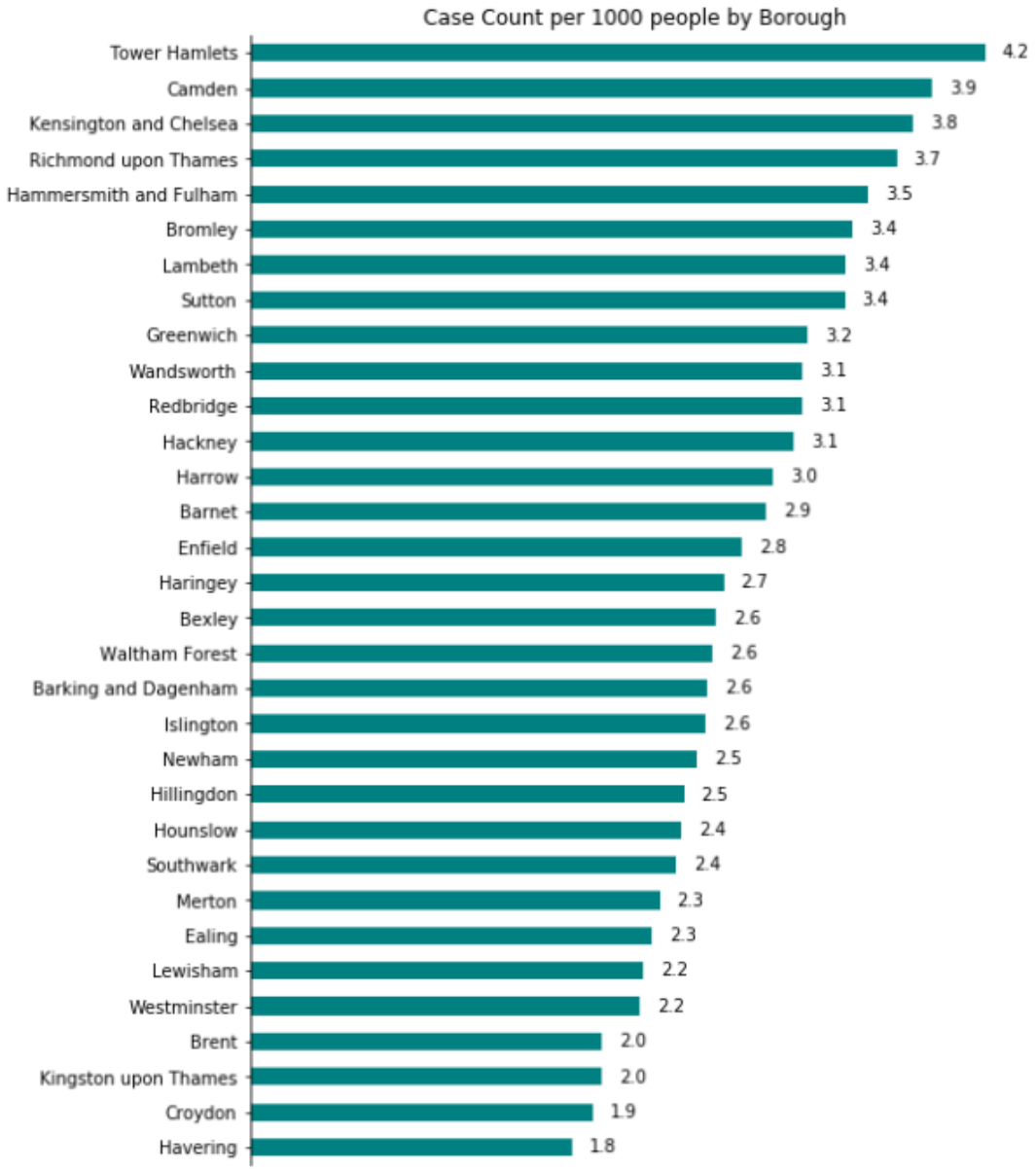

Figure 3: Cumulative Cases per 1000 People by Borough

4. Methodology

The data was analysed by clustering the boroughs to determine whether similar areas experienced similar-scale outbreaks. The clustering analysis was performed twice, once using the demographics data and once using the venue location data.

