Harris, R. J., & 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), 93-107. https://doi.org/10.1057/s41270-016-0003-1

(Journal of Marketing Analytics)

Title:

Putting the geography into geodemographics: using multilevel modelling to improve neighbourhood targeting – a case study of Asian pupils in London

Running title:

Multilevel modelling for neighbourhood targeting

Authors:

Richard Harris* and Yingyu Feng

School of Geographical Sciences, University of Bristol, University Road, Bristol, BS8 1SS

*corresponding author: rich.harris@bris.ac.uk

Richard Harris is Professor of Quantitative Social Geography at the University of Bristol, author of 'Quantitative Geography: the basics', co-author of 'Geodemographics, GIS and Neighbourhood Targeting', joint winner of the Market Research Society's 2006 David Winton Award for Technical Excellence, and recipient of the Royal Geographical Society's 2014 Taylor & Francis Award for excellence in the promotion and practice of teaching quantitative methods. He is also director of the Bristol Q-Step Centre, part of a £19.5 million programme to promote a step-change in quantitative social science training in the UK.

Yingyu Feng is a PhD researcher at the University of Bristol. Her research lies at the intersection of urban geography and housing economics, looking to understand housing market dynamics and neighbourhood effects on house prices through advanced model-based approaches. Recent work includes price predictions ('Chen Shu-Peng Youth Excellent GIS Award’, 2015 Institute of Electrical and Electronics Engineers Data Mining conference), housing market and neighbourhood segmentation ('Most Innovative Paper', 2016 Pacific Rim Real Estate Society), housing inequality at multiple geographical scales and the policy implications for urban planning and housing strategies (‘Best Graduate Paper’, 2016 Association of American Geographers conference).

Word count: 6,690

Putting the geography into geodemographics: using multilevel modelling to improve neighbourhood targeting – a case study of Asian pupils in London.

Abstract

This paper explores the use of multilevel modelling to provide a statistical framework for geodemographic analysis. It argues that combining a neighbourhood classification with a modelling approach to analysis allows the levels of the geodemographic hierarchy to be considered simultaneously, identifying those which are most appropriate to the analysis and allowing the apparent differences between neighbourhood types to be considered in regard to their statistical significance, and to the uncertainty of the estimates. The paper shows how the model can be extended to create a cross-classified multiscale model that makes better use of the locational information available and uses it to improve the efficiency of the neighbourhood targeting. The ideas are illustrated with a case study using a sample of data and the freely available London Output Area Classification to predict which neighbourhoods in London have the highest percentages of Asian school pupils. The multiscale model is shown to outperform the predictions made using geodemographics alone.

Key words: geodemographics, multilevel modelling, neighbourhood targeting, London, London Output Area Classification, segmentation

Introduction

This paper explores the use of multilevel modelling to provide a statistical framework for geodemographic analysis. It argues that combining a neighbourhood classification with a modelling approach that makes better use of the locational information available can better capture the geography of where a population subgroup is living, permitting improved targeting of it. To illustrate this, the paper takes a scenario common to users of geodemographics, which is to try and target as many of a group as possible from knowledge of where some of the group are living. In this application, the geodemographic classification identifies neighbourhood types where the group has highest probability to be living. The multilevel model enables the statistical significance of the probability to be assessed, and also allows extra geographical information to be included at a range of scales, benefitting the accuracy of the targeting.

Neighbourhood classifications have a long history in marketing, urban research, and in business and service planning. They are used to understand the spatial distributions of different groups of people, and to better target resources, marketing communications and public policy interventions to those that need or are most receptive to them (Birkin, 1995). There are many types of classification, designed for a wide variety of purposes (Harris, 2017). What they have in common is the knowledge that subgroups of a population rarely if ever are randomly nor uniformly distributed across a city, country or some other study region. Instead, the group disproportionately is found in some places more than others. This creates a link between location, often residential neighbourhood, and the probability of finding a member of the target group living there.

