Health Applications for Open Geodemographics

J. Petersen, D. Ashby and P. Atkinson

Geography, University College London
Pearson Building, Gower Street, London WC1E 6BT
Telephone: +44 20 7525 0267
jakob.petersen@southwarkpct.nhs.uk

1. Introduction

Health geodemographics is a fast developing research applications area that focuses on the development and application of neighbourhood classifications to targeted interventions in the health care arena. Population health intervention concerns immediate threats from infectious diseases or much longer term effects of chronic diseases like heart diseases or diabetes (Fielding 1999). In this paper we describe how the development of geodemographic tools both provides a framework for improving our understanding of chronic diseases and at same time presents a useful tool for targeting of public health intervention.

Health inequalities with regard to the incidence and effects of chronic diseases reflect complex social structuring in sex, age, life course exposure, phenotype and lifestyle (McKinlay and Marceau 2000, Davey Smith 2003, Siegrist and Marmot 2006). The concept of 'lifestyle' provides an umbrella term for the various behaviours that may have an impact on health: what we eat, our use of stimulants, our leisure time use and exercise levels, etc.

Previous research has sought to trace the underlying causes of many chronic diseases using simplistic statistical designs to produce circumstantial evidence linking a multitude of factors, under a 'black box' paradigm that has prevailed in the epidemiological literature since 1950s (Susser 2004). Rather than contemplating the seemingly infinite permutations of risk factors that might be put into the 'black box' we should also explore the opportunities arising from linking large georeferenced health care databases to socio-economic data. Geodemographic profiles offer a snapshot of the distribution of health outcomes in the population and allows for broader interpretations about the unfolding of chronic diseases.

Geodemographic systems have been developed for different purposes and at different scales (Sleight 2004, Harris et al. 2005). For marketing purposes geodemographics has been a convenient way to extrapolate consumer survey knowledge to new geographic markets. The creation of an open source geodemographic system by the Office for National Statistics (Vickers et al. 2005), Output Area Classifcation (OAC) marks a turning point in geodemographics, which now has an open source and is accessible to non-profit organisations.

Many questions however remain unanswered: how useful are these systems for the health care sector (public sector in general) and how do we improve and evaluate them for their new purposes? In this paper we present an alternative Output Area Classification for Greater London (LOAC) and a method to evaluate the actual and maximal market penetration potential of geodemographic systems.

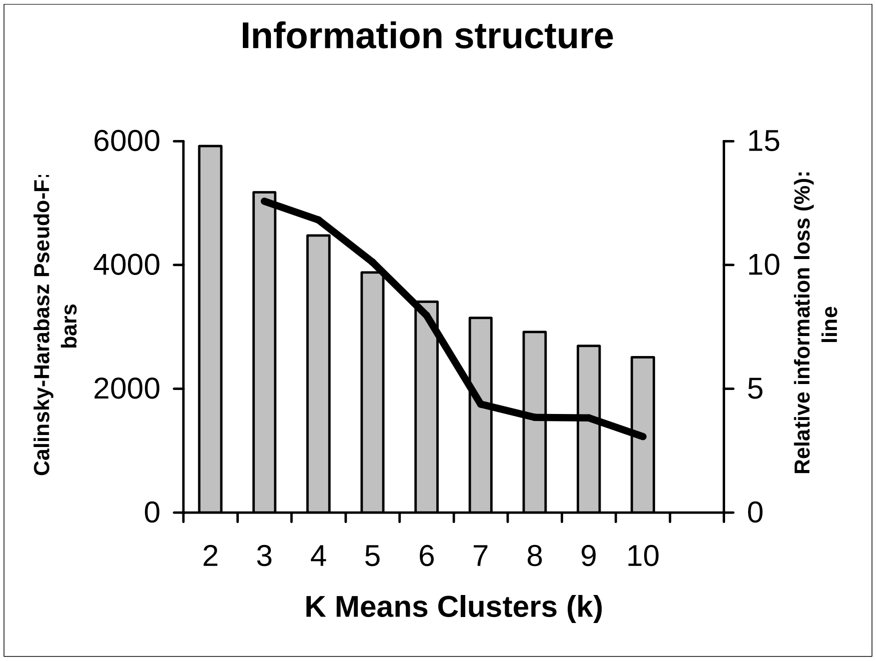

Figure 1 Stopping rule applied in the creation of London Output Area Classification

