TravelOAC: development of travel geodemographic classifications for England and Wales based on open data

Nick Bearman[*1] and Alex D. Singleton[†1]

[1]Department of Geography and Planning, School of Environmental Sciences, University of Liverpool,
Gordon Stephenson Building, Liverpool, L69 7ZQ, United Kingdom

Summary

This paper develops a custom geodemographic classification for travel in England and Wales. Travelling is an important factor in many life decisions, including home and work life. Variables for transport (distance to nearest airport, rail station, ferry station, tram stop and bus stop, number of cars owned, and mode of travel to work) and demographics (gender, age and social class) for each Output Area in England and Wales are used to create eight clusters of different transport characteristics. The characteristics of the different clusters are discussed, along with future improvements to be implemented in the classification method.

KEYWORDS:

Open geodemographics, travel, public transport, Output Area classification, open data

1. Literature Review

Geodemographics are "the analysis of people by where they live" (Sleight, 1997) and started in the early 1900s with Charles Booth's poverty map of London, a spatial representation of different social classes (Booth, 1902; LSE, 2014). The academic side of geodemographics developed through the 1950s and 1960s, when the commercial application of geodemographics also developed. These commercial developments in the 1970s were primarily led by Richard Webber (1977), involved in the creation of MOASIC and ACORN, two commercial classifications that are still in use today. This (and many other) commercial classifications are a "black box" process - there was little publicly known about the data sets used or how the classification was constructed.

Recent developments have brought geodemographic classification generation back into the academic sector, with the open geodemographics developing a type of classification that is academically rigorous and with the methods used open to inspection. This was partly driven by the availability of open data, particularly data for small areas related to the census. The OAC (output area classification) was developed for the 2001 and 2011 censuses (Vickers and Rees, 2007) and it is the 2011 OAC and 2011 Census data that this work builds upon.

Both the OAC classification and a number of commercial geodemographic classifications were built with a range of applications in mind. However, the validity of a geodemographic classification for a generic application has been questioned, as the factors influencing someone to pick a specific holiday destination are likely to be different to their opinions on private healthcare (Singleton and Longley, 2009). Additionally, it is likely that for a specific application (e.g. travel) that there is additional open data available which could contribute useful information to the geodemographic classification. From an analytical point of view, it has been found that "differences between [geodemographic] classes are generally smaller than the differences found within any particular class" (Voas and Williamson, 2001). These factors, alongside the fact that while generating a specific geographic used to be a complex undertaking, it is now easier as a result of the maturing of spatial data technologies (Adnan et al., 2010; Singleton and Longley, 2009), create a compelling argument to develop application specific classifications. Below, the area of travel and transport is introduced, and this will be followed

[*] n.bearman@liverpool.ac.uk

[†] alex.singleton@liverpool.ac.uk

by the creation of a travel geodemographic classification.

Everyone needs to travel for a variety of reasons (work, school, shopping, etc.) and the factors behind the choice of a specific mode of travel for a specific journey are varied and complex. A travel geodemographic classification will show how transport provision and usage varies across the country, and highlight any relationship it has with other demographic factors such as gender, age and SES (socio-economic status). It could also be used to target transport improvements that are particularly important to increase uptake of public transport and reduce reliance on private car use. Understanding the geodemographics of travel accessibility and travel use will help development of transport options and contribute to the task of reducing transport CO_2 emissions. Transport accounts for about 28% of the UK's total CO_2 emissions (Hickman and Banister, 2007) so understanding more about this issue can help reduce carbon emissions.

2. Methods

The theoretical framework adopted for this work assists with selecting variables by considering the domains that this work is interested in, the concepts within each domain and then the variable within each concept. Table 1 shows how this is applied to the travel geodemographic, and also shows which census variable was used. Gender, age and social class were included in the clustering process to represent a proxy for income (social class) and different transport needs (e.g. working vs. non-working).