Geodemographics draw on this link. It is the analysis of people by where they live (Sleight, 2014). It is sometimes said that 'birds of a feather flock together' (Leventhal, 1993; Kestle, 2011) – that different types of places are occupied by different types of people who, although not identical (and therefore not reducible to the places they live in), may still display shared characteristics, common behaviours or be exposed to similar socio-economic circumstances within the same type of neighbourhood. Flocking implies free movement; in reality, choices and decisions are constrained. This means people are brought together not only through their own preferences and behaviours but also by differential opportunities to participate in and benefit from housing markets, labour markets, consumer markets and so forth. Either way, the geographical outcome is what spatial statisticians call positive spatial autocorrelation, memorably described by Tobler as 'the first law of geography' – everything is related to everything else but near things are more related than far things. (Tobler, 1970). The 'law' is not intended in a strict scientific sense (Goodchild, 2004; Sui, 2004) but instead acknowledges that there are often patterns of spatial clustering within human societies and nature.

Geodemographic analysis draws on this clustering but goes beyond the assumption that similar types of people reside in close spatial proximity to each other. It also works on the principle that similar types of people are found in similar types of neighbourhood, regardless of whether those neighbourhoods are themselves situated close to each other or not. It therefore assumes there are geographical clusters of like-minded people within neighbourhoods, and also geodemographic clusters of like-minded people that span across neighbourhoods with similar socio-economic and demographic profiles. To some degree this is an appeal to economic determinism – expressed crudely, 'you are where you can afford to live' – but lifestage, social background, employment, lifestyle, culture, ethnicity, urban morphology, industrial change and other processes or socio-spatial stratification can be cited as other linking factors.

In practice, geodemographic analysis is not dictated by any strong theory of what sifts and sorts people into different types of places. The usefulness or otherwise of a neighbourhood classification is tested empirically: it stands or falls by its ability or otherwise to differentiate between different types of consumer or to target the groups it is supposed to. We would not, for example, have much confidence in a classification targeting the most deprived neighbourhoods if the internal heterogeneity of those neighbourhoods meant they contained very few of the most deprived people (Voas and Williamson, 2002; Harris, 2002). In contrast, a successful application is one for which the spatial structure of the classification well captures the geographical patterning of the target group of interest.

This paper is not a critique of geodemographics. It takes as given the utility of neighbourhood classification in a wide range of applications (see below). Instead, it is interested in considering how the geographical patterning of the target group can be better identified and evaluated within the statistical framework of multilevel modelling, to assess the variations within and between neighbourhood types, to allow the most important scales of analysis to be identified, to measure the statistical significance of any one or more neighbourhood types in comparison to the others, and to better accommodate uncertainties in small samples of data to more accurately target a population subgroup. The paper suggests that there are gains to be made in the efficiency of neighbourhood targeting in ways that are easily achieved and at little if any additional cost to the analyst. After a brief discussion of the history and state of the art in geodemographics, the paper proceeds with an illustrative case study aiming to use a neighbourhood classification to capture the geographical patterning of Asian pupils in London from a small sample of them.

The past and present of geodemographics

Various histories of geodemographics have been written, amongst them Birkin (1995), Sleight (2004), Harris et al. (2005), Troy (2008), Singleton and Spielman (2014) and Leventhal (2016). Two origins commonly are identified. The first, a pre-cursor to modern approaches, is Charles Booth's classifications of the streets and census areas in turn-of-the- twentieth century London (Booth, 1888, 1902–3). His Descriptive Map of London Poverty, for example (the 1898–1899 revision of which can be viewed at http://booth.lse.ac.uk) shades streets according to the general socioeconomic conditions of the residents, including Class A, "the lowest class – occasional laborers, loafers, and semicriminals", and Class B – "the very poor – casual labor, hand-to-mouth existence, chronic want." From a modern-day perspective the descriptions seem a little patronising although it is hard to argue that they really are any worse that more recent descriptions such as “claimant cultures,” “struggling estates,” and “shotguns & pickups” (Goss, 1995; Curry 1998; Burrow et al., 2005).