2. Output Area Classification

The OAC is based on a selection of 41 Census 2001 variables ranging from age to ethnicity, family structure, tenure, education, occupation, transportation, and health (Vickers et al. 2005)[1]. The variables were logarithmically and range-transformed to reduce the impact of outliers. The data set was first divided into 7 Supergroups using k-means clustering. The stopping rule for the generation of Supergroups was guided by the decreasing mean centroid distance in subsequent clustering (k+1) and to the fact that it is difficult to differentiate between much more than 7 colours when the classification is mapped.

3. Creating an Output Area Classification for London

A single national system enables a level of comparison across the system, so that e.g. national surveys may be projected to most areas. Yet in our work with planning at a local level in an inner-city borough of London (Southwark) the OAC appears overly vague, with swathes of central London ascribed to a single dominant "Multicultural" category. The greatest differentiation seems in contrast to be in the suburban boroughs. We suggest that this - in part - is an artefact in the way the OAC was created. While most of London is unified in being different to most other parts of the country, local level differences become subordinate in the classification. This is as much a general observation for capital cities as explained by rank size rules (Batty 2006) and central place theory (Webber 2004), but is also well-documented in the UK (Dorling and Thomas 2004). In order to explore whether greater differentiation could be obtained for London neighbourhoods, we repeated the classification with two modifications:

1. Reduced the data set and its range standardisation to Greater London

2. Deployed a cluster stopping rule based on within and between cluster variability, Callinsky-Harabaz pseudo-F, (Rabe-Hesketh and Everitt 2004)

Seven Supergroups were formed according to a distinct threshold in the information structure (Figure 1). Each LOAC Supergroup was divided into a Group level following the same procedure. The 7 LOAC Supergroups was thus subdivided into a total of 49 LOAC Groups. This makes the LOAC comparable to the OAC Subgroup level with 50 subgroups contained in Greater London.

4. Market penetration potential for OAC vs LOAC

The market penetration potential of geodemographic systems can be assessed in *the degree to which a given system can represent the overall variation in a market with the lowest number of groups*. This minimum of groups can again be measured by the degree the system uncovers variability for a given attribute[2]. Rank order inequality measures have been suggested to measure this variability (Callingham 2006). For this study we have used the Gini-coefficient as a way of assessing this market penetration potential (Novak et al. 1992). In the extreme a system with 24,140 groups for London's 24,140 output areas would have the *maximal* market penetration potential (shown as the outer boundary in the radial diagrams, Figure 2 and Figure 3). However a large number of groups would be impractical and would work against the quintessential force of geodemographic systems: *to 'borrow' strength from other areas to extrapolate knowledge about particular attributes*. In assessing the market penetration potential we could have chosen any number of variables of relevance to social structuring and health. We have chosen to evaluate the OAC (and LOAC) variables, because it at the same time tells us something of the leverage effects of these variables in the classifications.

Examining the market penetration potential of the OAC (and LOAC) attribute variables in this way we found evidence for the two main gradients: ethnicity and tenure (Figure 2, Figure 3). The LOAC has a greater market penetration potential than the OAC at Supergroup level (n=7) and also at the higher level (n=49 and 50, respectively) (Figure 2 and Figure 3).

A comparison of the actual vs the maximal market penetration potential also revealed variables for which the OAC (and to a lesser degree the LOAC) system would be less efficient, *viz*. 0-4 yr olds, single pensioner households, no central heating, working from home, provide unpaid care, manufacturing, hotel/catering and health/social care employment.

5. Longterm Limiting Illness

The prevalence of limiting longterm illness is of particular interest for health care authorities, because it represents population groups with high and complex health care needs (Wagner 1998, Bodenheimer et al. 2002, DH 2004, Saxena et al. 2006). Better social and geographical detection of this group will aid outreach activities, facilitate case management and reduce emergency hospital admissions.

