Geodemographic Classification

Course Work 2

Abstract

This study builds on the Geodemographic Classification based on the 2011 UK Census data. Geodemographic classifications are a way of categorizing people based on their geographic location and demographic characteristics. They are used in market research, advertising, and urban planning to help understand the lifestyles, behaviours, and preferences of people in different regions. Geodemographic classifications use a variety of data sources, including census data, consumer surveys, and lifestyle data, to create profiles of different areas and the people who live there. These profiles can be used to target marketing campaigns, identify potential customers, and plan services and infrastructure. The aim of this study is to develop a classification based on socio-demographic characteristics of the 3,223 Output Areas of the city of Birmingham using the recently published 2021 UK Census data from the Office of National Statistics. One of the most common techniques used in similar studies – the k-means algorithm, a form of partial clustering unsupervised learning approach is deployed on processed data of 8 socio-demographic categories – Ethnicity; Car or Van availability; Socio-Economic activity; Housing type; Dimensions of Household Deprivation; Age structure; Highest educational qualification; and Population density. For the percentage normalised and z-score transformed data, the 34 variables chosen from the 8 categories resulted in creation of 6 clusters in the R programming language, indicating the trends in the demographics seen in other similar cities.

Introduction

Geodemographic classifications offer concise measures of the demographic, economic, social, and built features of small regions. Some common examples of geodemographic classifications include the ACORN system in the UK and the PRIZM system in the US. These systems use a combination of demographic, socioeconomic, and behavioural data to create detailed profiles of neighbourhoods and the people who live in them. (Burns *et al.,* 2018) One can argue that all geodemographic classifications are variations of a common procedure involving data acquisition, manipulation, transformation, and cluster analysis. Therefore, geodemographic classifications can be distinguished based on factors such as the data source, data standardization and transformation techniques, clustering methods, and the existence of accompanying descriptive "pen portrait" materials. (Gale *et al.,* 2016) The 2021 Census data on ethnicity disclosed that Birmingham is among the earliest "super diverse" cities in the United Kingdom, where over 50% of the population is composed of citizens from ethnic minority backgrounds, with 187 nationalities being represented. Birmingham, being a core city and employment hub in the West Midlands, is intricately connected with its surrounding regions, forming an interdependent network that constitutes the broader functional economic geography of the city-region. This connection is particularly significant in terms of the city's economy and labour market. (Birmingham City Council, 2022) (Greater Birmingham Chambers of Commerce and University of Birmingham, 2018) The diverse nature of the city and an optimal geographic location with a bustling economic culture is the reason for the choice for this study. The process of clustering refers to an unsupervised machine

learning task, which autonomously separates the available data into distinct clusters, and is used for knowledge discovery by finding natural groupings within data. (Sabbata, 2022) The k-means clustering algorithm is used building on previous studies, although other methods like Fuzzy C-means, m-logit model, etc may be applied to achieve similar results. (Chris Brunsdon and Alex Singleton,2015)

Literature Review

(Gale *et al.,* 2016) presented an updated version of the 2001 Output Area Classification (OAC) methodology, which incorporates data from the 2011 UK Census to provide a summary of the social and physical makeup of neighbourhoods, using k-means clustering algorithm. (Yang, Dolega and Darlington-Pollock, 2022) described the implementation of the Ageing in Place Classification (AiPC) in England, a geodemographic classification system that operates in multiple dimensions and employs an extensive array of spatially relevant attributes to represent the sociodemographic features and living conditions of older individuals in LSOAs. This classification uses k-means clustering algorithm, and is updatable and reproducible for other countries as well. (Harris, Johnston and Burgess, 2007) used the results of k-means clustering approach to study the effects of background of neighbourhoods, distances and the ethnicities on pupils attending schools in Birmingham. (Singleton and Longley, 2015) built on the criticism of national level geodemographic classification to conduct a classification on higher resolutions for Greater London based on k-means clustering approach, and thus being the motivation to choose Output Areas as the resolution for the analysis in Birmingham. This study takes variables from 4 attributes – Demographic, Housing Characteristics, Socio-economic traits, and Employment attributes. (Vickers, 2008) The case study of Liverpool most closely resembles the current study in terms of variable selection and elimination, data manipulation and processing, while the interpretation of the spatial autocorrelation of patterns played an important role for interpreting the clusters in this study. (Alexiou and Singleton,2015)

