London Boroughs and their Business: A Clustering Analysis

CS910 Foundations of Data Analytics

Estefania Magana
U1856249

***Abstract*—The Analysis of industry clusters has become a popular focus in local and regional development practice. Identify and nurture industry clusters is of great interest to policy makers. As such, this paper seeks to understand the business distribution and find clustering patterns across London boroughs. A dimensionality reduction analysis was implemented followed by a number of cluster selection. An evaluation of different clustering algorithms was made and a deeper analysis into the outperforming clustering methods came after. Findings show that certain London boroughs have a high agglomeration of some industries proving that business clusters can be found in this polarized city.**

I. Introduction

Business clusters have been observed in all sorts of regions (urban and rural areas) and around many core industries (e.g. service, manufacturing, education) [7]. The identification and nurturing of industry clusters is of great interest to policy makers. The clustering phenomena and industry interdependence is of high value in economic development policy, yet it is viewed as a problem to be overcome rather than an advantage in some regions [7]. Incentives for diversification of economies are most times given, but clustering of firms can also represent less dense and less dominant concentration of firms [7]. Cluster's competitiveness may translate into positive factors such as firms paying close attention to quality and trying to invest heavily in innovation and advance technology [7]. However, Business clustering and diversification are not contradictory. Most successful clusters are able to generate new clusters by creating new markets or products based on their core competencies or building the capacities of related sectors [7].

Recasting the state of London in terms of clusters of related industries provides a unique view of its relative specializations, strengths, and weaknesses. The present document reports the implementation of data analysis techniques on the Standard Industrial Classification (SIC) system (i.e. industries and firms similar in product) to detect clustering trends within London boroughs. It also lays the groundwork for subsequent studies of specific clusters and industries.

The structure of this research is as follows: in Section II we define business clustering and outline London's business specialisation. Section III describes some of the works and main findings of previous business clustering applications. Section IV lists the datasets used in the current analysis. Section VII explains data pre-processing. Then, Section VIII explains the first steps of the analysis and shows the number of cluster selection. An evaluation of multiple clustering methods and an analysis of the outperforming ones is shown in SectioN IX, which also includes an implementation of a model-based. We finalize by concluding the findings of the clustering analysis and lay groundwork for subsequent studies of specific clusters in Section X.

II. Background

A. Defining Business Cluster

Researchers and policy makers have given different definitions to the term "cluster" in the industry world [1]. Porter (1998) defines business clusters as "geographical concentrations of interconnected companies, specialist suppliers, survice providers, firms in related industries and associates institutions". Business clusters' have been grouped in many ways such as scale of employment, perceived growth potential and political support [1]. Academic and researches have focused on conducting statistical analyses favouring SIC, agglomeration features and business transactions [7].

B. A deeper look into London

London's current gathering together of related businesses in specific areas is one of its economic landscape defining features [5]. Urban planners are able to divide London's commercial, industrial and residential development zones and incentivise businesses to settle in specific locations [5]. An example of this business clustering is given by the City of London, which is commonly described as a leading international hub of financial trade [5]. Similar stories of traditional and non-traditional business focus and development can be found all over London [5].

London as a major world city has a particular attraction weight [8]. This city is a major metropolis and cultural centre, which makes it an economic attraction [8]

This paper will analyse relevant data sets relating to all types of active businesses on each borough of London, and see if there are industries agglomerations that leads to clustering patterns.

III. Literature Review

Freser and Bergman [4] developed a national set of benchmark or template technological clusters that effectively represent strategically important alignments of underlying detailed sectors in the US. Specifically, the authors deried a set of 23 US manufacturing clustes and employed them as templates in an illustrative analysis of the manufacturing sector in a single US state and one of the top 10 manufacturing ones, North Carolina. This template clusters helped them to detect gaps and specialization in extended product chains and therefore constitute a useful first step in more comprehensive examinations of local cluster patterns.

Taylor et al. [8] analysed the intense concentration and clustering of financial services in the City of London and Canary Wharf. The authors described these areas as the premier European financial district zone and one of the three global international financial centres. The concluded that the financial clustering in City of London and Canary Wharf is going to continue disproportionately in relation to other areas in central London.

