Exploring social dynamics: predictive geodemographics

Jennie Gray
Leeds Institute for Data Analytics
University of Leeds
Leeds, UK
gyjhg@leeds.ac.uk

Lisa Buckner
School of Sociology and Social Policy
University of Leeds
Leeds, UK
L.J.Buckner@leeds.ac.uk

Alexis Comber
School of Geography
University of Leeds
Leeds, UK
a.comber@leeds.ac.uk

Abstract

Geodemographics is the analysis of people by where they live, underpinned by the notion that people residing within proximity exhibit similar characteristics. Geodemographic classifications are typically area classifications built with composite variables, in order to segment populations into homogenous groups. However, such composite variables are often collated from static data such as population census, and other administrative and or commercial sources which are not regularly updated. Furthermore, most commercial systems are black-boxes: they do not provide information about their data, methods, updates or clusters. Consequently, this means that geodemographic classifications offer out of date and opaque representations of the populations that they seek to segment. Since such classifications have a range of public and private sector uses such as education, policing, health, targeted advertising, and location optimisation, such static data sources create many temporal limitations, and drawbacks. This research describes the social processes that suggest changes in socio-economic status and therefore geodemographic classification. Using a case study, it explores how social dynamics and geodemographic trajectories over time can be predicted from data that capture social process and enable area's future geodemographics to be predicted. Socially dynamic processes include gentrification (area improvement, thus an improved geodemographic cluster trajectory), urban decay (area deterioration, thus a worsened geodemographic cluster trajectory), and area stability (thus a stagnated but stable geodemographic cluster trajectory). A number of future research areas are identified.

Keywords: Geodemographics, spatiotemporal analysis, predictive analytics, spatial dynamics.

1 Introduction

Geodemographics is often described as the analysis of people by where they live (Harris et al, 2005). However, in reality it provides a characterisation of areas through the typical characteristics of the people that live there. Geodemographics is underpinned by the notion that people residing within close proximity of another exhibit similar characteristics than those further apart (Harris et al, 2005).

Contemporary geodemographic classifications have origins in the work of human ecologists (Singleton and Spielman, 2014). The first contemporary geodemographic classification was developed in the UK in the 1970's by Richard Webber, who created a ward level classification called ACORN (Webber, 1977). These early activities were targeted towards local authorities with an interest in understanding the distribution of people, housing, and social deprivation (Birkin and Clarke, 2009), with ACORN developed from work exploring inner-city deprivation in Liverpool. The development of commercial geodemographic classifications was to support spatially targeted marketing (e.g., Baker et al, 1997) with increased spatial resolution (Sleight, 2004).

Geodemographic classifications have a number of limitations, (Gale and Longley, 2013; Harris and Feng, 2016). Two major ones are that they are typically developed using census data, published 2 to 3 years after the census period, and are temporally static. These temporal limitations mean that geodemographic classifications fail to capture the dynamic nature of an area (Gale and Longley, 2013).

The aims of this paper are to explore methods for generating dynamic classifications that give indication of area trajectories. It uses a case study in Sheffield, UK and explores the utility of open data to capture dynamic demographical processes.

2 Background

Geodemographic classifications have traditionally been developed by academics for commercial interests. Publication of their methodologies has been rare, consequently being described as 'black boxes' (Susser, 2004) since the subjective choices made during their development are undocumented. Recently, open geodemographics have been developed such as the Output Area Classification (OAC) in the UK, which are free to access, and publish their methodology (Gale et al, 2016).

However, despite data being increasingly recognized as spatiotemporal (Cressie and Wikle, 2011), the temporal dimension of data has generally been overlooked within the classification systems.