Methods

The following socio-demographic categories were chosen for the geodemographic classification for 3,223 Output Areas (OAs) of Birmingham:

1. Ethnicity
2. Car or Van availability
3. Socio-Economic activity
4. Housing type
5. Dimensions of Household Deprivation
6. Age structure
7. Highest educational qualification
8. Population density

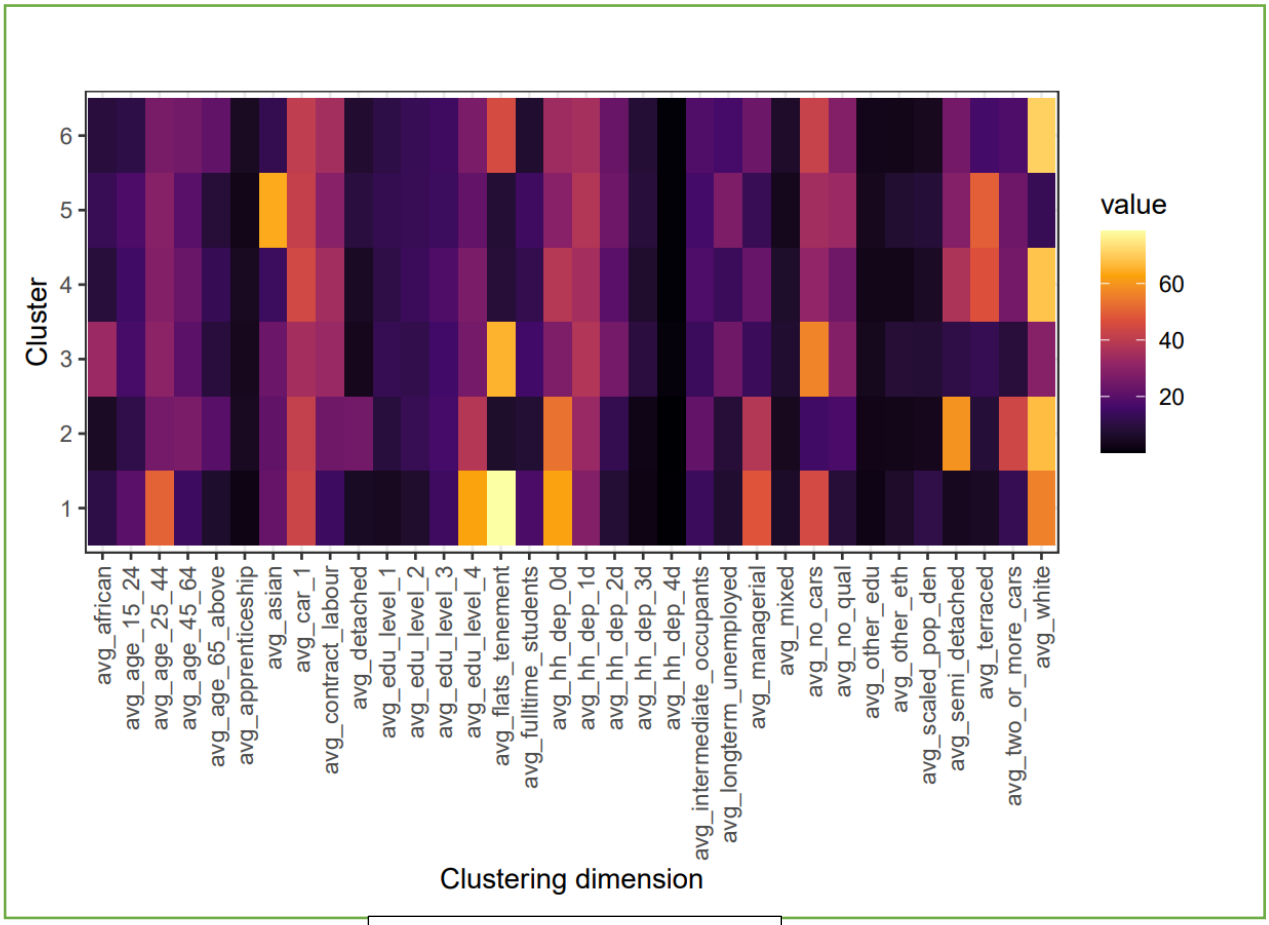
The OA geojson file and the variables in percentages for the above categories was obtained from. (Office for National Statistics, 2023) The data was pre-processed, and an exploratory analysis in terms of correlation test for all the variables was conducted in R. Some of the highly correlated variables were dropped along with outliers, while others that helped retain the variability in the data were either retained or aggregated for further analysis. For creating a composite indicator to compare the demographic situation of the communities, so as to have the indicators portray an equal influence, a z-score or normal score transformation was carried