The OAC uses the Standardised Illness Ratio (SIR) for longterm limiting illness. This emphasises illness above the effects of ageing. Correlating the SIR with the Census variables used in OAC shows a positive correlation with variables associated with low-income housing, e.g. renting public, lone parent household, routine occupation, etc. These areas are on the other hand not associated with the level of all unpaid care (r=0.03). Comparing the standardised (SIR) with the proportion of longterm ill, i.e.

Variables	Longterm limiting ilness (Standardised Illness Ratio)	Variables	Longterm limiting illness (proportion, no age standardisation)
Longterm limiting illness (proportion)	0.73	Longterm limiting ilness (Standardised Illness Ratio)	0.73
Rent (public)	0.72	Age 65+	0.61
Lone parent household	0.59	Single pensioner household (pensioner)	0.58
Routine/Semi-Routine occupation	0.58	Rent (public)	0.48
Unemployed	0.58	Rate of unpaid care providers (50+ hr/week)*	0.43
People per room	0.56	Routine/Semi-Routine occupation	0.37
Black African, Black Caribbean or Black Other	0.49	Divorced	0.34
Divorced	0.48	Rate of all unpaid care providers*	0.22
Rate of unpaid care providers (50+ hr/week)*	0.41	Age 25-44	-0.36
Hotel & catering employment	0.38	Two adult no children	-0.41
Economically inactive looking after family	0.37	HE qualifications	-0.42
All Flats	0.31		
Rate of all unpaid care providers*	0.03		
Age 45-64	-0.34		
Financial intermediation employment	-0.42		
HE qualifications	-0.43		
Rooms per household	-0.49		
2+ Car household	-0.52		
Two adult no children	-0.55		

Table 1. Pearson's population weighted correlations between the prevalence of Longterm Limiting Illness (Standardised Illness Ration vs unstandardised rate) and other Census variables used in the creation of OAC. Only correlations above +/- 0.30 are shown. *) Calculated from Census 2001 health data (ONS 2006)