Table 1. Variables included in the classification

Domains	Concepts	Variable	Census table used
Demography	Gender	Gender	KS101 Usual resident population
	Age	Age groups	KS102EW Age structure
	Social class	National Statistics socio-economic class	KS611EW NS-SeC
Transport	Travel to work	Mode of usual travel to work	QS701EW Method of travel to work
	Ease of access to car	Car ownership	KS404EW Car or van availability
	Ease of access to public transport	Distance to closest bus/train/ferry/airport stop	*NA (distance calculated from NaPTAN data)*

The transport stop data was generated using stop locations for England and Wales from NaPTAN[‡], with the distance to each type of nearest stop calculated for each Output Area population centroid. OA population centroids were used because these more appropriately reflected the location of the population within an OA, and the use of point data allowed a much simpler spatial calculation to take place (measuring distance points to points) than calculating distances from points to polygons (Output Areas).

Walking routes were modelled in Routino for distances from each OA centroid to the nearest rail station, tram stop and bus stop using previously discussed methods (Bearman and Singleton, 2014). Straight-line distances were used for airport and ferry ports because in the vast majority of cases a walking route to an airport or ferry port would not make sense. The transport types chosen were based on the existing categories in NaPTAN, and will have different weightings in the classification according to the number of different stops of that type to reflect their different levels of importance.

A k-means classification was performed using the data described above. Both an initial k-means cluster analysis and a clustergram analysis (Galili, 2010; Schonlau, 2002, 2004) were run to establish the number of clusters, and based on these results a classification was run for eight clusters.

[‡] NaPTAN, National Public Transport Access Nodes, http://data.gov.uk/dataset/naptan

3. Results & Clusters

The output from the cluster analysis shows interesting characteristics, with each of the eight clusters showing distinct characteristics. A number of parameters within each category showed similar relationships in the initial analysis, so were collapsed together in the final analysis. Table 2 and Figure 1 shows the clusters derived from this analysis, along with their variables and spatial distribution. A number of the clusters (e.g. 3 and 5) have strong patterns within the variables, whereas others have a weaker relationship.

Table 2. Clusters and data patterns for the transport geodemographic classification

