Clustering London Boroughs

Andrew Green

14 July 2019

1 Introduction

The Greater London area is comprised of 32 local government Boroughs and the City of London, which forms a separate adminstrative district.[1]. Greater London covers an area of $1569km^2$ and has an estimated population of 8.8 million in 2018 (see Wikipedia, 2019). The London Boroughs range in character from the glass office blocks of the financial districts in the City of London and Canary Wharf in Tower Hamlets to Bromley with large semi-rural areas.

While many of the landmarks in central London are world famous, many of the London Boroughs will be unfamiliar to non-residents. Can we group the London Boroughs into a small number of clusters which display similar character and can we do so using a data driven approach? Here a combination of data describing the types and age of housing in each area are used together with geospatial information about the types of local facilities in each area to analyse London Boroughs and cluster them. Such an anlysis would be of interest to anyone seeking relocate to London or from one district to another and trying to understand the character of the wider Greater London area.

2 Data

2.1 Data Sources

The analysis uses three sources of data, two files provided by the *London Datastore*, a website maintained by the Greater London Assembly (London Assembly, 2019) and *venue* data accessed through the *Foursquare* API (Foursquare, 2019).

2.1.1 Dwelling Period Data

A csv file containing information about the data of construction of dwellings in the UK was downloaded from the London Datastore, `dwelling-period-built-2014-lsoa.csv`. This file contains the number of properties in each area that were built in one of twelve periods or where the construction date is unknown.[2] The file contains data for the

[1] Henceforth The City of London will be referred to as a London Borough even though technical it is not

[2] The categories are: Pre-1900, 1900-1918, 1919-1929, 1930-1939, 1945-1954, 1955-1964, 1965-1972, 1972-1982, 1983-1992, 1993-1999, 2000-2009, 2010-2014, Unknown.

whole of the UK are various levels of granularity, including data for each London Borough. The file contains data up to the end of 2014.

2.1.2 Dwelling Type Data

The dwelling type data was also downloaded as a csv file from the London Datastore, `dwelling-property-type-2015-lsoa-msoa.csv`. This file contains the number of dwellings of each of five types, bungalows, flats / maisonettes, detached houses, semi-detatched houses and terraced houses. Each of these categories is further subdivided in the file, although the total is provided as a column. As with the dwelling period data, the file provides data for the whole UK at various granularities including data for each London Borough. In addition the file subdivides the data for each area by council tax band as well as providing a summary *All* category.[3] The data was compiled in 2015.

2.1.3 Foursquare Venue Data

Data on local facilities was taken using *venue* search with the Foursquare API. Venues were found within a 1km radius of the centre of each London Borough with a limit of 100 venues retrieved per area. Note that the latitude and longitude of each London Borough was found using the Python `geopy.geocoders` library.

2.2 Data Preparation

No data cleaning was required as all datasets were complete. However, considerable data processing was required to prepare features for clustering analysis.

2.2.1 Dwelling Period Data

As noted earlier, the dwelling period file contained rows relating to areas across the UK and so the first step was to drop all rows that did not relate to London Boroughs. After this, given that the *Unknown* column did not provide any data information this was also dropped. The counts in each period were then aggregated into three new categories, *Pre-20th Century*, *20th Century* and *21st Century.* The counts were then converted to a percentage of the total dwellings per area. Hence the final features were the proportion of all dwellings in each Borough constructed in each of the three catagories.

2.2.2 Dwelling Type Data

The dwelling type data was processed in a similar manner to the dwelling period data, with all rows unrelated to London Boroughs being dropped. In addition the council tax band information was considered superfluous and hence only the *All* category was retained. The sub-categories of dwelling type were not considered useful and hence only the columns relating top the total in each of the five main dwelling types were retained. Finally the counts of each type were converted to the percentage of the total in each London Borough.

[3] Council Tax is the main form of local taxation in the UK.

2.2.3 Foursquare Venue Data

The longitude and latitude of each Borough were used to search for *venues* using the Foursquare API. This yielded 250 unique types of venue across the data set, which were filled very sparsely. To increase the statistical significance of each feature, the venues were processed into each of the 10 parent venue categories defined by Foursquare; Arts & Entertainment, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, Travel & Transport. To do this a map was built between the venue sub categories and the ultimate parent category. The map was constructed by parsing the json file retrieved from the venue categories API `https://api.foursquare.com/v2/venues/categories`. Once the sub-categories were mapped to the parent categories, the sub-categories were dropped. The frequency of each parent venue category was then calculated for each London Borough.

2.2.4 Final Feature Selection

The final set of features for cluster analysis are listed below:

- Arts & Entertainment
- Food
- Nightlife Spot
- Outdoors & Recreation
- Professional & Other Places
- Residence
- Shop & Service
- Travel & Transport
- Pre20C
- 20C
- 21C
- Detached%
- Semi-det%
- Terraced%
- Apartment%
- Bungalow%

All features were defined numerically as perentages and hence lay between 0 and 1.0. Hence no further feature scaling was performed.

References

Foursquare (2019). *Foursquare API.*

London Assembly (2019). *London Datastore.* `https://data.london.gov.uk/`.

Wikipedia (2019). Greater london. `https://en.wikipedia.org/wiki/Greater_London`.