The data was standardised in both cases by the mean and the standard deviation to prevent dominant features in the clustering. The appropriate number of clusters was determined by plotting the sum squared distance for an increasing number of clusters and using the elbow method. To avoid settling towards a local minimum, the algorithm was run 12 times for each number of clusters, with the best result taken.

The significance of the clustering was then determined by comparing the cumulative lab-confirmed cases per 1000 people across clusters. This evaluation is normalised by population size to eliminate the disparity in population size between boroughs.

5. Results

5.1. Study 1: Clustering by Demographics

The appropriate number of clusters to segment the London boroughs by demographics was determined to be 6 (Figure 4). Geospatially, the clusters tend to be neighbours rather than distributed across the city (Figure 5). Demographic features of the clusters are summarised in Table 2 and Figure 6. Comparing coronavirus prevalence between demographic clusters there does not appear to be a significant difference (Figure 6).

Table 2: Summary of Demographic Clusters

Cluster	Boroughs	Description
0	Hackney, Islington, Lambeth, Lewisham, Southwark, and Tower Hamlets	High-density, inner-city boroughs. Populations are of average age, affluence, and education with average numbers of BAME groups.
1	Bromley, Bexley, Havering, and Sutton	Low-density, outer-city boroughs. Populations are older and less educated with average affluence and low numbers of BAME groups.
2	Camden, and Westminster	High-density, inner-city boroughs. Populations are of average age, more affluent, and more educated with average numbers of BAME groups.
3	Barnet, Croydon, Ealing, Enfield, Harrow, Hillingdon, Hounslow, and Redbridge	Low-density, outer-city boroughs. Populations are older, of average affluence and less education, with low numbers of BAME groups
4	Hammersmith and Fulham, Kensington and Chelsea, Kingston upon Thames, Merton, Richmond upon Thames, and Wandsworth	Mix of high- and low- density inner- and outer- city boroughs. Populations are of average age, more affluence, and more education with low numbers of BAME groups.
5	Barking and Dagenham, Brent, Greenwich, Haringey, Newham, and Waltham Forest	Mostly outer city boroughs of average density. Populations are younger, less affluent, and less educated with high numbers of BAME groups.

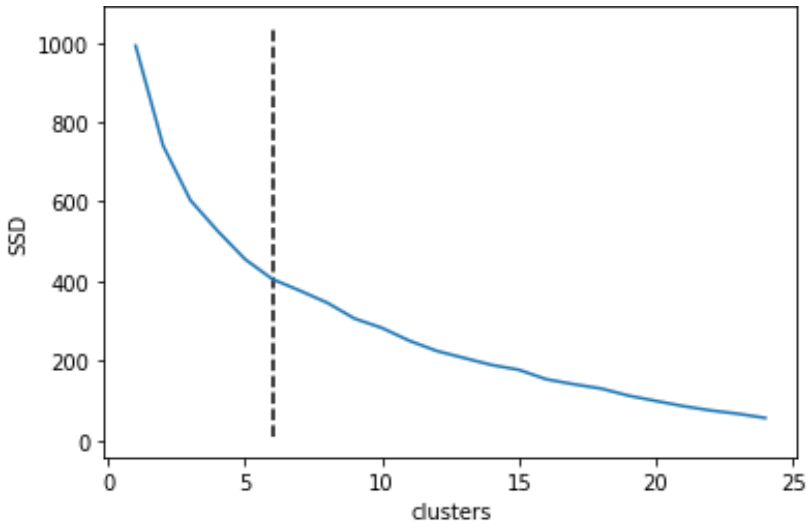

Figure 4: Sum squared distance versus the number of clusters, with the elbow point defined at n=6.

Figure 5: London boroughs clustered by demographics.