IV. The Dataset

The main data sources used to cluster London Boroughs by their business composition were: a directory of London businesses, London Boroughs profiles and London Boroughs boundary file information from the London Data Store. This paper also relied on the Standard Industrial Classification (SIC) from the Office for National Statistics.

A. Directory of London Businesses

The Directory of London Businesses is a dataset containing businesses located in London showing a range of information including company name, address, postcode and SIC code. This data needed further formatting and cleaning that is going to be explained on Section VII

B. London Borough Profiles

London Borough Profiles database includes a wide description of each of the boroughs in London. The use of these information was restricted to obtained London boroughs codes. This database was available for immediate implementation.

C. Statistical GIS Boundary Files for London

London Borough 2014 boundaries shape file from the London Data Store contains geographic information that was mainly used for visual representation purposes. This document was immediately accessible and ready to use.

D. Summary of SIC structures

The current SIC classifies business establishments and other statistical units by the type of economic activity in which they are engaged. Of particular interest was the wides industrial classification. This data needed further formatting and cleaning that is going to be explained on Section VII

V. Software & Techniques

The data obtained is a mixture of Comma Separated Value files, GIS shape files and Microsoft Excel formatted files. This document applies the following tools to analyse the data:

A. R

The main use of this statistical computing software was: cleaning, merging and handling databases (packages: data.table, plyr, stringr), principal component analysis (package: stats), clustering analysis (packages: cluster, hmisc, mclust, factoextra, clValid), and data visualization (packages: ggplot2, gridExtra)

B. Microsoft Excel

Microsoft Excel main use was to treat data bases that could not be automatized and required manual cleaning and formatting.

C. QGIS

The use of this geographic information system was for visual representation of clusters distributions.

VI. Hypothesis

With the knowledge that London is a polarized city, it follows that there is going to be a high level of clustering patters between London's Boroughs. We expect to find an intense administrative driven businesses int the central boroughs and more diverse business sector in the surrounding areas.

VII. Data Cleaning

We began by reviewing that all the borough codes included in the directory of London business database were included in London Borough Profile database, both provided by London Database. We followed to check the status of the included companies. Some of the listed companies in this directory database had liquidity problems, issues with creditors or have not traded or sold off any stock in the three months or more. For these reasons, we decided to restrict this paper's analysis to active businesses.

Once the active businesses were selected from the directory dataset, we continue to review SIC businesses classification. London Database managed their businesses information by classes or sub classes, which are the most granular division according to the Office for National Statistics. Given that the aim of this research is to find business clusters in London wards, which are a reduced geographic area, a wider and more general classification was needed. The Office for National Statistics provides a summary of SIC structures which contains: (i) Industry Section, (ii) Industry Division, (iii) Business Group, (iv) Business Class, and (v) Business Sub Class. This research was particularly focused on the Industry Sections. The office for National Statistics provide an excel file with a complicated format that needed manual work in order to be converted into a workable csv for further analysis in R.

Around 200 businesses had and industry class that was not defined in SIC business classification so manual classification had to be made by looking the directly into UK's Company Database.

VIII. Data Processing

A. Standarisation

To begin the exploratory analysis of the active businesses in London included on the directory, the distribution of the industries and the distribution of London's Bouroughs were checked. In Figure 1 one can see that businesses categorized as "Professional, scientific and technical activities" are the most common ones all over the city.

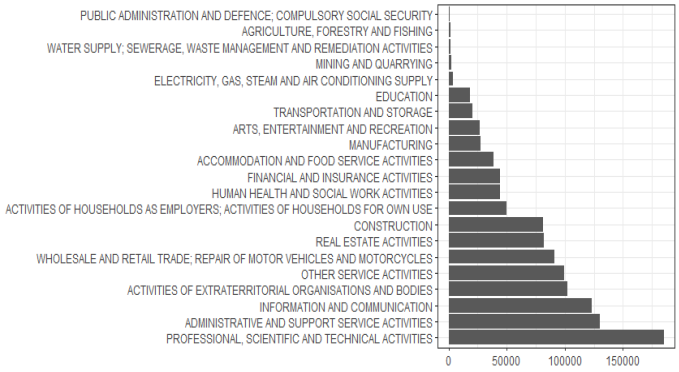

Fig. 1. "Standard Industrial Classification (SIC) Distribution"

In Figure 2 we can observe that Westminster has considerably more businesses than the rest of London's Borough. Is also noticeable that the difference of businesses presence between boroughs is significant, standardisation is needed in order to avoid the size of the Borough's to skew the clustering analysis.

