Open Geodemographics: Open Tools and the 2011 OAC

Gale, C.G., Adnan, M., Longley, P.A.

University College London, Department of Geography, Gower Street, London, WC1E 6BT.
Tel: +44 (0)20 7679 0510 Fax: +44 (0)20 7679 0565
Email c.gale.10@ucl.ac.uk, m.adnan@ucl.ac.uk, p.longley@ucl.ac.uk

ABSTRACT

Historically geodemographic classifications have been created as closed systems – or 'black box' environments. This results in a user being presented with a classification with little or no understanding of the decisions made and the processes involved in its creation. The 'GeodemCreator' is a free general purpose geodemographic decision support tool, making the processes of building a classification transparent to the user. The 2011 OAC will be a free, open source classification with a fully published methodology. This will allow techniques to be easily transferred and applied using 'GeodemCreator' to make the 2011 OAC adaptable and possibly updateable in the future.

KEYWORDS: Geodemographics, GIS, Open Tools, Geoweb 2.0, OAC, Open Data

1. Introduction

Commercial geodemographic classifications are created as 'black box' systems (Longley and Singleton 2009). Experts use closed methods and provide little documentation of the data inputs, weighting and normalisation procedures or the specific methods of clustering. The 2001 Output Area Classification (2001 OAC: Vickers and Rees 2007), by contrast, is an open geodemographic classification built using 2001 Census data, and that has been widely used in applications. Notwithstanding arguments that many neighbourhoods 'filter out' successive residents with similar characteristics to their predecessors, the 2001 OAC is clearly increasingly marginal to measuring the geodemographic patterning of neighbourhoods today. Moreover, there is a need for more open geodemographic classifications which reflect the changing dynamics of population characteristics. An important issue that contributes to the lack of open geodemographic classifications is the unavailability of free software tools which remove the technical complications of creating them. In this paper we present our work of creating a more responsive and open geodemographic classification by using the 'GeodemCreator' software tool. A case study is presented by using 'GeodemCreator' to build an open 'Socio-economic and ethnic' classification of Greater London. The paper also describes preliminary work towards the creation of 2011 Output Area Classification (2011 OAC) which will use 2011 Census data when they become available.

2. Need of Open, Transparent and Flexible Geodemographic Classifications

The use of geodemographic classifications has become popular in different areas with applications in health (Farr and Evans 2005; Shelton *et. al.* 2006), policing (Ashby and Longley 2005), education (Singleton 2010) and local government (Longley and Singleton 2009). Census data have remained the core data source for creating geodemographics segmentations. The current expansion of 'open data' initiatives has resulted in an ever increasing amount of data sources becoming available to the public. This has allowed, in addition to general purpose classifications, bespoke local and national area

classifications to be created by public and private organisations in addition to academic researchers. The Office for National Statistics (ONS) NeSS Data Exchange (Office for National Statistics 2009) is an important open data source, where users can get feeds of Census data through the API. The London Data Store (London Data Store 2010) has been created by the Greater London Authority as an initiative to make London's data free and accessible to all. Programmers and data analysts can now use thousands of data sources in addition to Census data to create their own local area classifications. Crime Data have also been made public by police forces in the UK as part of the open data initiative (http://www.police.uk). This enables general users, data analysts and programmers to map latest crime data either by downloading the data or getting live feeds from the http://www.police.uk website.

By contrast, there has been considerable critique arising from the 'black box' nature of commercial geodemographic classifications. Users of commercial classifications only receive final classifications after areas have been grouped into different classes, and as such have to accept what they are given. Longley *et. al.* (2009) critique the 'black box' nature of geodemographic classifications and form the view that there is a need for more open methods. These open methods are expected to be transparent in explaining all the procedures employed to build a geodemographic classification. Thus there is a need of a clear documentation about the methods of selecting variables and their weightings, the normalization techniques employed, and the clustering algorithms used. Open methods ensure that users have more confidence in the geodemographic classifications they are using.