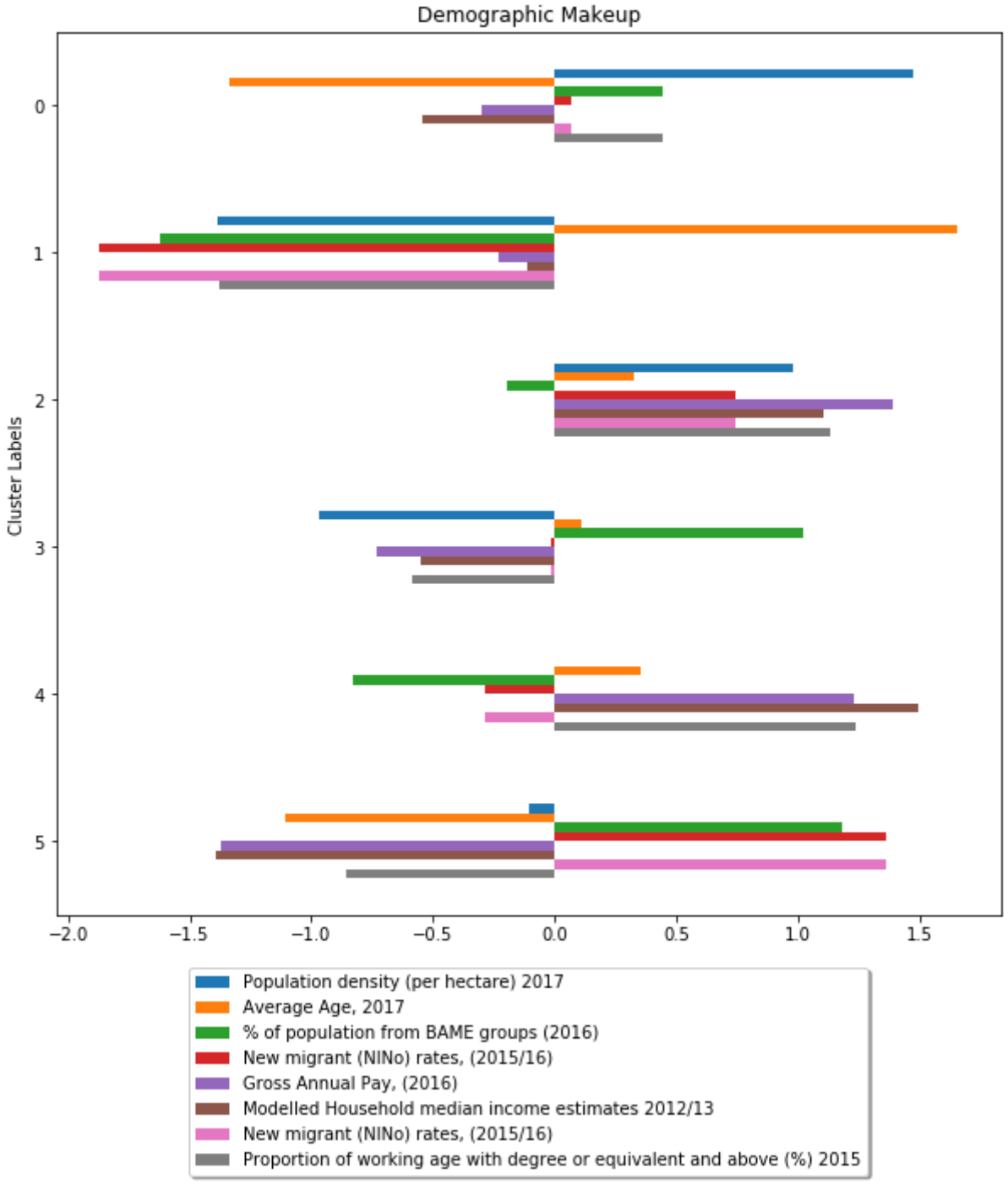

Figure 6: Demographic features of the clustered London boroughs.