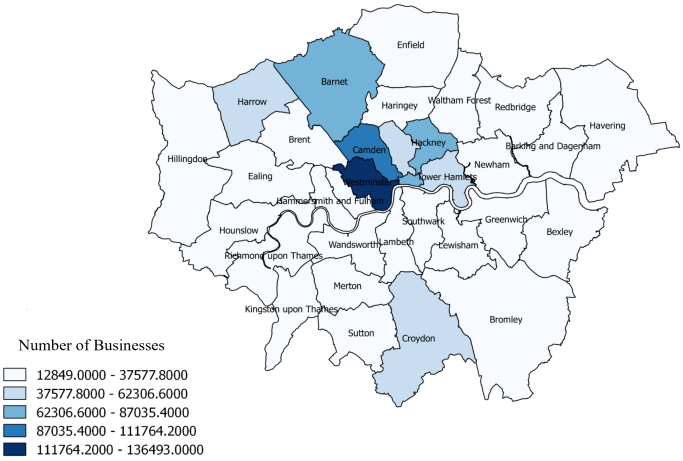

Fig. 2. "Number of Businesses in London"

B. Correlation

We then review the correlation between the industries classifications. Even though collinearity is a problem normally related to regression analysis, it also can complicate any clustering study. In clustering, high level of correlation between two variables (collinearity), implies a higher weight of whatever concept they are both (or more) explaining. Therefore, the solution is likely to be skewed in the direction of the concept that the correlated variables are explaining.

The correlation coefficient and p-values were computed for each pair of SIC. Figure 3 shows the level of the statistically significant correlations. One can see that industries A, B, D, E and O have a strong positive correlation. These industries are: (i) Agriculture, forestry and fishing, (ii) Mining and quarrying, (iii) Electricity, gas, steam and air conditioning supply, (iv) Water supply; sewerage, waste management and remediation activities, and (v) Public Administration and defence; compulsory social security. The analysis of these correlation is beyond the scope of this research.

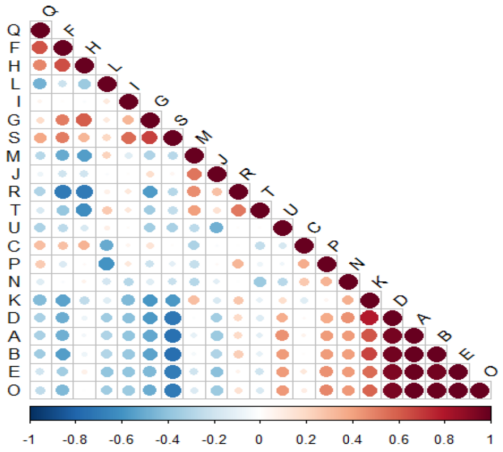

Fig. 3. "Correlation Between Industries"

C. Dimensionality Reduction

In order to work with uncorrelated vectors and find stronger patterns to link London's boroughs by their industries compositions, we now reduce the dimensionality of the data set by applying a Principal Component Analysis (PCA). PCA picks up the dimensions with the largest variances among the dataset generating a noise and dimensionality reduction [3].

Table 4 shows the results of the PCA subspace summarizing the standard deviation, proportion of explained variance and the cumulative proportion of each of the vectors. For this task we selected those the first 7 vectors to get more than 95% cumulative proportion of variance.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation	1.154	1.027	0.827	0.524	0.504	0.41	0.33
Proportion of Variance	0.328	0.26	0.168	0.068	0.063	0.041	0.027
Cumulative Proportion	0.328	0.588	0.757	0.824	0.887	0.928	0.955

Fig. 4. "Principal Component Analysis Table"

D. Optimal Number of Clusters

Next, we apply three methods to determine the optimal number of clusters: (i) elbow method, (ii) silhouette method and (iii) gap statistic.