2.1 Temporal Limitations

The reliance of geodemographic classifications on static data sources such as censuses generates several spatial and temporal limitations such that geodemographic classifications can easily be anachronistic. While seemingly current they can present out of date and obsolete social patterns and result in mis-guided decision making (Gale and Longley, 2013). Such temporal limitations include the Modifiable Temporal Unit Problem (MTUP), comparable to the MAUP effect (Coltekin et al, 2011). There are three significant considerations regarding the temporality of data; their persistence and duration, resolution, and point in time (Coltekin et al, 2011). These can have important aggregation, segmentation, and boundary effects, which have impacts on the space-time clusters detected in analysis (Cheng and Adepeiu,2014). Temporal issues thus add an extra layer of uncertainty into geodemographic classifications (Gale and Longley, 2013).

Little work has been conducted to overcome these temporal classification limitations, with the exception of Singleton et al (2016). OAC classifications from 2001 and 2011 were used to create a Temporal OAC in order to analyse the stability of geodemographic clusters. The results indicated that 39% of OAs in 2011 were reassigned from their 2001 counterpart (Singleton et al, 2016), suggesting some level of geodemographic cluster instability, or geodemographic change. In reality, local areas are dynamic and may undergo changes in that are not captured by decadal censuses. However, this study used only two points in time separated by a decade, and the results may severely misrepresent the actual social dynamics within this period.

Thus, research is needed to explore in greater depth the dynamics of geodemographics through the analysis of intercensal data in order to develop critical understandings of the characteristics of geodemographic change, and the drivers of that change.

2.2 Geodemographic Processes

In order to get a handle on how to deal with dynamic data for predictive geodemographics, it is important to explore the social processes that potentially drive the changes seen in geodemographic classifications, such as the 2001 and 2011 OACs. Geodemographic classifications are generally hierarchical in nature, and since they are an accumulation of socio-economic and demographic variables, can be indicative of socioeconomic class (Burrows and Gane, 2006). Geodemographic trajectories are likely to be both in upward and downward directions thus, regarding social mobility, moving either to a higher or lower geodemographic cluster. Subsequently, social processes that represent both upward and downward socio-economic and geodemographic trends, as well as those maintain current conditions, need to be considered.

Gentrification is attributed to much of the positive change that has occurred in city and town centre locations and is defined as the reinvestment of capital into the urban centre, designed to create spaces of affluence (Hackworth and Smith, 2001). It is a highly dynamic process (Smith, 1986) and meaning of the term has grown and changed. Gentrification can now be attributed or associated with several geodemographic processes including, a reduction in male working class and an increase in female employment, the loss of manufacturing employment, and an increase in service employment (Short, 1989), alongside a rapid rise of managerial and professional occupations (Bell, 1973). Gentrification is thus is a proxy for upward geodemographic change.

Conversely, urban decay is attributed to much of the negative changes that occur in urban neighbourhoods. Detroit famously suffered from intense deterioration caused by social processes such as racial segregation, loss of employment, and population reduction, all of which contributed to its deterioration, and abandonment (Sugrue, 2014). Therefore, urban decay is a proxy for downward geodemographic change.

In order to explore how these processes may impact geodemographic cluster assignment, data associated with these processes are included within the study.

3 Data and Methods

The study area is Sheffield, a local authority in South Yorkshire, UK, was selected due to its geographical and demographic diversity. Sheffield encompasses urban and rural areas, and the uplands of the Peak District (SCC, 2016). Demographically, Sheffield is ethnically diverse with 19% of the population from minority groups (SCC, 2018). Sheffield is also ranked 26th in England and Wales by the proportion of LSOAs in the 10% most deprived (SCC, n.d), suggesting a range of scales of deprived neighbourhoods.

3.1 Datasets

The OAC is used within this study, alongside the Temporal OAC (TOAC) since it provides estimated annual cluster assignments. The OAC has 3 hierarchical levels, while TOAC clusters are available at the highest level ("supergroup") only.

Six datasets at LSOA level (see Table 1), alongside TOAC assignment (OA level), are used for analysis. They are updated annually, and cover a period of 10 intercensal years, in order to determine which data best explain the observed changes in geodemographic cluster assignment through 2001-2011, and the processes they represent. Furthermore, this study period enables demographic trajectories to be identified and validated by the 2011 OAC.