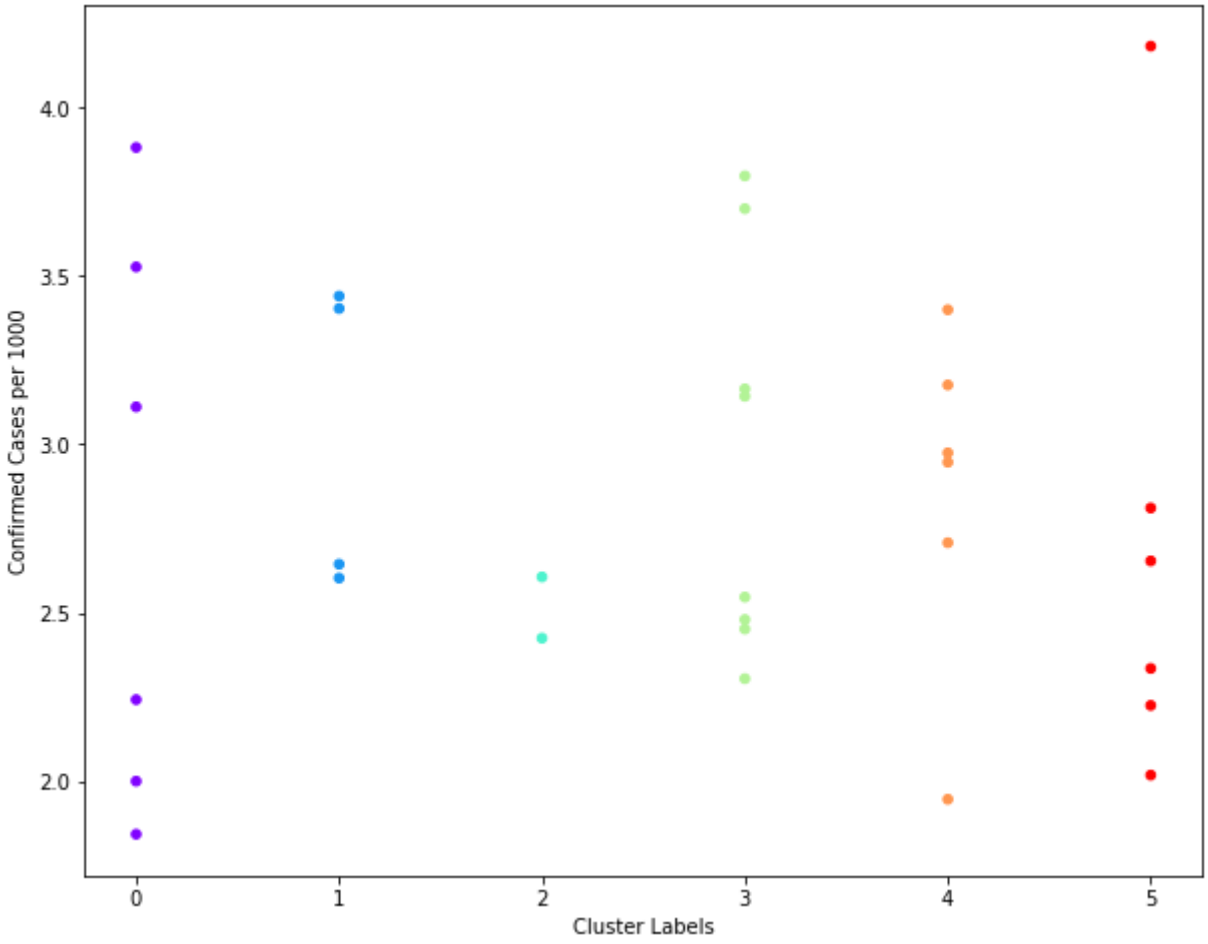

Figure 7: COVID-19 prevalence per 1000 people for each demographic cluster.

5.2. Study 2: Clustering by Venue Profile

The appropriate number of clusters to segment the London boroughs by demographics was determined to be 6 (Figure 8). Again, geospatially, the clusters tend to be neighbours rather than distributed across the city (Figure 9), with a high degree of overlap between the demographic clusters (vs. Figure 5). The most common venues in each the cluster are summarised in Table 3. Comparing coronavirus prevalence between demographic clusters there does not appear to be a significant difference (Figure 6).

Table 3: Most common venues in each cluster.