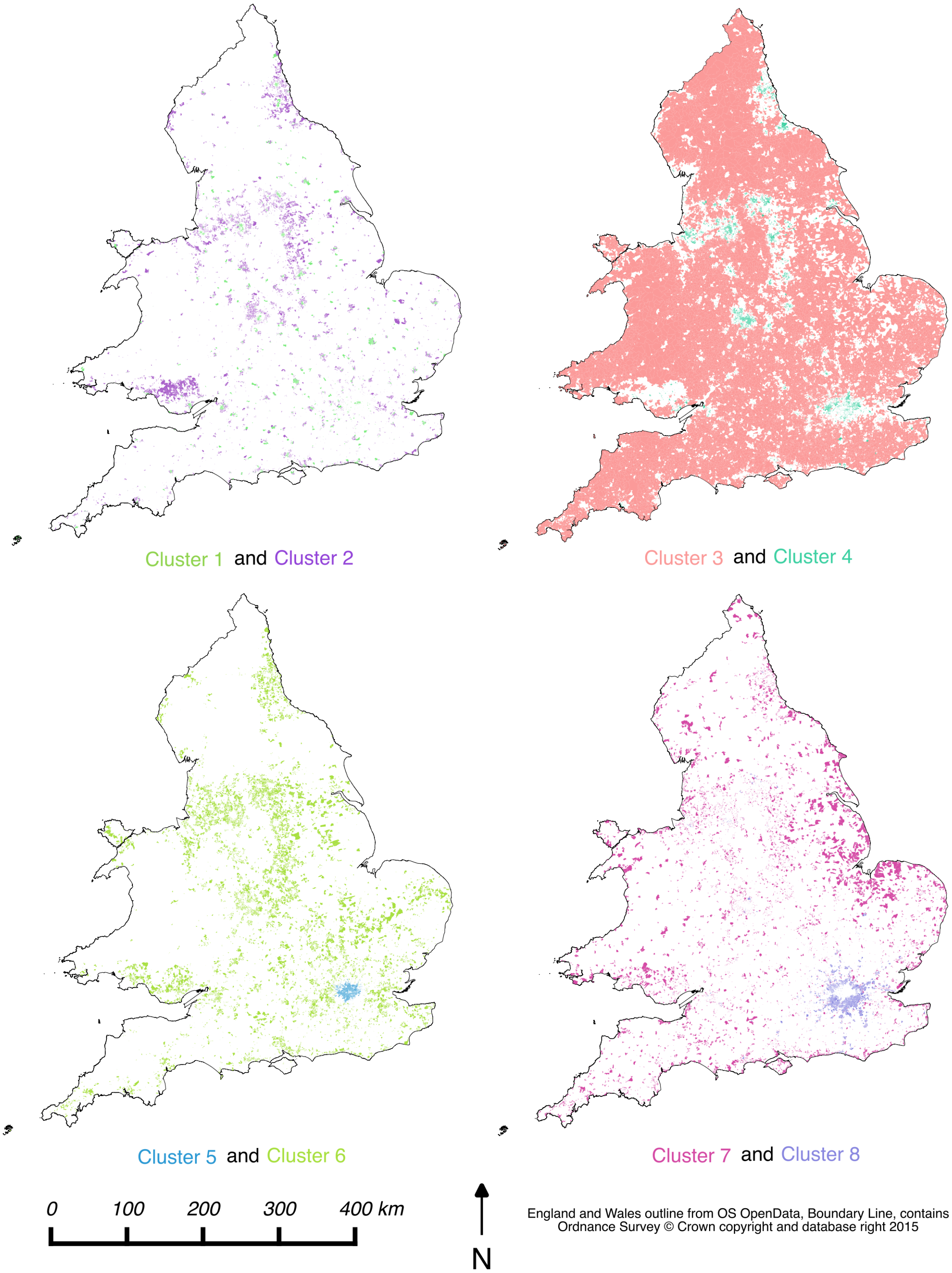
Cluster		1	2	3	4	5	6	7	8
N (total N = 181,408)		8,755	38,634	24,076	21,161	9,480	42,445	23,541	13,316
Distance to public transport		Mid	Far	Far	Close	Close	Mid	Far	Far
Car ownership		0 car	1 car	2+ cars	0 car	0 car	2+ cars	1 car	1 car
Gender		Male	Mix	Mix	Mix	Slightly male	Mix	Female	Slightly female
Age		15-44	Under 45	45+	Under 45	15-44	Mix	45+	Slightly <45
National Statistics Socio-economic classification#	1	▨		▨		▨	▨		▨
	2	▨		▨		▨	▨		▨
	3						▨	▨	▨
	4			▨				▨	▨
	5		▨					▨	
	6		▨		▨				
	7		▨		▨				
	8		▨		▨	▨			
Usual mode of travel to work		Cycle / Walk	Passenger	Home work / drive	Bus & taxi	Tram /bus	Drive	Slight car	Train

# 1. Higher managerial, administrative and professional occupations 2. Lower managerial, administrative and professional occupations, 3. Intermediate occupations, 4. Small employers and own account workers 5. Lower supervisory and technical occupations, 6. Semi-routine occupations, 7. Routine occupations, 8. Never worked and long-term unemployed

Table 2 summarises the analysis of the results, showing the different characteristics of the different clusters from this analysis. The full table (including variable index scores) is available in the appendix. Clusters 1 and 5 are professional workers, more likely to be male and in management. Cluster 5 is predominantly located in London, with a higher use of public transport, whereas cluster 1 is located in other urban areas, with a higher rate of cycling and walking. Clusters 2 and 3 are located in rural areas, with cluster 3 in higher management with a higher income and more likely to have 2 or more cars, and cluster 2 with lower incomes, with 1 car, in lower supervisory, technical and routine occupations. Cluster 4 is located in urban areas, with an emphasis on the more deprived areas in the north of England and London, with a combination of low wages and a reliance on bus and taxis. Cluster 6 is primarily distributed in suburban and rural areas, with a mixed demographic, higher income and multiple cars. People within clusters 7 and 8 are more likely to be female and to have a lower income. Cluster 8 has a London commuter focus and much exhibits higher train usage, where as cluster 7 covers rural areas in the rest of England with a reliance on a car.