There are a number of statistical packages (R, SPSS, and Microsoft Excel) available which can be used for building geodemographic classifications. However, there has hitherto been no unified software utility that can be used for building geodemographic classifications which are open, transparent, and flexible. Created by Adnan (2011), 'GeodemCreator' is one such general purpose geodemographic decision support tool. 'GeodemCreator' is a free software utility with no license fees. The software can be used for building national or region-specific bespoke geodemographic classifications. In the current version, 'GeodemCreator' allow users to build geodemographic classifications at any bespoke spatial levels.

The next section explains the use of 'GeodemCreator' in building a new geodemographic classification of Greater London. It is proposed that this tool can be used in conjunction with the creation of the 2011 OAC. Some of the background to the 2011 OAC is set out in Section 4 of this paper.

3. Using 'GeodemCreator' to build an open geodemographic classification

This section shows the results of a case study by using 'GeodemCreator' to build a geodemographic classification. 'GeodemCreator' is a cross-platform tool that requires only Java (http://java.com) and R (www.r-project.org) installed on the machine, and can be used by both experienced and inexperienced users for building their local area geodemographic classifications. Figure 1 shows a screen shot of the software.

Figure 1: A screen shot of 'GeodemCreator'

In response to the observation that the 2001 OAC (Vickers and Rees 2007) ascribes too many neighbourhoods to the blanket 'multicultural' category, 'GeodemCreator' has been used to create a software environment for socio-economic and ethnic classification of Greater London. The 41 2001 OAC variables (Table 1) are supplemented with 12 other ethnicity variables (Table 2), created using the UCL Worldnames database (http://worldnames.publicprofiler.org).

Variables	Domains
V1: Age 0-4 V2: Age 5-14 V3: Age 25-44 V4: Age 45-64 V5: Age 65+ V6: Indian, Pakistani or Bangladeshi V7: Black African, Black Caribbean or Other Black V8: Born Outside the UK V9: Population Density	**Demographic**
V10: Divorced V11: Single person household (not pensioner) V12: Single pensioner household V13: Lone Parent household V14: Two adults no children V15: Households with non-dependent children	**Household Composition**
V16: Rent (Public) V17: Rent (Private) V18: Terraced Housing V19: Detached Housing V20: All Flats	**Housing**

V21: No central heating V22: Rooms per household V23: People per room	
V24: HE Qualification V25: Routine/Semi-Routine Occupation V26: 2+ Car household V27: Public Transport to work V28: Work from home V29: Limiting Long Term Illness (SIR) V30: Provide unpaid care V31: Students (full-time) V32: Unemployed V33: Working part-time V34: Economically inactive looking after family V35: Agriculture/Finishing employment V36: Mining/Quarrting/Construction employment V37: Manufacturing employment V38: Hotel & Catering employment V39: Health and Social work employment V40: Financial intermediation employment V41: Wholesale/retail employment	**Socio-Economic**

Table 1: 41 2001 Census variables used for building the 2001 Output Area Classification

V42: ‘European’ ethnic group
V43: ‘East Asian & Pacific’ ethnic group
V44: ‘Muslim’ ethnic group
V45: ‘Greek’ ethnic group
V46: ‘English’ ethnic group
V47: ‘Nordic’ ethnic group
V48: ‘African’ ethnic group
V49: ‘Japanese’ ethnic group
V50: ‘Hispanic’ ethnic group
V51: ‘Celtic’ ethnic group
V52: ‘Jewish’ ethnic group
V53: ‘South Asian’ ethnic group

Table 2: Ethnicity variables for creating the geodemographic classification

‘GeodemCreator’ produces the final classification and their corresponding radial charts. The software uses the standard implementation of *k*-means clustering algorithm to cluster the data in homogeneous groups. The radial charts are helpful in naming and identifying characteristics of individual clusters. The final classification produced is shown in Figure 2.