out. (Zach, 2021) While the population density being already a normalised variable did not show much variation along with the percentage based classification, the z-score standardization technique helped in population density being a discerning factor in clustering.

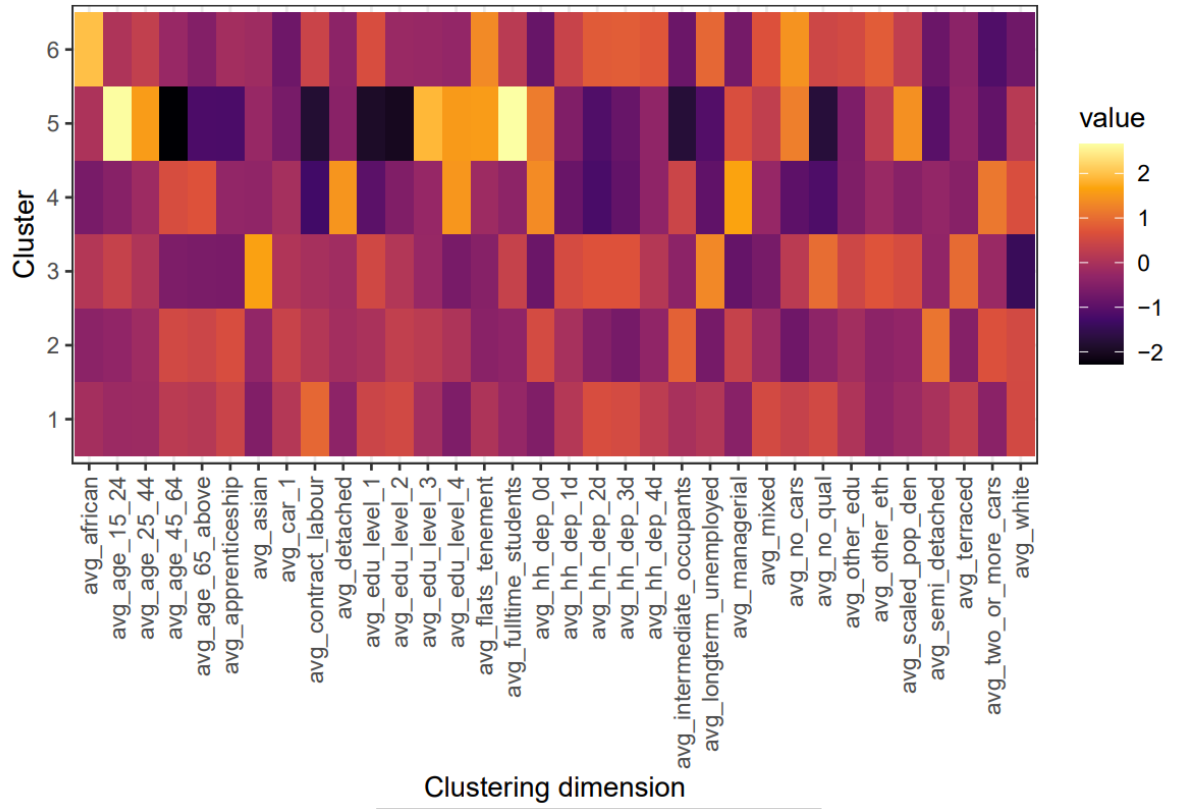
The k-mean method clusters observations into k clusters by minimizing the within-cluster sum of squares (WCSS) through an iterative procedure. In this approach, the algorithm determines the distance between each observation (i.e., case, object, row) and the centroid of its respective cluster. It then sums the squared values of these distances for each cluster and subsequently for the entire dataset. (Ahmed, Seraj and Islam, 2020) The number of clusters = 6 were chosen based on 3 heuretics – the "elbow method", the silhouette and the gap statistic, all 3 indicated a value between 3 & 6 clusters, with 6 being the optimal choice. (Sabbata, 2022) The pair plots, heatmaps, radar charts and the geodemographic classification maps for both the percentage data and z-transformed data were generated using appropriate R libraries.