1) Elbow Method: The idea of this method is to define the optimal number of clusters by minimizing the intra-cluster sum of squares (intra-ss). The optimal number of clusters is reached when the decrease of the intra-ss of one more cluster is lower than the previous ones. In Figure 5 one can observe the elbow method result comparation for 1 to 10 clusters made by K-mean, K-centres and Hierarchical clustering methods. In this, we have that 2, 3 and 4 clusters are the optimal ones for k-means, k-centres and hierarchical clustering method, respectively.

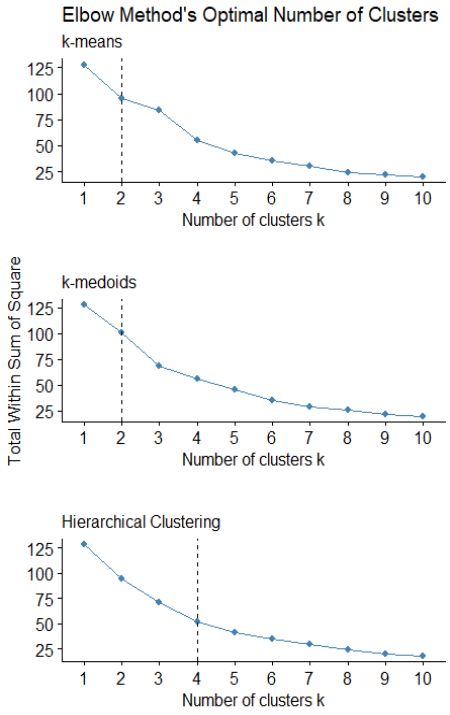

Fig. 5. "Elbow Method's Optimal Number of Clusters for k-meand, k-medoids and Hierarchical Clustering"

2) Silhouette Method: Silhouette method determines how well each object lies within its cluster, so a high average silhouette width indicates a good clustering. The optimal number of clusters is the one that reaches the maximum level of the average silhouette within a range of possible cluster numbers. In Figure 6 one can see that the optimal number of clusters between 1 to 10 are 2, 7 and 2 for k-means, k-centres and hierarchical clustering methods, respectively.

3) Gap Statistic: Gap Statistic Method compares the total intra-cluster variation for different cluster numbers with their expected values of data coming from a distribution with no obvious clustering (null reference distribution of the data). The reference dataset is generated using Monte Carlo simulations of the sampling process. That is, for each variable (x_i) in the dataset, n points uniformly distributes are created from the range $[min(x_i), max(x_i)]$. The gap statistic for a given k number of clusters is defined as:

$$Gap_n(k) = E_n^*\{log(w_k)\} – log(w_k) \quad (1)$$

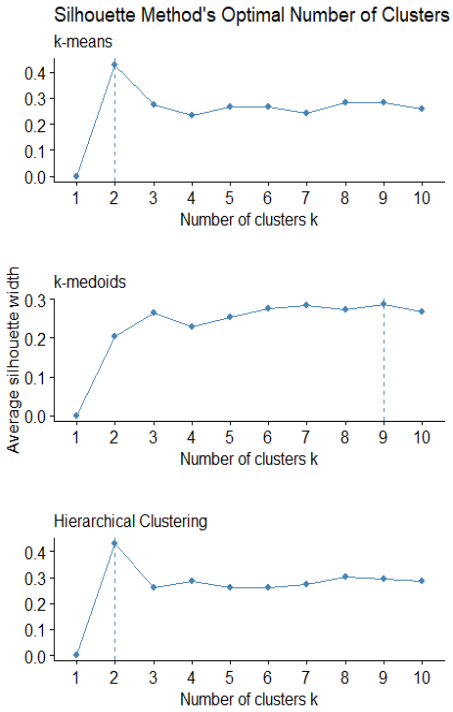

Fig. 6. "Silhouette Method's Optimal Number of Clusters for k-meand, k-medoids and Hierarchical Clustering"

Where w_k is the total intre-cluster variation for a given k number of clusters (total within sum of square) and E_n^* is the expectation under a sample of size n from the reference distribution. In Figure 7 one can see that according to Gap Statistic's method, the optimal number of clusters for k-means, k-centre and hierarchical clustering methods is one.