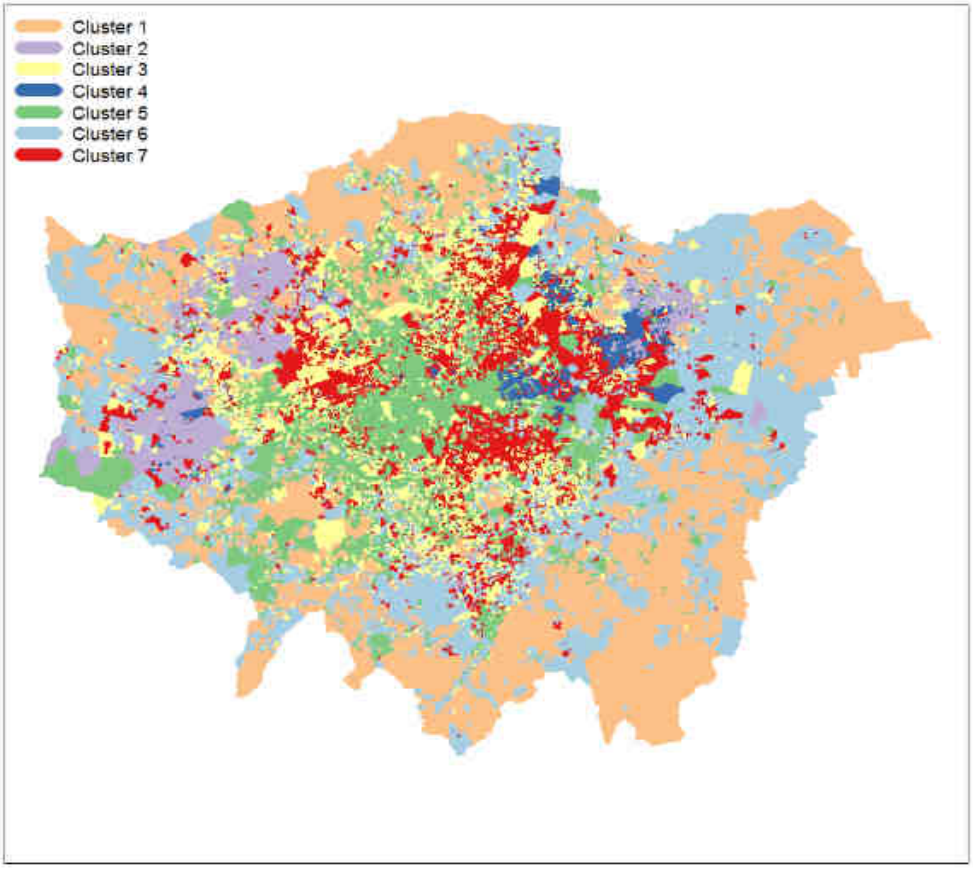
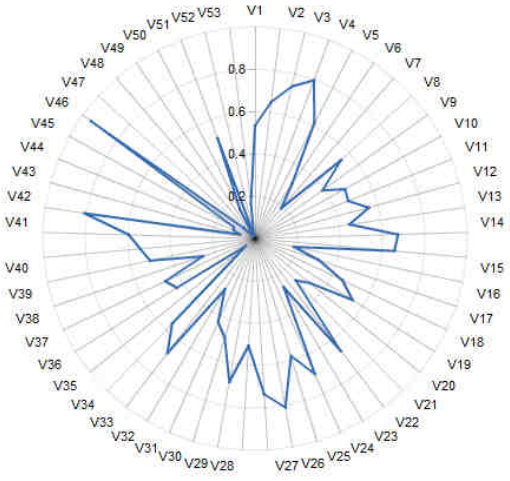


Figure 2: A socio-economic and ethnicity classification of Greater London

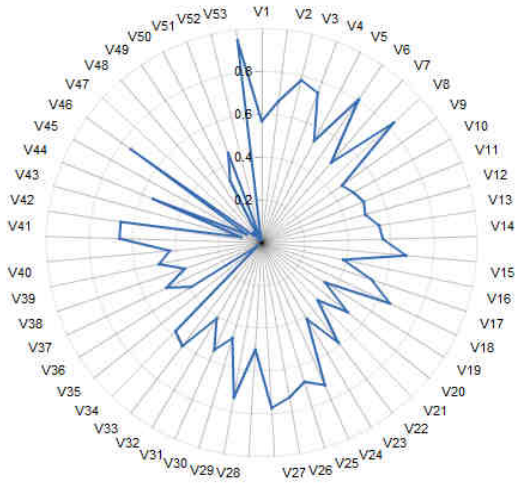
The following Figures (3 to 9) show the radial charts of individual clusters. Based on the values of the selected variables, each cluster was given a unique name.