Cluster	Boroughs	Most Common Venue Types
0	Bromley, Bexley, Havering, and Sutton	Park, Grocery Store, Italian Restaurant, Train Station, Indian Restaurant, Pub, Bakery, Fast Food Restaurant, Burger Joint, Bar
1	Hammersmith and Fulham, Kingston upon Thames, Merton, Richmond upon Thames, and Wandsworth	Pub, Coffee Shop, English Restaurant, Home Service, Playground, Bus Station, Tram Station, Café, Thai Restaurant, Clothing Store
2	Hackney, Haringey, Islington, Lambeth, Lewisham, Newham, Southwark, Tower Hamlets	Pub, Coffee Shop, Café, Bar, Park, Italian Restaurant, Fast Food Restaurant, Restaurant, Hotel, Gym
3	Barnet, Brent, Croydon, Ealing, Enfield, Harrow, Hillingdon, Hounslow, Redbridge, and Waltham Forest	Pub, Coffee Shop, Fast Food Restaurant, Indian Restaurant, Hotel, Pharmacy, Clothing Store, Park, Chinese Restaurant, Grocery Store
4	Camden, Kensington and Chelsea, and Westminster	Pub, Café, Coffee Shop, Italian Restaurant, Burger Joint, Plaza, Clothing Store, Outdoor Sculpture, Garden, Historic Site
5	Barking and Dagenham, and Greenwich	Bus Stop, Grocery Store, Convenience Store, Boat or Ferry, Pub, Pizza Place, Bar, Market, Garden

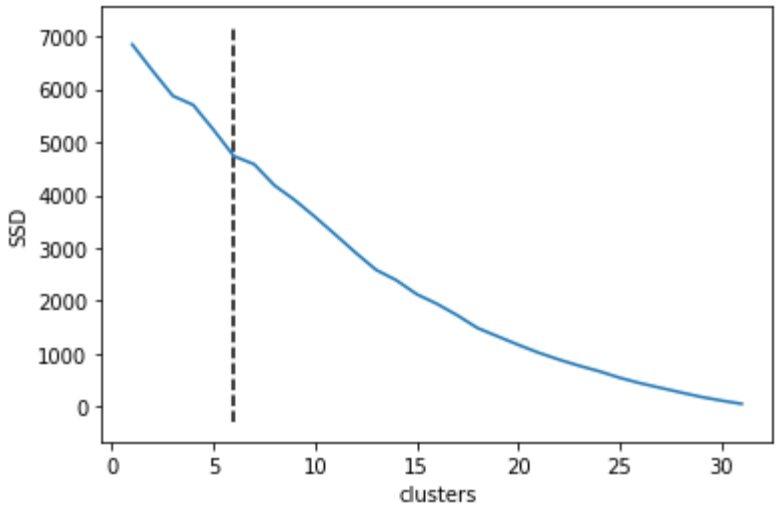

Figure 8: Sum squared distance versus the number of clusters, with the elbow point defined at n=6.