Table 1: Datasets

Dataset	Source
DWP Benefits: State Pension, Income Support, Job Seekers Allowance	ONS (Nomisweb)
Median House Price	CDRC
Number of Cars	DFT/DVLA
Population Estimates	ONS
Temporal Output Area Classification	CDRC

Data were collected via availability sampling, since the ideal administrative data that are used as census inputs at LSOA level simply did not exist. However, they still represent social processes of interest for example, Job Seekers Allowance provides a proxy for unemployment, while population estimates may suggest migratory pressures.

However, some datasets are collated at the 2001 LSOA boundary, while others at 2011 LSOA boundary. These differences in spatial boundaries adds extra complexity.

3.2 Methods

This study takes a spatiotemporal regressive model, a restricted maximum likelihood (REML) (e.g., Welham et al., 2004) approach for identifying the socio-demographic drivers of geodemographic change from dynamic data sources and predicting future cluster assignments.

A change in geodemographic cluster assignment is a collective property of the accumulation of socio-demographic processes. Understanding geodemographic change is critical to planning for allocative efficiency in numerous fields including health, education, policing etc. With this basis, the understanding of geodemographic change would provide insight into how the mechanisms of many socio-demographic phenomena interact to influence change, and how different dominant processes will lead to different geodemographic changes.

With the collation and collection of dynamic data, geodemographic changes reveal the spatial and temporal outcomes of socio-demographic processes. Subsequently, the study workflow in Figure 1 was devised.

Figure 1: Study workflow

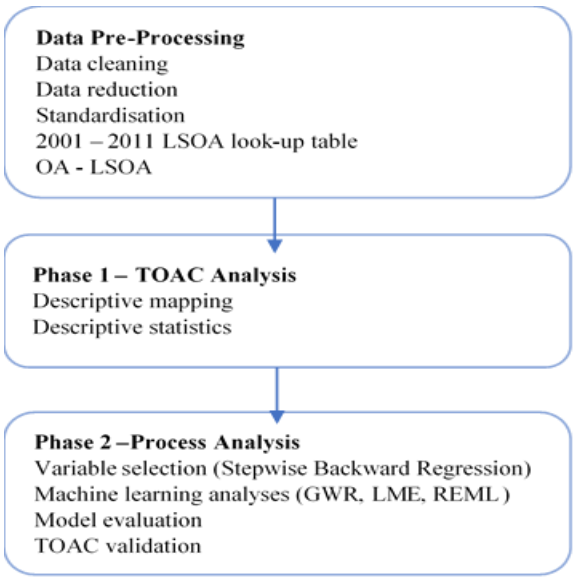


Figure 1 displays the main steps of the study, which was conducted in R. Pre-processing was a major component of the study. The initial phase generated descriptive analyses of the annual TOAC cluster reassignments within Sheffield OAs. The second phase comprises machine learning analyses upon the dynamic data in order to develop predictive geodemographic models and clusters. Variable selection is performed in phase 2 in order to reduce the number of variables used from 53. Once conducted, regression (geographically weighted, and linear mixed models) are run to generate regression residuals, that are then modelled in REML.

REML offers a flexible estimation framework, and allows for spatial and temporal correlations, and combining series data (O'Neill, 2010; Nabugoomu, 1994). Thus, REML is suitable for this study, and is utilised to account for the spatiotemporal variability of the socio-demographic processes within geodemographic change.

