Open Data Sources for Domain Specific Geodemographics

Matthew D. Pratt[1], Paul A. Longley[1], James Cheshire[2], Chris Gale[1]

[1]Department of Geography, University College London, Gower Street, London, WC1E 6BT

matthew.pratt.12@ucl.ac.uk; p.longley@ucl.ac.uk; c.gale.10@ucl.ac.uk

[2] Centre for Advanced Spatial Analysis, University College London, 90 Tottenham Court Road, London W1T 4TJ

james.cheshire@ucl.ac.uk

KEYWORDS: Open Data, Geodemographic Classifications, Scale and Aggregation

1. Introduction

Geodemographics has been defined as *'the analysis of people according to the places where they live'* (Sleight 2004). Although most widely used in commercial applications, geodemographic classifications are relevant across many sectors, including health, crime and policing, and education, with classifications devised for these domains relevant to applications in retailing. The wider availability of 'open data' presents an expanding range of data sources for geographical analysis, potentially adding value to geodemographic classifications. However, there remains a need for a more strategic approach to understanding the potential of open data, as recognised in proposals for a National Information Framework (NIF: APPSI 2012). The spatial dimension of a UK NIF has the potential to spawn many applications and leverage resource, although it seems probable that the onus will be upon developers in order to bring this about. While few open datasets are likely to be made available at levels of granularity comparable to the Census of Population, Harris and Johnston (2008) have suggested that coarser granularities may nevertheless provide acceptable frameworks for analysis of collectively consumed goods and services in the public sector domain. Additional issues arise from the absence of linkage between open data resources, as well as limited information regarding their provenance and quality. This paper evaluates the use of open data within different sectors, specifically retailing, health, crime and policing, with respect to their provenance, refresh rate and granularity.

Our motivation therefore is to identify the most promising elements of the emerging NIF for the construction of bespoke classifications within these sectors. One way of addressing issues of spatial heterogeneity within coarse-grained open data series is to use the UK Output Area Classification (OAC) to examine local patterns of variation in socioeconomic and demographic conditions. This discussion leads us to consideration of issues of specification, estimation and evaluation of geodemographic classifications in different applications domains. Finally, and more broadly still, we consider issues of availability (or otherwise) of different data sources, notably national address registers, to support the development of a NIF. The PSI glossary from the National Archives defines open data as *'data which can be used, re-used and re-distributed freely by anyone – subject only at most to the requirement to attribute and share-alike. There may be some charge, usually no more than the cost of reproduction'.* At the time of writing, it remains unclear what the infrastructure underpinning a NIF based upon open data will be.

2. An Overview of Geodemographic Classifications

Geodemographics are based on the premise that internal homogeneity prevails amongst the residents of a neighbourhood, i.e. people with similar characteristics and lifestyles incline to reside in similar places (Goss 1995). Two of the most prominent commercial geodemographic classifications are Mosaic (Experian, Nottingham) and ACORN (CACI, London). The Mosaic family of classifications is perhaps the most widely use across the world. Both classifications are based upon demographics, property values and attributes, socio-economic and consumption characteristics, location attributes and financial measures (Experian 2010). One of the most important issues with the deployment of these classifications is that the content, provenance and weightings assigned to most of the data remain commercially sensitive secrets (Longley and Goodchild 2008). This essentially 'black box' characteristic of geodemographics means that it is not possible for independent analysts to evaluate and question the ways in which clusters emerge, or to assess the quality of many of the data sources that are used. While both classifications offer more tailored versions they still in part can be seen to be 'general-purpose'. In academia, this means that detailed and informed critique of the geodemographic classification is not possible, relegating evaluation to general assessments of how they are used and the broad types of methods that are employed. Furthermore, the awareness of classifications in academia is fragmentary, with access to the results of classifications limited; and the manner in which categories are represented is often more appropriate to marketers, since after all, they are the people for whom the classifications were built (Webber 2007).

Singleton and Longley (2009) anticipate the supplementation and partial replacement of general purpose geodemographic classifications with bespoke purpose built systems for different applications domains. Geodemographic classifications that are open, both with regard to the data sources that they use and the methods that they apply, can engage a wide stakeholder base and stimulate a problem centred approach to market segmentation amongst analysts. The classic exemplar of this approach is the OAC classification, first developed in 2006, that provided a UK wide social classification that is both freely available and fully documented. It produces area classifications for key characteristics of people living in each UK Output Area, forming hierarchal classes (Vickers, Rees et al. 2005). Its transparency enables users to assess the accuracy and techniques used and to potentially combine their own geodemographics to customise OAC.

3. Geodemographic Data Sources

Conducted once every ten years, the UK Census is the most thorough survey in the UK and offers a wealth of household-level geo-demographic data. In highly dynamic regions however the UK Census can rapidly become out-of-date. Open source data are consequently attractive, with government departments working to make more available to the public. While these datasets have much higher repeatability, often several times a year, these published demographics are often at much coarser resolutions. Most commonly in the retail sector 'Big Data' such as those obtained from loyalty card use and card transactions, present a third avenue for demographic information. An issue here however is the bias of loyalty card information to certain demographic groups such as large households and those with children (Wright and Sparks 1999). Geodemographic classifications have hitherto often been heavily based on census and open data. Census data have poor temporal resolution, and although completion of the survey is compulsory, some systematic aspects of under reporting (e.g. of student populations and recent migrants) are likely to be more of an issue than with administrative datasets. Furthermore even the future of the census is in doubt; although more recently MPs have shown their concerns over the loss of this vast and detailed population data source (House

of Commons Science and Technology Committee 2012). Many open data sources exist with high temporal resolution which could add great value to geodemographic classifications, thereby refreshing census based data (Harris and Longley 2002). One example of this is the registers of GP practices updated in real-time. There has been some movement in this area with the development by CACI of 'HealthACORN'. The 'soon to be released' data from the Department for Work and Pensions (DWP see - http://statistics.dwp.gov.uk/asd/index.php?page=tabtool) also presents another interesting and potentially useful form of open data.