Figure 3



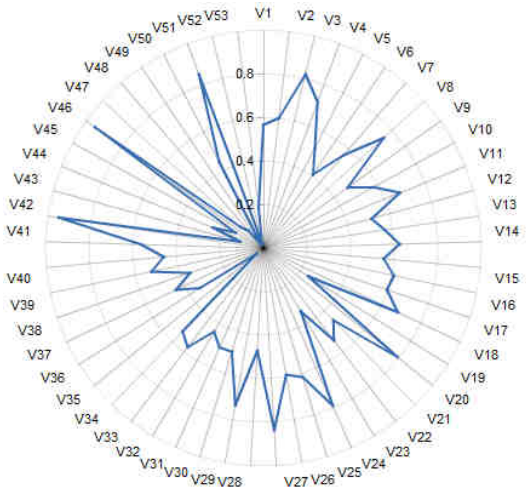
Cluster 1: English and European ethnic groups living in suburban areas

Figure 4



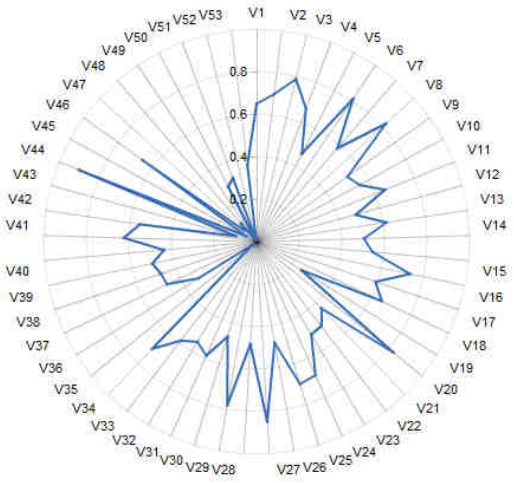
Cluster 2: Well off and educated Asian Families

Figure 5



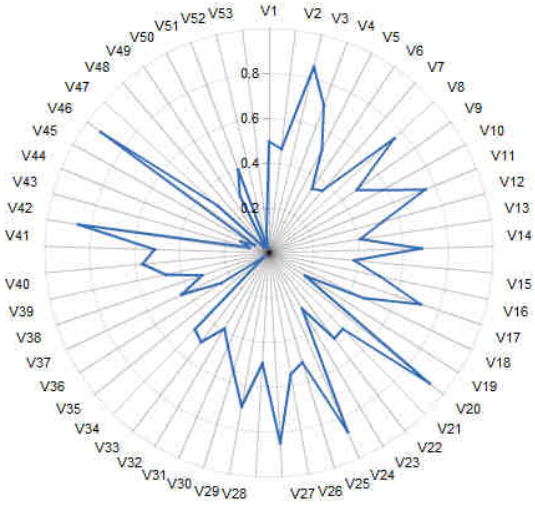
Cluster 3: English, European, and Celtic fringe city commuters