IX. Data Analysis

With the previous analysis, we now have the option to cluster our data base from 2 groups up to 9 groups. Now the question is which combination of method and number of clusters is the optimal one for London's businesses.

A. Internal Validation

The goal of clustering is to group objects in the same cluster that are as similar as possible and have differences between clusters as high as possible [2]. Internal validations indexes reflect how compact is the average distance within clusters and separated the clusters within each other. The three commonly applied measures are: compactness, separation and connectedness [2].

a) Compactness measure: evaluates how close are the object within the same cluster, so a lower intra-cluster variation is a good compactness indicator.
b) Separation measure: determines how well-separated are all the clusters with each other by computing: (i) distance between cluster centres, and (ii) the pairwise minimum distances between objects in different clusters.

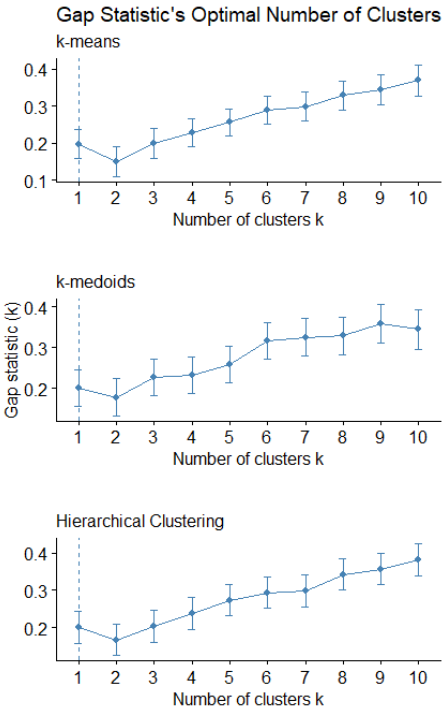

Fig. 7. "Gap Statistic's Optimal Number of Clusters for k-meand, k-medoids and Hierarchical Clustering"

c) Connectedness measure: reflects the extent to which items that are placed in the same cluster are considered their nearest neighbour.

Since compactness and separation demonstrate opposing trends, the majority of the internal clustering validation indices combine these two measures into a single score as follows [2]:

$$Index = \frac{\alpha\ Separation}{\beta\ Compactness} \tag{2}$$

With α and β weights.

With the R package $clValid$ we now computed the three following internal validation indexes:

1) Connectivity: reflects the connectedness measure of the clusters and it should be minimised.

2) Silhouette Width: estimates the average distance between clusters by measuring how close each point in one cluster is to points in the neighbouring clusters. Values as close to 1 as possible are most desirable.

3) Dunn Index: represents the ratio of the smallest distance between observations in different clusters to the larges intra-cluster distance. This index should be maximized.

Figure 8 shows the optimal internal validation scores and one can observe that hierarchical clustering with two clusters performs the best according to Connectivity an Silhouette indexes, and k-medoids the outperforming clustering method with 9 clusters according to Dunn Index.

B. Stability

Stability measures assess the suitability of the clustering algorithm by testing how sensitive it is to perturbations in the input data [2]. $clValid$ package removes one column of the data at a time and re-run the clustering. The included measures are: average proportion of non-overlap (APN), average distance (AD), average distance between means (ADM), and figure of merit (FOM), all of which should be minimised [2].

	Score	Method	Clusters
Connectivity	5.42	Hierarchical	2
Dunn	0.50	k-medoids	9
Silhouette	0.43	Hierarchical	2

Fig. 8. "Internal Validation Optimal Scores by $clVal$ R package"

1) APN: measures the average proportion of observations not placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed. Is a number in the interval $[0, 1]$, with values close to zero corresponding with a highly consistent clustering result.

2) AD: measures the average distance between observations placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed. The AD measure is a balue between zero and $\inf$ and smaller values are preferred.

3) ADM: measures the average distances between cluster centers for observation placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removes. It has a value between zero and $\inf$, with smaller values equalling better performance.

4) FOM: measures the average intra-cluster variance of the observations in the deleted column of the data, where the clustering is based on the remaining (undeleted) samples. The final score is averaged over all the removed columns and it has a values between zero and $\inf$, and smaller values are preferred.