4 Results and Discussion

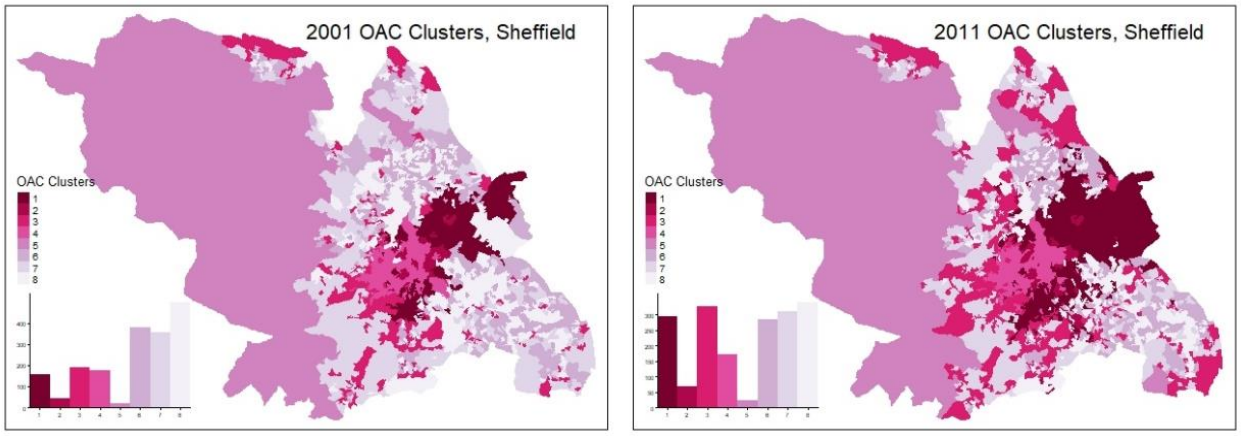
Initial results show that a total of 511 of 1817 Sheffield OAs was reassigned from their 2001 OAC cluster in 2011, equivalent to 28.1%. Figure 2 displays a map of OA cluster change from the start of the study period, to the end. The legend histogram shows that during this period, the number of clusters 1 and 3 increased. Spatially, these are within the city centre, expanding into the areas previously classified as 6, 7, and 8, and to the East of Sheffield. Smaller changes are recognized towards the North East, South East, and South West of Sheffield. This suggests that these areas, specifically those reclassified to cluster 1 (Suburban Diversity), may have experienced processes associated with urban decay, since they are associated with diverse ethnicities, an aging population, and overcrowded, rented living, with high unemployment (Gale et al, 2016). Those being reclassified from cluster 7 (Professional Propensity) and 8 (Hard-up Households) would be of particular interest, since these two clusters represent both high and low sociodemographic status respectively. Thus, there is no hierarchical clustering of the TOAC like in other geodemographic classifications, 1 does not represent the highest sociodemographic status, incrementing down to the lowest, rather the TOAC clusters are somewhat randomly numbered. However, these clusters though not hierarchical, they are still indicative of sociodemographic standing.

With this understanding, those reassigned from cluster 7 or 8 to cluster 3 (Intermediate), may be in the process of relative decline or improvement, since Intermediate clusters have few defining characteristics. However, the above decennial change is not representative of the annual changes in cluster assignment as expressed by the TOAC's yearly OAC cluster assignments. The greatest change with 91 (5%) of OAs being reassigned to a new cluster was 2001-2002, whilst 2006-2007 had the least change with only 45 (2.5%) reassignments. Table 2 shows the number of cluster reassignments for each year of the study period, alongside their percentage.