Figure 6



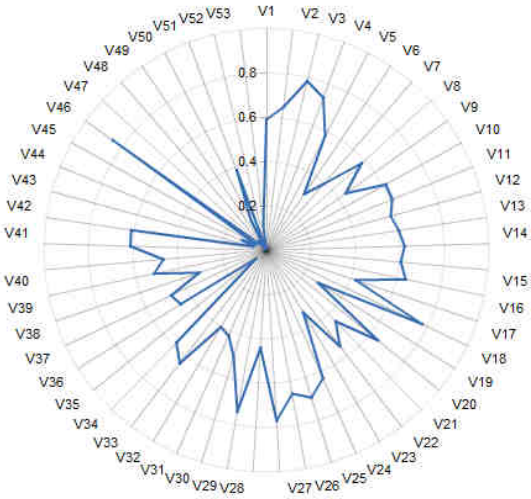
Cluster 4: Poor Asian Families

Figure 7



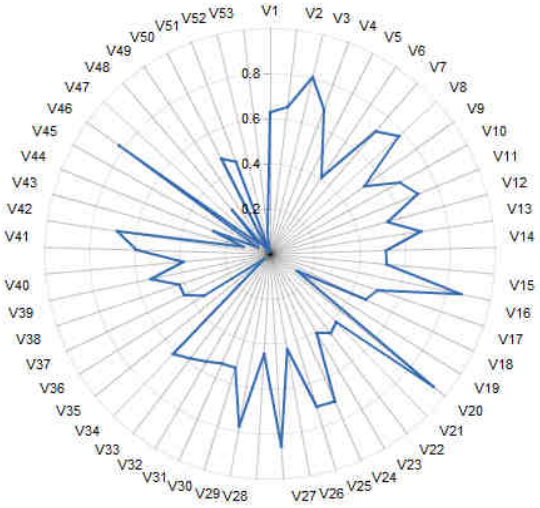
Cluster 5: Childless European city dwellers

Figure 8



Cluster 6: Native Blue Collar communities

Figure 9



Cluster 7: English and Europeans living in council properties

This section has explained the use of 'GeodemCreator' for building an open geodemographic classification. Because 'GeodemCreator' is a free software utility, the process of building a geodemographic classification is transparent to the user. This will result in improvements in research practice, through clear documentation of the procedures employed in producing classifications.

4. The 2011 Output Area Classification

The 2011 OAC will be a free, open source geodemographic classification of the United Kingdom. Like its predecessor, the 2001 OAC (Vickers and Rees 2007), it will be considered a "general purpose" classification that is not designed with any specific application in mind. Through the development of the 2011 OAC it is foreseeable that specialised variants of the classification will be made, but the responsibility to do so will increasingly shift towards the user community. As CACI have shown, in addition to their standard ACORN product, variants – such as HealthACORN and GreenACORN – can co-exist (CACI 2011), demonstrating a desire to create both general purpose and specialised geodemographic classifications.

It is proposed that the 2011 OAC methodology will be published in full, including all code used. This will result in the 2011 OAC being a completely open source geodemographic classification; with techniques that can be easily transferred and applied by any user to create their own bespoke classification. Decisions that in the past were based upon 'black box' solutions (Longley and Singleton 2009), where specific details of the methodologies used remained unpublished, create the greatest problem when trying to create an open and accountable classification. This is especially true of variable selection as this is particularly hard to describe. Vickers (2006) describes the creation of a geodemographic classification as "as much of an art form as a science". This is in reference to the long-held assumption that the analyst needs to have experience in the field of geodemographics to

make decisions about how best to create a classification – a process in which variable selection is of fundamental importance. In the past, even if users have been aware of which variables have been included, they do not fully understand why those particular variables were used. To an extent this is true even of the 2001 OAC, considered an open classification. Due to the subjective nature of variable selection the final variables selected will vary depending on the analyst creating the classification. The 2011 OAC will attempt to resolve this issue by using a methodology that will rely on an algorithm to sort through all variables and decide for *n* cluster groups which variables offer the greatest discriminatory power. This is only something that has recently become a possible, in the wake of continuing improvements in computational processing.

For the advanced user, this should be sufficient for the creation of a bespoke classification using the 2011 OAC methodology. Other users will require assistance in creating their own bespoke classifications. The 'GeodemCreator' is therefore a potentially useful tool for both this particular element of the geodemographics community as well as those who have previously created their own bespoke geodemographic classifications, such as Hull City Council (AGI 2011). 'GeodemCreator' will also extend the shelf life of the 2011 OAC, through its provision of a number of useful functions. Firstly users will be able to modify aspects of the 2011 OAC methodology to their choosing – such as the number of variables used and/or the extent of coverage required across the United Kingdom. Secondly the 2011 OAC variables can be used in conjunction with other data sources to introduce new variables, as is the case with the socio-economic and ethnicity classification of Greater London. Such modifications and adaptions will allow the possibility of updating the 2011 OAC over time, albeit rather crudely, and thus maintain its relevance as the Census data elements age.