Figure 9 shows the optimal stability scores and one is able to notice that in line with internal validation scores, hierarchical clustering with two clusters and k-medoids with nine clusters are the outperforming clustering methods.

	Score	Method	Clusters
APN	0.02915	Hierarchical	2
AD	1.13198	k-medoids	9
ADM	0.15516	Hierarchical	2
FOM	0.60513	k-medoids	9

Fig. 9. "Stability Optimal Scores by $clVal$ R package"

C. k-medoids

We now clustered London's Boroughs into nine groups with k-medoids algorithm. Figure 10 shows the formed clusters and one is able to notice that Westminster and City of London are clustered into their own group each, and the rest of the groups go from two to eight boroughs.

In Figure 11 the distribution of each SIC within each cluster is shown. SIC names were changed into shorter names to

allow a better visualization. In here, one can observe that businesses with professional, scientific and technical activities are relatively common in all boroughs. Cluster 1 (City of London) as expected, is defined by a high level of financial and insurance activities and administrative support activities related businesses. Cluster 2 is mainly defined by construction businesses which matches with high percentage of homes owned boroughs. Cluster 3, the biggest one, in general is formed by boroughs whit a more varied businesses activities but none is particularly higher that the rest (besides professional, scientific and technical activities). Cluster 4 has a higher level of information and communication businesses, which relates with the fact that Islington borough is part of this cluster where Old Street, the Silicon Roundabout and house of tech and web-focused companies, is located. Cluster 5 is formed by the two furthest boroughs (northern and southern) and is greatly explained by businesses with activities of extraterritorial organisations and bodies. Cluster 6 does not have a particularly strong industry and is not that different from cluster 3. Cluster 7 is characterized by administrative and support service activities businesses while cluster 8 has a particularly greater amount of professional, scientific and technical activities than the rest of the clusters. Cluster 9 (Westminster) is mainly defined by businesses with activities of extraterritorial organisations and bodies.

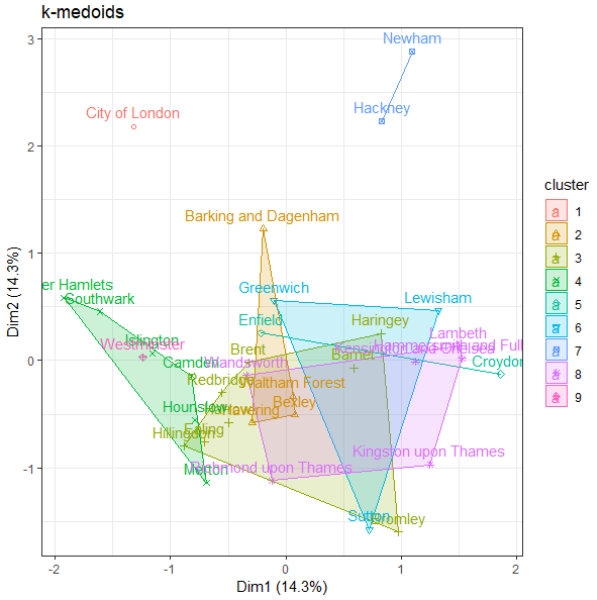

Fig. 10. "Map of k-medoids Clusters"

D. Hierarchical Clustering

For the hierarchical clustering algorithm, the three distance measure were applied and results are shown in Figure 12. Single-link give really unbalance clusters by neighbouring City of London in a one borough cluster and all the other boroughs in another. Complete-link and Average-link cluster London's boroughs the same and, even though they are still far from

Fig. 11. "SIC Distributions in k-medoids Clusters"

balance, the borough distribution is better. For this reason, we stick to this clustering method for further analysis.

Figure **??** shows the distribution of each SIC within each cluster. When analysing the two clusters formed by hierarchical clustering algorithm, one is able to identify that cluster 2 is mainly characterized by businesses with activities of extraterritorial organizations and bodies and Westminister borough is included. Also, cluster 2 SIC distribution variation is lower than in cluster one which might be explained by the fact that cluster 2 is the smallest group, formed by three boroughs. Cluster 1 strongest businesses presence are those related to professional, scientific and technical activities. In general, this cluster does not seem to have a really strong industry but it one can observe that the presence of administrative and support service activities related businesses is stronger in these boroughs than the ones in cluster 2.