R being an open source, universally accessible and license-free programming language, along with the freely available census data and boundaries, were the choice for this study, making it reproducible.

Results



Percentage Heatmap



Z-Transformation Heatmap

Table 1	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Ethnicity	White	White	White/African/ Asian	White	Asian	White/Other
Car-van availability	<=1	>=1	<=1	<=1	<=1	<=1
Socio-economic	Managerial	Managerial	Contract/Full-time students	Contract	Long-term unemployed/Full-time students	Long-term unemployed
Housing type	Flats	Semi-detached	Flats	Terraced/Semi-detached	Terraced/Semi-detached	Semi-detached/Flats
Dimensions of Deprivation	0-1	0-1	1-2	0-1	0-2	0-2
Age structure (yrs)	25-44	25-64	25-44	25-65+	25-44	15-44
Highest level of Qualification (Level)	4	4	0	0,4	0	4
Population Density	Low	Low	Low	Low	Low	Low
Table 2	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**

Ethnicity	Mixed	White	Asians/Other	White	Other/Mixed	African/Other
Car-van availability	<=1	>=1	<=1	>=1	0	0
Socio-economic	Long-term unemployed	Intermediate-managerial	Full-time students	Managerial	Full-time students/Managerial	Long-term unemployed/ Contract
Housing type	Terraced	Semi-detached	Terraced	Detached	Flats	Flats
Dimensions of Deprivation	1-4	0-1	1-3	0	0	1-4
Age structure (yrs)	Mixed	25-64	15-44	25-65+	15-44	25-44
Highest level of Qualification (Level)	0-3	1-3	0-1	4	3-4	0-1
Population Density	Medium	Medium	High	Low	High	High