Table 2: Annual TOAC Reassignment of Sheffield OAs

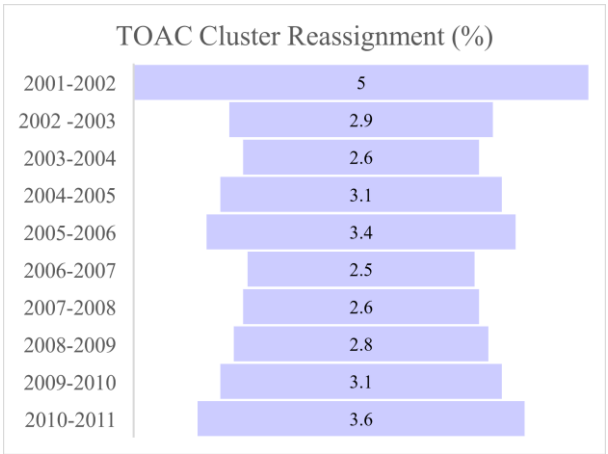
Year	No. of TOAC Cluster Reassignments	% of Change
2001-2002	91	5.0
2002 -2003	53	2.9
2003-2004	48	2.6
2004-2005	56	3.1
2005-2006	62	3.4
2006-2007	45	2.5
2007-2008	48	2.6
2008-2009	51	2.8
2009-2010	57	3.1
2010-2011	66	3.6
	Total: 577	Total: 31.6
2001-2011	511	28.1

Figure 2: OAC Cluster Reassignment, 2001 to 2011 for Sheffield



Analysis of the TOACs has shown that analysis at a temporal resolution of 10 years, does not account for the small patterns of change within the 10-year window. Figure 3 seemingly shows a five-year cycle to the geodemographic change within this 10-year study period. This alongside the decadal observations, suggest that there are several socio-demographic processes influencing the level of change seen within the geodemographic cluster reassignments. Therefore, examining these processes further would gain insight into the interaction of these processes and how they influence change, which processes hold more influence for different kinds of change, thus best explain that specific geodemographic change.

Figure 3: Geodemographic Change in Sheffield 2001-2011, represented by TOAC cluster reassignment



The most stable cluster was 5, with only 6 TOAC reassignments throughout the period, just 1% of the total. The least stable was cluster 8, which saw the most change with 158, or 27%, TOAC reassignments. Since the reassignment of OAs is not equal throughout the geodemographic clusters, this analysis could provide the basis for further examination of the stability of geodemographic clusters regarding the probability of OA membership to specific clusters. Subsequently, such results suggest that a fuzzy geodemographic classification with probability of belonging may be more representative of real-world experiences.

Phase 2 of the research explores the relationship between the data variables via regressive REML, and how they represent the geodemographic processes of upward and downward trends. It determines which of those best explain the geodemographic cluster changes. These variables may therefore be able to be used as proxies for geodemographic future cluster change, enabling future projection, and creating the foundations of predictive geodemographics. The results additionally provide a validation of the TOAC, when analysed against the geodemographic processes' variables.

Though this research is not yet finalized, with the exploration of the geodemographic process variables in phase 2 to be completed, this research has shown that the temporal resolution of data greatly impacts results, and MTUP should therefore be an important consideration in the development of geodemographic classifications.

Furthermore, these results have shown that when analysing the temporal change of geodemographic clusters, intercensal data and clusters provide a more robust analysis. This therefore has the power to suggest that intercensal administrative data should be implemented wherever possible in current general-purpose geodemographic classifications.

These results will suggest several areas of future work; how to select and handle data with greater temporal dynamics (e.g. real time data)? How to quantify changes within high-level cluster (i.e. changes in condition and quality)? When is the threshold of change enough to warrant a change in geodemographics class label? Can the early warnings / signals of change be identified? How should the changes in the characteristics of areas (i.e. their attributes in the database) and the associated impacts on statistical segmentation routines be handled? This research, and the wider questions it suggests have the potential to support decision-making practice in several fields.

5 Acknowledgements

This research was funded by the ESRC Centre for Doctoral Training – Data Analytics and Society, ES/P000401/1.

References

Baker, K., C. McDonald and J. Bermingham. (1997) The utility to market research of the classification of residential neighbourhoods. *International Journal of Market Research*, 39(1). Available at: https://www.warc.com/ [Accessed 6th September 2018].

Bell, D. (1973) *The Coming of Post-Industrial Society*. New York: NY Basic Books

Birkin, M. and Clarke G. (2009) *Geodemographics*. *International encyclopaedia of human geography*. Oxford: Elsevier.