E. Model Based Clustering

We now continue to apply a model-based clustering method: Expectation Maximization (EM) algorithm. For this, we used *mclust* R package which allows modelling of data as a Gaussian finite mixture. Figure 14 shows a summary of the best model selected using the Bayesian Information Criterion (BIC) and Figure 15 shows the clusters obtained by the selected model.

When analysing the SIC distributions within each cluster showed in Figure 16, one can identify that cluster 1 is mainly characterized by a strong presence of businesses with activities

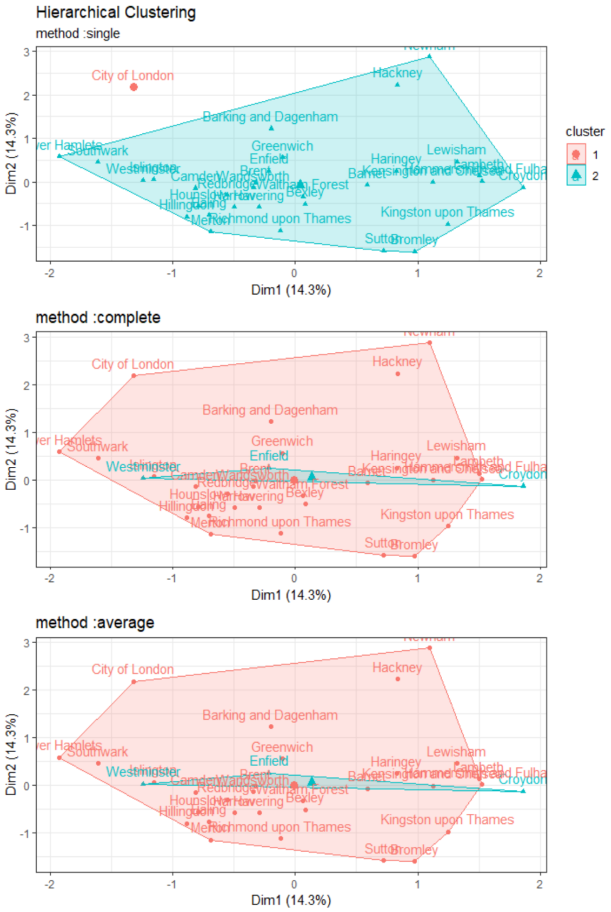

Fig. 12. "Hierarchical Clustering"

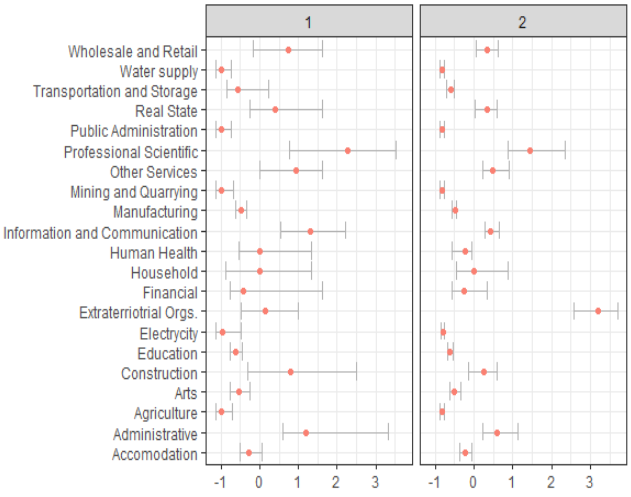

Fig. 13. "SIC Distributions in Hierarchical Clustering"

```
-------------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
-------------------------------------------------------

Mclust VEV (ellipsoidal, equal shape) model with 6 components:

 log.likelihood  n  df       BIC       ICL
        152.748 33 185 -341.3579 -341.3579

Clustering table:
1 2 3 4 5 6
4 7 8 6 6 2
```

Fig. 14. "Expectation Maximization Clustering"

of extraterritorial organisations and bodies and we find Westminster borough. Cluster 2 seems to have a greater presence of professional, scientific and technical activities than the other clusters. Cluster 3 does not seem to have any particularly strong industry in their boroughs, but the SIC distribution in this cluster seem to have a reduced variation. Cluster 4 has a similar SIC structure than cluster 3 with no relatively high agglomeration of any industry, besides professional, scientific and technical activities related businesses which is strong in all clusters. Cluster 5 shows a high presence of information and communication businesses and we again find in this cluster London's tech hub, Old Street. Cluster 6 is mainly defined by administrative and support service activities and, surprisingly, this time City of London is not included. In this case Hackney and Newham boroughs are the selected ones.