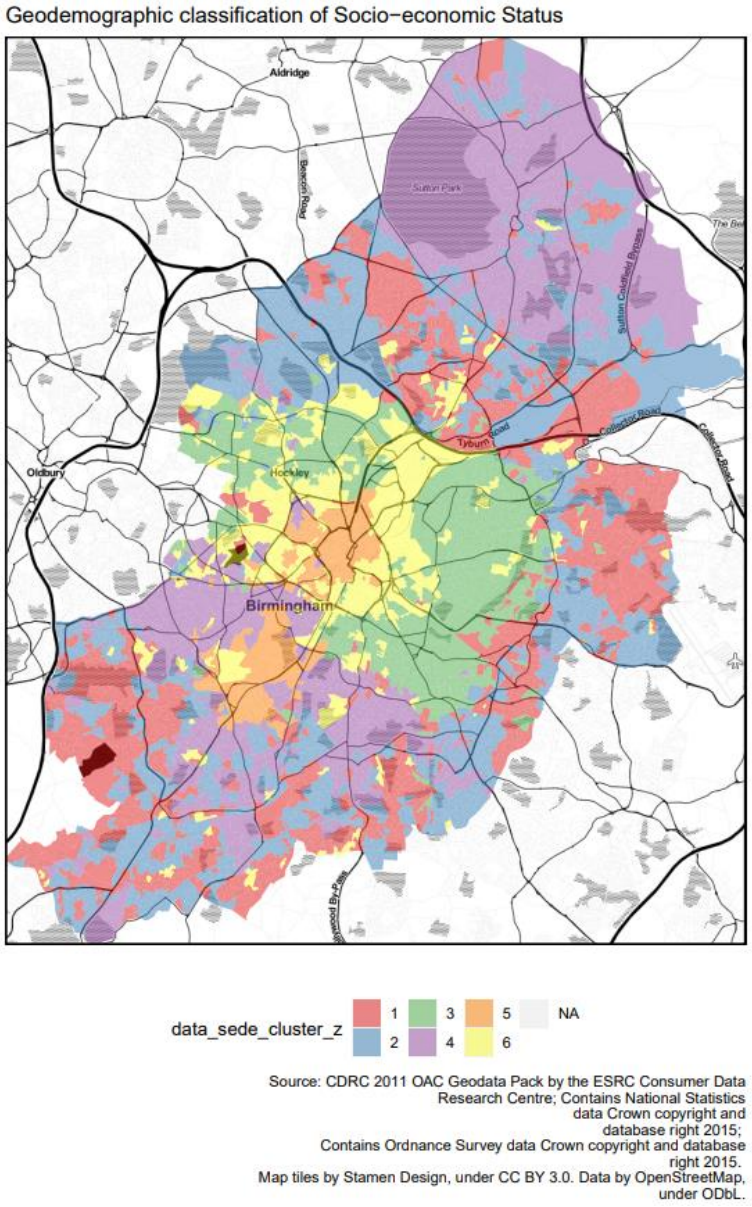
Table 1: Percentages. Table 2: Z-Score Transformation

The heatmaps and the clustering tables are provided for both Percentage based classification and z-score transformation based classification, however, the cluster formation and spatial autocorrelation turned out to be inexplicable.

Discussion

The names and descriptions of clusters are a crucial aspect of the user interface of geodemographic classifications. They aim to reflect the intricate nature of cluster compositions. (Gale *et al.,* 2016) However, since this study uses the recent 2021 census data, and is not required to be in a public domain with users interacting with the results and analyses, the naming of the clusters is not considered vital, and therefore has been omitted. For the sake of convenience, only the Z-score transformation based classification (map below) is being discussed. The map and radar plot for the Percentage based classification is present in the Rmd. File. Perhaps the most prominent cluster that explains the socio-demographic characteristics of the area is Cluster 5. It is noticed that ages 15-44 years reside in these areas, with many ethnicities living together usually in flats and in high population density areas. These are either full-time students or people in managerial positions having low car ownership, high level of education qualifications and very low deprivation. The areas containing the cluster have two major universities – University of Birmingham and Aston University, which explains the presence of young people with high level of education and no car ownership living in flats. Being close to the city centre, these could also possibly be the locations where young managers reside. Similarly, Cluster 4 explains the sub-urban dwellings away from the city centre with affluent white population containing retirees living in detached houses and owning one or more cars. These people live in low population density areas with the residents having a high level of education and negligible deprivation in the areas, a trend seen in other cities as well, with other factors like low levels of pollution as well. Clusters 1,3 & 6 contain a majority

of other ethnicities, with a struggling population living in areas of high population densities and deprivation, and with lower levels of education which might also explain the socio-economic activities and therefore their low levels of car ownership. These populations live near the city centre, with mostly long-term unemployed, contract labour and full-time students, quite possibly immigrants. After manipulating the data in percentages, it would have been beneficial to conduct either a log-transformation or an inverse hyperbolic sine transformation before z-score transformation, in order to normalise the variables and obtain better results. (Sabbata and Liu, 2019) suggest the use of deep neural networks for the classification, wherein the substitute of variable selection and pre-processing – dimensionality reduction, along with geo-convolution is used for analysing the geographic patterns. The results of the study were comparable for the city of Leicestershire with the 2011 Geodemographic Classification by (Gale *et al.,* 2016), albeit there being concerns about the difference in the resolutions of the two studies.