4. Discussion

These results show the potential benefits of generating a travel geodemographic classification. There are both travel patterns associated with age and income (SES), as well as additional ones associated with access to and use of different modes of transport. In addition, there are particular geographic patterns, both for rural and urban locations, which also reflect income and age distributions. Gender is a strong factor in two clusters (cluster 1 and 7, and to a lesser extent cluster 5) and is likely to reflect income and social make up differences. Socio-economic classification also features strongly in the classification, and we use this data as a proxy for income levels.



Cluster 1 and Cluster 2

Cluster 3 and Cluster 4

Cluster 5 and Cluster 6

Cluster 7 and Cluster 8

England and Wales outline from OS OpenData, Boundary Line, contains Ordnance Survey © Crown copyright and database right 2015

Figure 1. Spatial variation of a selection of the clusters from the travel geodemographic classification.

This data indicates particular clusters which might benefit from targeting to reduce CO_2 emissions, either by promoting more public transport use, or acting on reasons for low levels of public transport usage. In addition it highlights areas that have low public transport provision as well as low public transport usage levels, which could assist in targeting new public transport provision.

There are a number of refinements to be made to the classification, both in the data included and the processing of the classification. Currently, distances to most forms of public transport are calculated as walking distances. However this will vary between individuals, and is also not appropriate for some types of multimodal transport (for example, when someone drives to a train station to catch a train). Additionally, this work does not take into account the frequency of public transport or the routes followed, which are important factors when considering transport accessibility. In addition, data from the National Travel Survey could provide more contextual information for the analysis of the cluster behaviour patterns.

5. Acknowledgements

The authors of this paper would like to acknowledge support from the ESRC for this research, as well as acknowledging the open data sets used in this work, including 2011 Output Area Classification, OS Open Data Boundary Line, 2011 Census data and National Public Transport Access Nodes. The authors would also like to thank the reviewers for their useful comments.

6. Biography

Nick Bearman is a Research Associate and University Teacher at the University of Liverpool, previously working at the University of Exeter Medical School. He is interested in the use of GIS to solve novel problems, particularly in the areas of secondary data reuse, health and big data.

Alex Singleton is a Professor in Geographic Information Science; his research concerns how the social and spatial complexities of individual behaviour can be represented and understood within a framework of quantitative social science and computer modelling, extending from a geographic tradition of area classification.

References

Adnan, M., Longley, P.A., Singleton, A.D., and Brunsdon, C. (2010). Towards Real-Time Geodemographics: Clustering Algorithm Performance for Large Multidimensional Spatial Databases. Trans. GIS *14*, 283–297.

Bearman, N., and Singleton, A.D. (2014). Modelling the school commute for 7.5m students over 4 years using data from the school census. In GISRUK2014, (University of Glasgow, UK),.

Booth, C. (1902). Life and Labour of the People in London (London: Macmillan).

Galili, A.T. (2010). Clustergram: visualization and diagnostics for cluster analysis (R code).

Hickman, R., and Banister, D. (2007). Looking over the horizon: Transport and reduced CO2 emissions in the UK by 2030. Transp. Policy *14*, 377–387.

LSE (2014). Poverty maps of London (Charles Booth Online Archive).

Schonlau, M. (2002). The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses. Stata J. *2*, 391–402.

Schonlau, M. (2004). Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. Comput. Stat. *19*, 95–111.

Singleton, A.D., and Longley, P.A. (2009). Geodemographics, visualisation, and social networks in applied geography. Appl. Geogr. *29*, 289–298.

Sleight, P. (1997). Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business (Henley-on-Thames, Oxfordshire England: NTC Publications).

Vickers, D., and Rees, P. (2007). Creating the UK National Statistics 2001 output area classification. J. R. Stat. Soc. - A *170*, 379–403.

Voas, D., and Williamson, P. (2001). The diversity of diversity: a critique of geodemographic classification. Area *33*, 63–76.