X. Conclusion

As hypothesised, it can be concluded that there is an agglomeration of some particular industries in the London's boroughs. While the Standard Industrial Classification is quite general, there is a consistently high level of differentiation in administrative and support service activities related businesses and businesses with activities of extraterritorial organisations and bodies. However, the strength of the business agromeration was not as strong as expected. In some of the clustered algorithms with larger amount of clusters (k-medoids and EM), two specific clusters were found to have similar business profiles, which is worth for a further analysis to explore the option of joining the together and make the proper evaluations.

Clustering, as an unsupervised technique, complicates the method selection and the number of clusters determination. A variety of measure we applied to validate the results of the clusters, but the objective of the clustering analysis would determine which algorithm performs the best. For. This research presented a variety of methods for validating the results from multiple clustering methods and a deeper analysis into de outperforming ones showed that each formed cluster presented a particular industry structure.

An extension can be made to find not only single industries clusters but to expand the analysis to related industries as businesses find it advantageous to be close to their suppliers, customers, services and competitors.

References

[1] Bergman, E., Feser, E., & Sweeney, S., 1996. "Targeting North Carolina manufacturing: understanding the state's economy through industrial cluster analysis." *University North Carolina Institute for Economic Development, Chapel Hill, NC.*

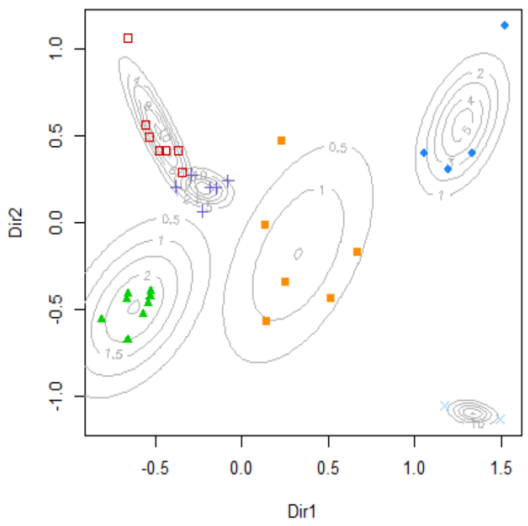

Fig. 15. "Expectation Maximization Summary obtained by *mclust* R package"

Fig. 16. "SIC Distributions in Expectation Maximization Clusters"

[2] Brock, G., Pihur, V., Datta, S. and Datta, S., 2011. "clValid, an R package for cluster validation." *Journal of Statistical Software* (Brock et al., March 2008).

[3] Ding, C. and He, X., 2004, "July. K-means clustering via principal component analysis." *Proceedings of the twenty-first international conference on Machine learning* (p. 29). ACM.

[4] Feser, E.J. and Bergman, E.M., 2000. "National industry cluster templates: a framework for applied regional cluster analysis." *Regional studies, 34(1)*, pp.1-19.

[5] Nelson, S. 2016. Pearl & Coutts. [Online]. [17 December 2018]. Available from: https://www.pearl-coutts.co.uk/the-insider/resources/types-of-business-clusters-in-london/

[6] Porter, Michael E., 1998. "Clusters and competition. new agenda for companies." *Governments and Institutions, in: Ibid., On Competition.*.

[7] Rosenfeld, S.A., 1997. "Bringing business clusters into the mainstream of economic development." *European planning studies*, 5(1), pp.3-23.

[8] Taylor, A. J., et al., 2013. "Financial services clustering and its significance for London."

[9] London Datastore.

[10] Office for National Statistics

[11] UK Company Database