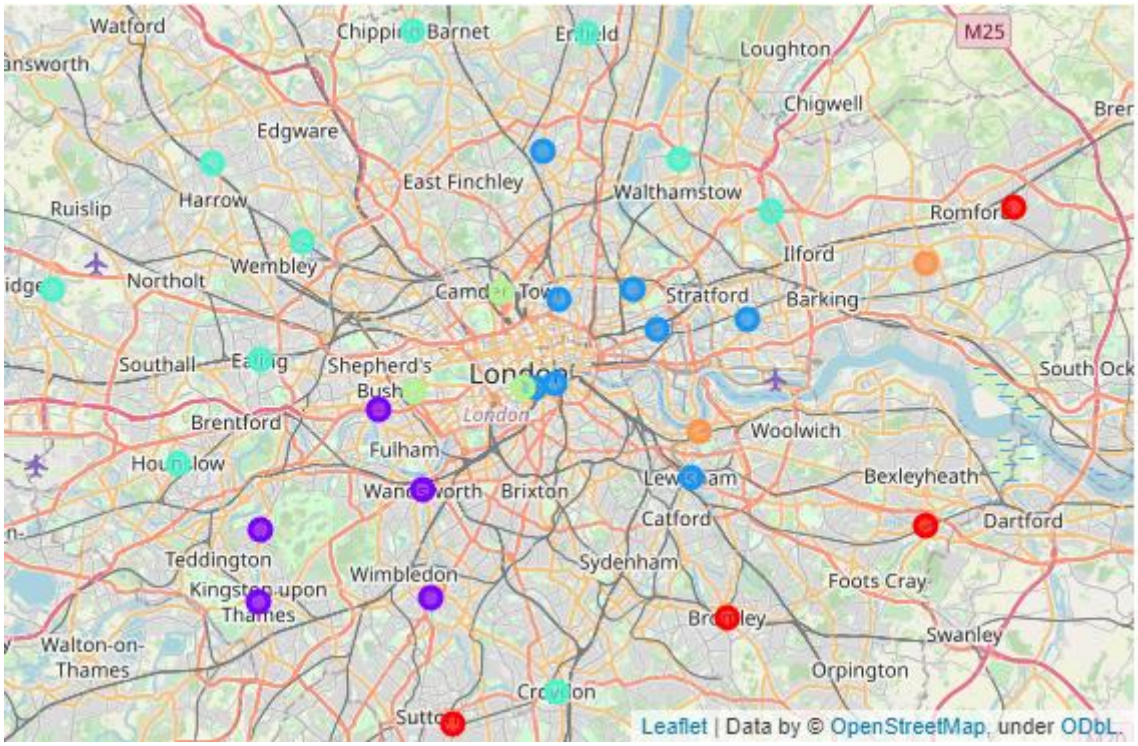

Figure 9: London boroughs clustered by venues.