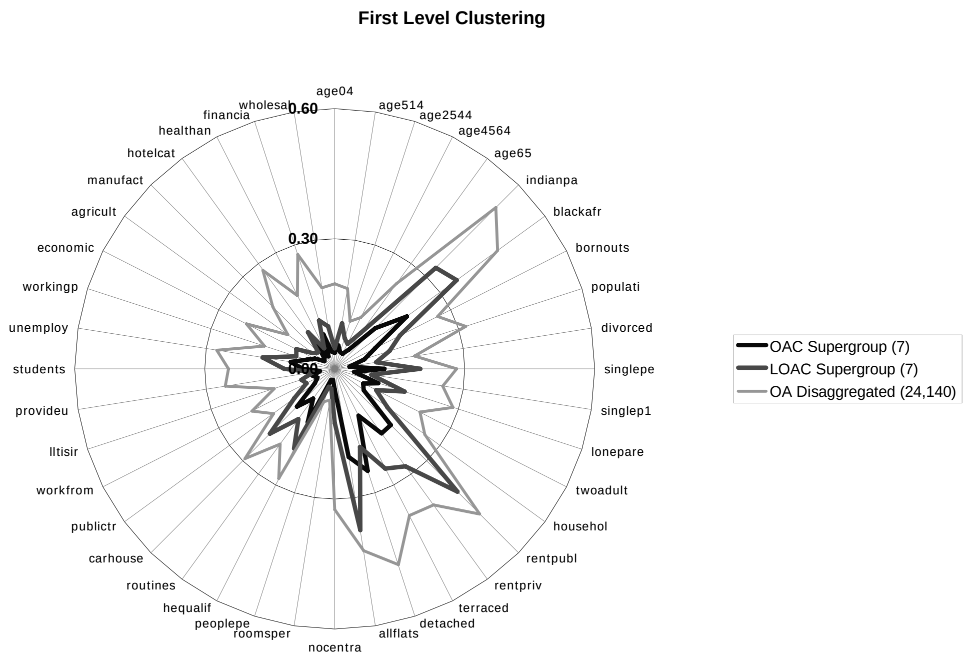

Figure 2. Market penetration potential at 1st level clustering: OAC vs LOAC

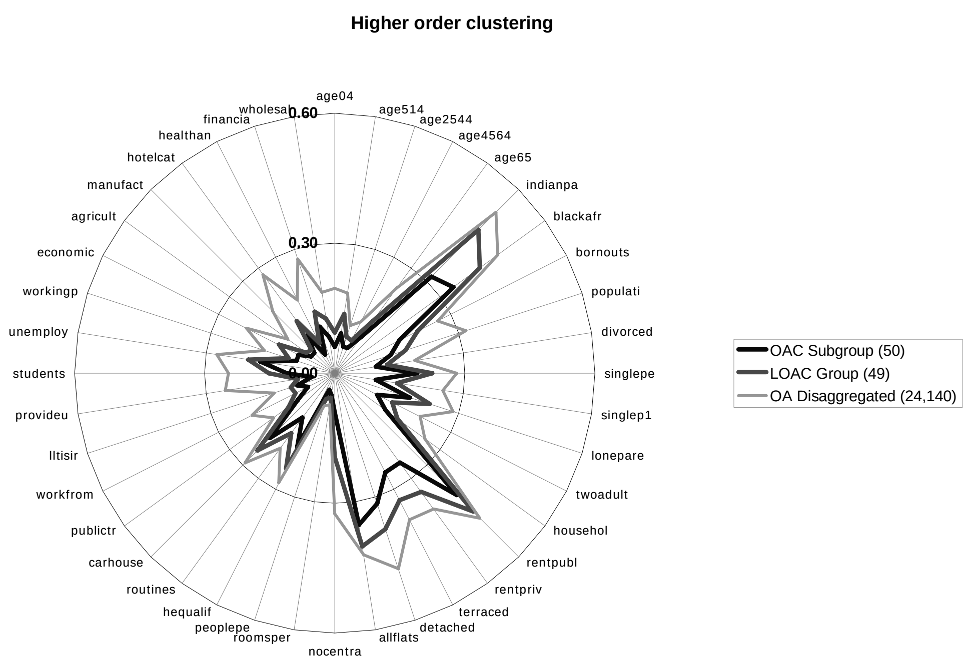

Figure 3. Market penetration potential at higher order clustering: OAC vs LOAC

not standardised for age,[3] we see greater association with old age, i.e. age 65+ and single pensioner households. In both cases the rate of unpaid care of 50+ hr weekly is positively associated (r=0.41 and 0.43) and so to a degree confirms the 'positive care law' for informal care in the home or the neighbourhood as defined by a small area unit like the OA (Shaw and Dorling 2004): *those with the greatest needs live near those most committed to provide unpaid care*. Comparing the SIR and crude proportions indicates that they represent different, but genuine groups with high health care needs.

6. Conclusion

The national Output Area Classification system (OAC) did not represent the variation measured as market penetration potential across 41 Census variables in Greater London very well. A simple classification only at regional level (Greater London) outperformed the OAC across all attribute variables. Of course, this improvement is achieved at the expense of making Greater London a 'special case', but the results of this preliminary analysis and health care application suggest that this is a price worth paying in exchange for the benefits of detecting greater variability.

We suggest that geodemographics provides a useful framework for studying the prevalence of diseases. The power of geodemographic systems, however, is likely to arise from extrapolating behaviours from surveys of health, consumption and leisure time use. Future work will lie in validating the predictions that can be made in combining lifestyle surveys and large georeferenced health care databases as well as developing new geodemographic systems based on health care data.

7. References

BATTY, M. 2006 Rank clocks. Nature 444: 592-596

BODENHEIMER, T., H. WAGNER, AND K. GRUMBACH. 2002. Improving primary care for patients with chronic illness: the chronic care model, part 2. JAMA **288**:1909-1914.

CALLINGHAM, M. 2006 Exploring the Use and Value of the ONS output area classification. http://www.rss.org.uk/rssadmin/uploads/577303_Callingham%20Descriptions%203%20Oct%202006.pdf Accessed 30 Nov 2006

DAVEY SMITH, G. 2003. Health Inequalities - lifecourse approaches. The Policy Press, Bristol.

DH. 2004. Improving chronic disease management. Department of Health, London.