4. Methodology

The paper reports on an audit of the most promising open data for creating bespoke classifications in the health, education and policing domains, with respect to issues of granularity, refresh rate, provenance, etc. Health demographics considered include prescriptions and GP practice records, crime through reported thefts and violence, education as literacy rates and qualifications and retail as consumer types and income. Crime data will present a particular issue as they are less associated with geographies of home addresses. These are then compared to OAC socio-economic categories in evaluating the degrees of variance. The R statistical tool is used both in this task and in the investigation of cluster analysis procedures. Scale and aggregation issues are investigated, connoting issues related to areal interpolation, the modifiable area units problem and the small areas problem.

Open data will also be used to investigate the representation of particular sub-populations in open geodemographics. Some demographic groups such as students have proved hard to reach, because of consistent under-reporting and short term population movements. Feasibility analysis is undertaken in the assessment of open data to issues such as scaling and aggregation. Some of the more general issues geodemographics face are the simplified representation of people living within classified areas, automated production of space through 'software sorting' and implications relating the surveillance and privacy (Goss 1995). Preliminary work will focus on the study area of Central Manchester, chosen due to its well dynamic socio-economic communities and therefore the difficulties in producing reliable geodemographic classifications of communities; 'reliable' meaning a geodemographic classification which best reflects the current characteristics of the population to a chosen zonal geography. The size of the study area is also appropriate to the scope of this paper.

5. Conclusion

The absence of linkage in open data presents a fundamental weakness within publicly available geodemographic classifications. The current and near-future release of large open-source datasets must also be explored through national information frameworks and therefore be considered strategically. Feasibility audits of the usefulness of currently available open-source data is undertaken within different sectors, in order to provide a clearer picture of the current state of open data for use in geodemographic classifications. It is hoped that this will stimulate a strong movement towards bespoke, freely accessible and transparent geodemographic classifications rather than the 'black box' traits of private geodemographic classifications.

References

Experian (2010). "Optimise the value of your customers and locations, now and in the future Mosaic UK - the consumer classification of the United Kingdom." Mosaic UK Brochure, June 2010.

Goss, J. (1995). "We know who you are and we know where you live: instrumental rationality of geodemographic systems." Economic Geography **71**(2): 171 - 198.

Harris, R. and R. Johnston (2008). "Primary schools, markets and choice: studying polarization and the core catchment areas of schools." Applied Spatial Analysis and Policy **1**: 59 - 84.

Harris, R. J. and P. A. Longley (2002). "Creating small area measures of urban deprivation." Environment and Planning A **34**: 1073 - 1093.

House of Commons Science and Technology Committee (2012). "Decision to scrap Census could hit UK social science according to MPs: 21 September 2012." Retrieved 24/10/2012, 2012, from http://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/news/120921science-and-social-science-report-published/

Longley, P. A. and M. F. Goodchild (2008). The use of geodemographics to improve public service delivery. Managing to improve public services. J. Hartley, C. Donaldson, C. Skelcher and M. Wallace. Cambridge, Cambridge University Press.

NIF: APPSI (2012). "A National Information Framework for Public Sector Information and Open Data: Advisory panel on public sector information." APPSI DISCUSSION PAPER, October 2012

Singleton, A. and P. Longley (2009). "Geodemographics, visualisation, and social networks in applied geography." Applied Geography **29**(3): 289 - 298.

Sleight, P. (2004). "An introductory review of geodemographic information systems." Journal of Targeting, Measurement and Analysis for Marketing **12**: 379 - 388.

Vickers, D., et al. (2005). "Creating the National Classification of Census Output Areas: Data, Methods and Results." University of Leeds, Working Paper 05/2.

Webber, R. (2007). "The metropolitan habitus: its manifestations, locations, and consumption profiles." Environment and Planning A **39**: 182 - 207.

Wright, C. and L. Sparks (1999). "Loyalty saturation in retailing: exploring the end of retail loyalty cards?" International Journal of Retail & Distribution management **27**(10): 429 - 440.

Biography

Matthew Pratt completed his MSc in Applied GIS and Remote Sensing from the University of Southampton in 2012. He is now undertaking a PhD relating to geodemographics at UCL under the supervision of Professor Paul Longley and Dr. James Cheshire.

Paul Longley's research interests concern the use of GIS and quantitative methods in urban analysis. He is a co-editor of Environment and Planning B and co-author of the book 'Geographic Information Systems and Science' and c. 150 other refereed publications.

James Cheshire currently works at CASA and has a background in mapping clusters of UK surnames and data mining, more specifically of twitter feeds and users.

Chris Gale is a third 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.