Z-score Transformation based Geodemographic Classification

Conclusion

Considering that the census data was released quite recently, there are no studies conducted, especially for the city of Birmingham, to confirm the results of this study. However, the results show a trend that is common in cities similar to Birmingham, where the rich, affluent majority lives in the suburbs away from the highly dense city centre, while the working or 'struggling' class flock towards the city centre in search of employment, which may result in an increase in deprivation of the areas. Moreover, students live in rented flats near major universities, while the young professionals may do likewise, in this case near the city centre which tends to be an employment hub. This analysis may be compared to the geodemographic classification for the 2011 data for confirmation or change in trends. Better data manipulation, choice of variables, and other techniques like Fuzzy C-means, DBSCAN or the use of Deep Neural Networks might have resulted in different and/or better results. Being a melting pot of cultures, generations and trades, Birmingham like London, results in multiple clusters based on the intended applications of a study.

REFERENCES

Ahmed, M., Seraj, R. and Islam, S.M.S. (2020) 'The k-means algorithm: A comprehensive survey and performance evaluation', *Electronics (Basel),* 9(8), pp. 1295.

Alexiou, A. and Singleton, A. (2015) 'Geodemographic analysis', in Brunsdon, C. and Singleton, A. (eds.) *Geocomputation: A Practical Prime* SAGE, pp. 137-151.

Birmingham City Council. (2022) 'Why birmingham's super-diversity is a strength, and not a surprise', .

Burns, L., See, L., Heppenstall, A. and Birkin, M. (2018) 'Developing an individual-level geodemographic classification', *Applied Spatial Analysis and Policy,* 11(3), pp. 417-437.

Chris Brunsdon and Alex Singleton (2015) 'Geodemographic analysis'*Geocomputation: A Practical Primer*. First Edition edn. 55 City Road: SAGE Publications, Inc, pp. 136.

Gale, C.G., Singleton, A.D., Bates, A.G. and Longley, P.A. (2016) 'Creating the 2011 area classification for output areas (2011 OAC)', *Journal of Spatial Information Science,* 2016(12), pp. 1-27.

Greater Birmingham Chambers of Commerce and University of Birmingham (2018) *Birmingham economic review 2018.* Available at: https://www.birmingham.ac.uk/Documents/college-social-sciences/business/research/city-redi/birmingham-economic-review-2018/ber-2018-ch-0-foreword.pdf (Accessed: 11 May, 2023).

Harris, R., Johnston, R. and Burgess, S. (2007) 'Neighborhoods, ethnicity and school choice: Developing a statistical framework for geodemographic analysis', *Population Research and Policy Review,* 26, pp. 553-579.

Office for National Statistics (2023) *Output areas (dec 2021) boundaries generalised clipped EW (BGC).* Available at: https://geoportal.statistics.gov.uk/datasets/ons::output-areas-dec-2021-boundaries-generalised-clipped-ew-bgc/explore?location=52.748611%2C-2.489483%2C7.84 (Accessed: 10 May, 2023).

Sabbata, S.D. (2022) *R for geographic data science.*

Sabbata, S.D. and Liu, P. (2019) *Deep learning geodemographics with autoencoders and geographic convolution.* . June, 2019.

Singleton, A.D. and Longley, P. (2015) 'The internal structure of greater london: A comparison of national and regional geodemographic models', *Geo : Geography and Environment,* 2(1), pp. 69-87.

Vickers, D. (2008) *Creating a geodemographic classification.* Department of Geography, University of Sheffield: Open and Free Geodemographics. Available at: https://www.mrs.org.uk/pdf/03_11_08_dan_vickers.pdf (Accessed: 6 May, 2023).

Yang, Y., Dolega, L. and Darlington-Pollock, F. (2022) 'Ageing in place classification: Creating a geodemographic classification for the ageing population in england.', *Applied Spatial Analysis and Policy,* .

Zach (2021) *Z-score normalization: Definition & examples.* Available at: https://www.statology.org/z-score-normalization/ (Accessed: 11 May, 2023).