DORLING, D. AND THOMAS, B. 2004 People and places. A 2001 Census atlas of the UK. Policy Press, Bristol

FIELDING, J. E. 1999. Public health in the twentieth century - advances and challenges. Annual Review of Public Health **20**:xiii-xxx.

HARRIS, R., P. SLEIGHT, AND R. WEBBER. 2005. Geodemographics, GIS and neighbourhood targeting. John Wiley & Sons Ltd., Chichester

LONGLEY, P. A., GOODCHILD, M. F., MAGUIRE, D. J. AND RHIND, D. W. 2001 Geographical Information Systems and Science. Wiley, Chichester

MCKINLAY, J., AND L. MARCEAU. 2000. US public health and 21st century: diabetes mellitus. Lancet **356**:757-761.

NOVAK, T. P., J. LEEUW, AND B. MACEVOY. 1992. Richness curves for evaluating market segmentation. Journal of Marketing Research **19**:254-267.

ONS 2006 The 2001 Census. http://www.statistics.gov.uk/census/. Accessed 1 April 2006

RABE-HESKETH, S., AND B. EVERITT. 2004. A handbook of statistical analyses using Stata. Chapman & Hall, CRC, Boca Raton.

SAXENA, S., J. GEORGE, J. BARBER, J. FITZPATRICK, AND A. MAJEED. 2006. Association of population and practice factors with potentially avoidable admission rates for chronic diseases in London: cross sectional analysis. Journal of the Royal Society of Medicine **99**:81-89.

SHAW, M., AND D. DORLING. 2004. Who cares in England and Wales? The positive care law: cross-sectional study. British Journal of General Practice **54**:899-903.

SIEGRIST, J., AND M. MARMOT, editors. 2006. Social inequalities in health - new evidence and policy implications. Oxford University Press, Oxford.

SLEIGHT, P. 2004. Targeting customers. How to use geodemographics and lifestyle data in your business. WARC, Henley-on-Thames.

SUSSER, E. 2004. Eco-epidemiology: Thinking outside the black box. Epidemiology **15**:519-520.

VICKERS, D., P. REES, AND M. BIRKIN. 2005. Creating the national classification of census output areas: data, methods and results. School of Geography, University of Leeds, Leeds.

WAGNER, H. 1998. Chronic disease management: What will it take to improve care fro chronic illness? Effective Clinical Practice **1**:2-4.

WEBBER, R. 2004 Central place theory and geodemographics. The application of central place theory to zones of residence. CASA Working Papers 89: 1-22

8. Author biography

Jakob Petersen is researching in health care applications and geodemographics in a knowledge exchange collaboration between Southwark Primary Care Trust (NHS) and University College London (UCL). He is studying for a phd at Centre for Advanced Spatial Analysis (UCL) and is supervised by professor Paul Longley, Dr David Ashby and Dr Philip Atkinson (NHS).

---

1 Clustering variables (Census 2001): Age 0 – 4, Age 5 –14, Age 25 – 44, Age 45 – 64, Age 65+, Indian/Pakistani/Bangladeshi, Black African, Black Caribbean or Black Other, Born outside UK, Population density, Divorced, Single person household (not pensioner), Single pensioner household (pensioner), Lone parent household, Two adult no children, Households with non-dependent children, Rent (public), Rent (private), Terraced Housing, Detached Housing, All Flats, No central heating, Rooms per household, People per room, HE qualifications, Routine/Semi-Routine occupation, 2+ Car household, Public transport to work, Work from home, Long Term Limiting Illness, (Standardised Illness Ratio), Provide unpaid care, Students (full time), Unemployed, Working part-time, Economically inactive looking after family, Agriculture/fishing employment, Mining/quarrying/construction employment, Manufacturing employment, Hotel & catering employment, Health/social work employment, Financial intermediation employment, Wholesale/retail employment

2 Is this variability spatial? In a system with a finite number of regions, greater attribute variability also has to infer a degree of spatial variability although the Gini measure deployed here would give the same result no matter the spatial arrangement of the same observations. In this way the proposed technique is not spatial by the definition given by Longley et al. (2001)

3 Proportions calculated from Census 2001 health table and denominators at OA level