Webber, R.J. (1977). An Introduction to the national classification of wards and parishes (London: Centre for Environmental Studies).

Appendix: Index Scores of Cluster Variables

Table 1. Clusters and variable index scores (0 = typical) for the transport geodemographic classification

Cluster		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
N (total N = 181,408)		8,755	38,634	24,076	21,161	9,480	42,445	23,541	13,316
Distance#	Rail	-0.35	-0.11	**0.58**	-0.20	-0.17	-0.05	0.26	-0.37
	Tram	0.04	0.24	0.08	-0.29	-0.74	0.04	0.28	-0.50
	Bus	-0.36	-0.10	0.38	-0.28	-0.20	-0.14	**0.53**	-0.05
	Airport	-0.32	0.21	0.44	-0.47	-0.63	-0.06	0.46	-0.61
	Ferry	0.05	0.17	0.07	0.10	-1.02	0.25	-0.09	-0.72
Car / van ownership	None	**0.75**	0.30	-1.00	**1.41**	**1.44**	-0.74	-0.14	-0.20
	1	0.07	0.44	-0.96	-0.57	-0.46	0.07	0.47	**0.61**
	2 or more	-0.78	-0.51	**1.46**	-1.13	-1.22	**0.71**	-0.08	-0.10
Age	0 - 4	-0.48	0.34	-0.62	**0.99**	0.18	-0.01	-0.88	0.34
	5 - 14	-1.31	0.22	0.06	**0.80**	-0.61	0.20	-0.82	0.07
	15 - 44	**2.14**	0.15	-0.87	0.42	**1.42**	-0.09	-1.06	0.24
	45 - 64	-1.45	-0.22	**1.08**	-0.77	-1.01	0.34	0.39	-0.19
	65+	-0.90	-0.21	0.46	-0.58	-0.81	-0.20	**1.59**	-0.28
National Statistics Socio-economic classification	SES1&2 Higher / middle management	0.17	-0.81	**0.89**	-1.18	**1.05**	0.35	-0.07	**0.77**
	SES3 Intermediate management	-0.83	-0.26	-0.12	-0.80	-0.77	**0.74**	0.28	0.49
	SES4 Self-employed	-0.86	-0.44	**1.39**	-0.62	-0.14	-0.08	0.31	0.12
	SES5 Lower supervisory	-0.76	**0.76**	-0.72	-0.06	-1.01	0.14	0.40	-0.71
	SES6&7 Semi-routine / routine	-0.82	**1.05**	-0.91	**0.83**	-1.03	-0.30	0.21	-0.84
	SES8 Never worked	-0.22	0.21	-0.65	**1.98**	0.31	-0.55	-0.40	-0.20
Usual mode of travel to work	Car (Driver)	-1.00	-0.01	**0.60**	-1.16	-1.92	**0.97**	0.11	-0.48
	Car (Passenger)	-0.46	**0.87**	-0.44	-0.15	-1.37	0.24	-0.11	-0.79
	Tram	-0.22	-0.31	-0.27	0.07	**3.44**	-0.29	-0.31	0.44
	Train	0.04	-0.38	-0.09	-0.19	**0.62**	-0.19	-0.37	**2.34**
	Bus	0.37	0.09	-0.82	**1.23**	**1.20**	-0.31	-0.51	0.07
	Taxi	-0.03	0.32	-0.41	0.45	0.11	-0.20	-0.21	0.02
	Motorcycle	-0.33	0.09	-0.22	-0.35	**0.78**	0.08	-0.12	0.31
	Cycle	**1.14**	0.02	-0.40	-0.23	**1.21**	-0.09	-0.13	-0.08
	Walk	**2.11**	0.33	-0.50	0.01	0.03	-0.36	-0.05	-0.22
	Other	0.15	-0.10	0.26	-0.06	0.29	-0.12	-0.01	0.01

Bold = more than or equal to **+0.5**, underline = less than or equal to -0.5.

# For distance, positive values are higher distances than average, and negative values are closer than average.