Burrows, R., and Gane, N. 2006. Geodemographics, Software and Class. *Sociology*. 40(5), pp.793-812. Available at: https://www.academia.edu/ [Accessed

Cheng, T. and M. Adepeju. (2014) Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection. *PloS one*, 9(6). Available at: https://www.ncbi.nlm.nih.gov/ [Accessed 10th September 2018].

Çoltekin, A., S. De Sabbata, C. Willi I. Vontobel, S. Pfister, M. Kuhn and M. Lacayo. (2011) Modifiable temporal unit problem. In: *ISPRS/ICA workshop "Persistent problems in geographic visualization" (ICC2011)*, Paris, France.

Cressie, N. and C. K. Wilke (2011). *Statistics for spatio-temporal data*. John Wiley & Sons.

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

Gale, C., and Longley, P. (2013) Temporal uncertainty in a small area open geodemographic classification. *Transactions in GIS*, 17(4), pp. 563-588.

Hackworth, J, and Smith, N. (2001) The changing state of gentrification. *Tijdschrift voor economische en sociale geografie*, 92(4), pp. 464-477.

Harris R and Feng Y. (2016) Putting the geography into geodemographics: using multilevel modelling to improve neighbourhood targeting - a case study of Asian pupils in London. *Journal of Marketing Analytics*, 4(2/3), pp. 93-107.

Harris R. P Sleight and Webber R. (2005) *Geodemographics, GIS and neighbourhood targeting.* Chichester: John Wiley & Sons.

Nabugoomu, F, (1994). REML and the analysis of series variety trials. Doctor of Philosophy. Edinburgh: The University of Edinburgh.

O'Neill, M. 2010. *ANOVA & REML: a guide to linear mixed models in an experimental design context*. Available at: http://stats.net.au/ [Accessed 5th April 2019].

Sheffield City Council. (2016). *Sheffield Trees and Woodland Strategy 2016-2030.* [Online] available from: www.sheffield.gov.uk.

Sheffield City Council. (2018). *Sheffield's Population.* [Online]. Available from: https://www.sheffield.gov.uk/

Sheffield City Council. (N.d) *Sheffield Factsheet*. [Online] available from: www.sheffield.gov.uk.

Short, J.R. (1989) Yuppies, yuffies and the new urban order. *Transactions of the Institute of British Geographers*, 14(2), pp. 173-188. Available at: https://www.jstor.org/ [Accessed 16th February 2019].

Singleton A Palvis M and Longley P. (2016) The stability of geodemographic cluster assignments over an intercensal period. *Journal of Geographical Systems*, 18(2), pp. 97-123.

Singleton, A., Spielman, S.E (2014) The past, present, and future of geodemographic research in the United States and United Kingdom. *The Journal of the Association of American Geographers*, 66(4). Available at: https://www.tandfonline.com/ [Accessed 15th February 2019].

Sleight, P. (2004) *Targeting customers: how to use geodemographic and lifestyle data in your business*. Oxford: World Advertising Research Centre.

Smith, N. (2000) Gentrification in R.J Johnston, D Gregory, G Pratt and M Watts, editors. *The Dictionary of Human Geography*. Oxford: Blackwell.

Sugrue. T.J. (2014) The origins of the urban crisis: Race and inequality in post-war Detroit. Oxford: Princeton University Press.

Susser, E. (2004) Eco-epidemiology: thinking outside the black box. *Epidemiology*, 15, pp. 519-520. Available at: https://www.ncbi.nlm.nih.gov/ [Accessed 16th August 2018).

Webber, R. (1977) *An Introduction to the National Classification of Wards and Parishes*. London.

Welham, S., Cullis, B., Gogel, B., Gilmour, A., and Thompson, R. (2004). Prediction in linear mixed models. *Australian & New Zealand Journal of Statistics*. 44 (3), pp. 325-347. Available at: https://onlinelibrary.wiley.com/ [Accessed 8th April 2019].