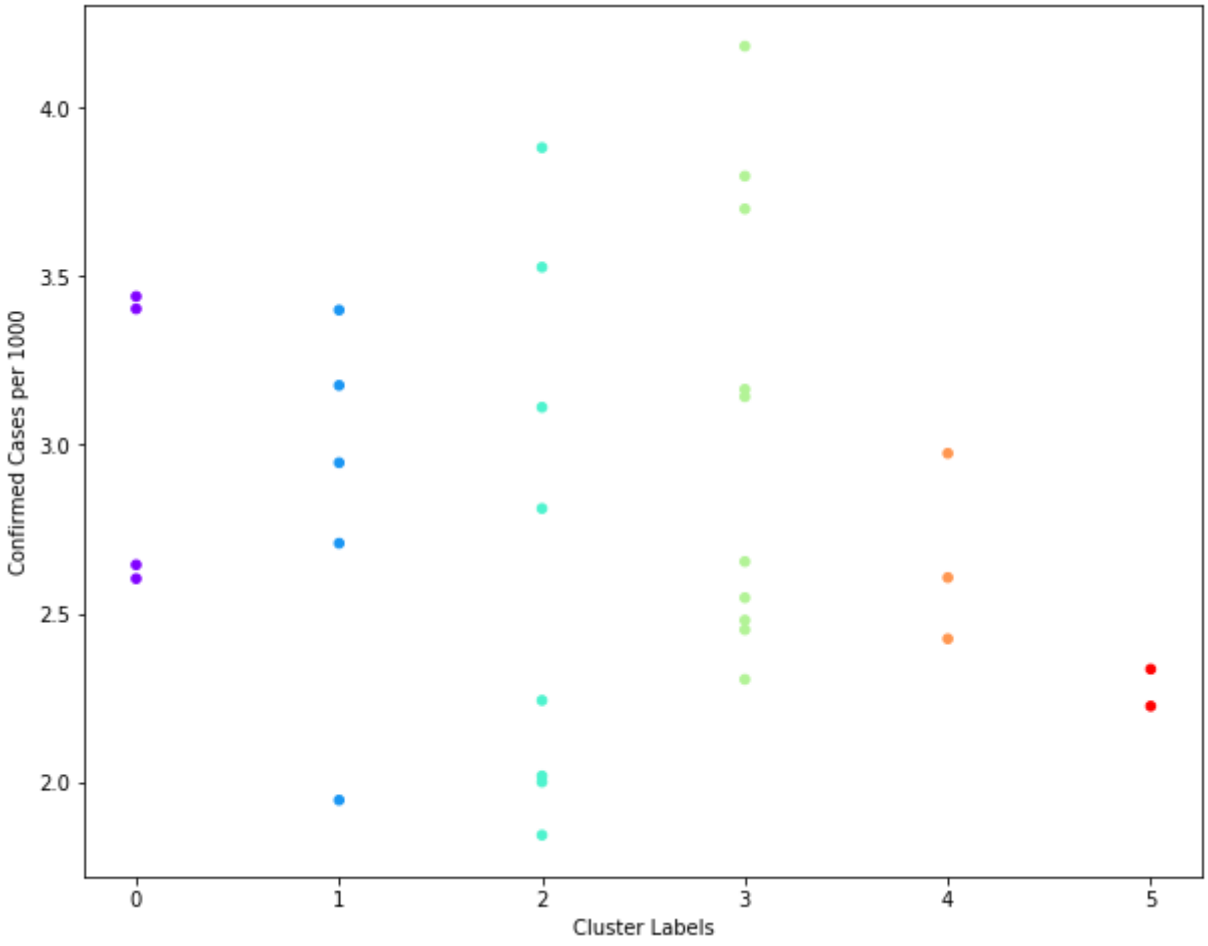

Figure 10: COVID-19 prevalence per 1000 people for each venue cluster.

6. Discussion

The thirty-two London boroughs have been clustered by their demographics and by their venues, with coronavirus prevalence rates compared across clusters. Results from both clustering schemes do not indicate any statistically significant conclusions (Figure 7 and Figure 10) – there is high intra- and low inter-cluster variation. When considering coronavirus management at the level of the boroughs, pre-empting outbreaks does not appear possible with all areas at similar susceptibility. For statistically significant trends to emerge, a greater number of samples would be required beyond the thirty-two London boroughs.

There is a high degree of overlap between the resulting clusters from either scheme, indicating that the demographics of a borough influence its venue types. In both cases, the clusters seemed to contain geospatial neighbours. As examples, demographic cluster 4 contains a set of neighbouring South West London boroughs, whilst venue cluster 2 contains a set of neighbouring boroughs around the East End. Whilst this is a logical result, it does affect the independence of the outbreaks between clusters as cross-transmission between neighbours is likely. For this reason, enlarging the study size to increase other European cities to silo the outbreaks may increase variation in prevalence and expose underlying trends.

Comparing Figures 4 & 8, which show the sum squared difference against the number of clusters in each study, the venue clustering curve is a lot shallower than the demographic clustering curve. This indicates that there is greater homogeneity between boroughs in terms of venue types than for the demographics. Again, this homogeneity is a logical result, the thirty-two London boroughs were created in the 1960s with the intention of grouping areas into town-like boroughs. It is however not helpful and indicates that using Upper Tier Local Authorities, the finest breakdown available from Public Health England, may be too coarse for adequate variation in the data.

Alternative methods for investigating the classification of high- and low-risk areas include Support Vector Machines (SVMs), logistic regression, and decision trees, among others. Such methods are however likely to fall prey to the same traps as this methodology, namely low sample size, sample interdependence and homogeneity in the dataset.

7. Conclusion

Using demographics and location-based social network data is of limited use when coordinating a coronavirus response at the level of an Upper Tier Local Authority in London. All boroughs seem to be at similar propensity to a large-scale outbreak, with the available data too coarse, interdependent, and homogeneous to expose statistically significant trends. The method may, however, produce more significant results when comparing between European urban centres, where characteristics are more varied and outbreaks are more siloed.

References

[1] Public Health England, “Coronavirus (COVID-19) cases in the UK,” 4 May 2020. [Online]. Available: https://coronavirus.data.gov.uk/.

[2] Greater London Authority, “London Borough Profiles and Atlas,” [Online]. Available: https://data.london.gov.uk/dataset/london-borough-profiles. [Accessed 4 May 2020].

[3] Foursquare, “Places API,” 4 May 2020. [Online]. Available: https://developer.foursquare.com/docs/places-api/.