5. Conclusion and Prospects

This paper describes work towards building more responsive and open geodemographic classifications. For a range of essentially legacy reasons, only expert users presently produce classifications. However, 'GeodemCreator' is a software tool that removes the technical complications of creating a geodemographic classification, and thus helps users in building open geodemographic classifications. However, in the current form, 'GeodemCreator' is a free software utility, but the future work will involve the launch of an open source version of the software. This paper also describes the preliminary work towards building the 2011 OAC, which will be an open source and free geodemographic classification. In the future, 'GeodemCreator' will make the 2011 OAC adaptable to developments in geodemographics, which in turn will require a methodological approach adaptable to the changing needs of users.

6. References

Adnan, M. (2011). Towards real time geodemographic information systems: design, analysis and evaluation. PhD Thesis. University College London, London. Unpublished.

AGI (2011) "Developing a 'free' customer classification system – The Alternative Approach!" Retrieved 4th September 2011 from http://www.agi.org.uk/lps/2011/5/6/developing-a-free-customer-classification-system-the-alterna.html

Ashby, D. I., Longley, P. A. (2005). "Geocomputation, Geodemographics and Resource Allocation for Local Policing". *Transactions in GIS*, 9 (1): 53-72.

CACI (2011) “Location Planning Data” Retrieved 20th May 2011, from http://www.caci.co.uk/LocationPlanningData.aspx

Farr, M., Evans, A. (2005). "Identifying `Unknown Diabetics' using Geodemographics and Social Marketing". *Interactive Marketing* 7: 47-58.

Harris, R., Sleight, P., Webber, R. (2005). Geodemographics, GIS and Neighbourhood Targeting. Wiley, London.

London Data Store (2010). "London Data Store" Retrieved 15th May, 2011 from http://data.london.gov.uk.

Longley, P. A., Singleton, A.D. (2009). "Classification through consultation: public views of the geography of the E-Society". *International Journal of Geographic Information Science*, 23 (6), 737-763.

Office for National Statistics (2009). "Ness Data Exchange." Retrieved 3rd June, 2009, from http://www.neighbourhood.statistics.gov.uk/HTMLDocs/downloads/NeSS-Data-Exchange-Technical-Implementation-Guide-v1.0.pdf.

Shelton, N., Birkin, M., Dorling, D. (2006). "Where not to live: a geo-demographic classification of mortality for England and Wales, 1981-2000". *Health and Place*, 12 (4).

Singleton, A.D., Longley, P.A (2008). Creating open source geodemographic classifications for Higher Education applications.

Singleton, A.D. (2010). "The Geodemographics of Educational Progression and their Implications for Widening Participation in Higher Education". *Environment and Planning A*. In Press.

Singleton, A. D., Wilson, A.G., O'Brien, O. (2010). "Geodemographics and Spatial Interaction: an integrated model for higher education". *Journal of Geographical Systems*, 1435-5930, 1-19.

Vickers, D. (2006) Multi-level integrated classifications based on the 2001 Census. PhD Thesis. University of Leeds, Leeds. Unpublished.

Vickers, D. and Rees, P.H. (2007). Creating the National Statistics 2001 Output Area Classification. Journal of the Royal Statistical Society, Series A. 170(2), 379-403.

7. Biography

Chris Gale *is a second year PhD student at University College London, funded by studentships from UCL and the Office for National Statistics. He is working towards creating better area classifications for the 2011 Census, specifically a 2011 version of the Output Area Classification, with particular focus on new modes of dissemination that better utilise web technologies and new advances in GIS and geodemographics.*

Dr Muhammad Adnan *is working as a Research Associate at University College London. His research focus is on Open Geodemographics, Data Mining, algorithm optimisation, and Visualisation of large spatio-temporal databases.*

Professor Paul Longley *is Professor of Geographic Information Science at the Department of Geography, University College London.*