The second origin is the Chicago school of urban sociology and its interest in the spatial structure of cities and processes of urban morphology, especially those forming what were termed natural neighbourhoods – geographical areas physically distinguishable from other areas within cities by the demographic and ethno-cultural characteristics of their populations. Much of the research undertaken by the Chicago school was ethnographic and qualitative in nature. However, from the 1960s/1970s onwards, an interest in measuring the underlying structures of cities, including the similarities and differences between the places where people live, was fuelled by the increased availability of small area data measuring neighbourhood populations, and the development of statistical techniques, including factor analysis, principal components analysis and data agglomeration techniques such as cluster analysis (Harris, 2017).

With the development of products such as PRIZM, ACORN, Mosiac and SuperProfiles, geodemographics emerged as a commercial industry within the market research sector from the late 1970s and early 1980s. At the same time there was interest amongst academics in creating free to use classifications (Charlton et al., 1985), an early inspiration for what has become known as open geodemographics – classifications that are free to obtain, and for which the input data as well as the clustering algorithms are also available and well documented, enabling the classification to be reproduced and customised as desired (Vickers and Rees, 2006, 2007; http://www.opengeodemographics.com/). The commercial growth led to resurgence in the use of neighbourhood classification in public service delivery (Longley, 2005) and also in academic research. Singleton and Spielman (2014) give a list of academic applications that includes research in urban policy, transport and utilities infrastructure, tourism and leisure, segregation, retail, migration, marketing and advertising, law enforcement and crime studies, access to internet and broadband, housing and real estate, health and well-being, finance, environmental and resource management, education, deprivation, and animal welfare. Leventhal (2016) provides applications in industry sectors.

The widespread interest in geodemographics has fostered innovation. Developments have tended to focus on the data side. The emergence of open geodemographics has been mentioned already. There has been interest also in the growing availability of administrative data and 'big data' to complement more traditional census based approaches (Leventhal, 2016), in part to draw on a richer range of data in the process of classification, and also to afford opportunity for real-time classification or to add a temporal dimension, including assessing the stability in geodemographic clusters over time (Furness, 2008; Singleton et al., 2016; for a critique, see Dalton and Thatcher, 2015).

Although new and interesting classifications have been devised and created, there has been less thought given to the analytical outcomes. Regardless of the classification or the ingenuity by which it was created, it is still usual to see a process by which simple index values are created: a value of 200 means there are double the expected number of a target group within a neighbourhood type; 50 means there is half; 100 means it is as expected. There is nothing especially wrong with this; it is simple and readily comprehensible. Nevertheless, what is less rarely considered is whether the differences between neighbourhood types can be judged statistically significant, at which scale of geodemographic classification the analysis is best undertaken or whether the differences between the neighbourhood types exist over and above that which can be explained by the use of other predictor variables. Rare exceptions to this are the paper by Harris et al. (2007) who use multilevel modelling to provide a statistical framework for geodemographic analysis and, in that case, to look for evidence of neighbourhood effects; and the paper by Nnoaham et al. (2010) who use multilevel modelling to ask whether geodemographic typologies explain variations in uptake in colorectal cancer screening.

This paper extends that previous work with a case study targeting Asian pupils in residential parts of London. It builds an argument for multilevel modelling as a natural complement to geodemographic analysis. Neighbourhood classifications are hierarchical: people live in places that are classified into neighbourhood types that are then further aggregated into 'super groups' (see below). Adopting multilevel modelling permits us to consider that hierarchy and to assess at what geodemographic scales the differences between people and places most matter for targetting. It also allows for the locational information to be more fully exploited, with the potential to improve the neighbourhood targeting.

An illustrative example

The rest of the paper provides an illustrative scenario where information has been collected about a sample of a larger population and the aim is to use a geodemographic classification to target neighbourhoods where members of the population most probably live. For the purpose of demonstration, the data are a 2 per cent random sample of all pupils that attended a state-funded state school in Greater London in 2011 (or, more correctly, are recorded in the National Pupil Database at either primary or secondary level; the vast majority of pupils are included). The neighbourhoods are UK Census small areas (known as Output Areas, OAs) and the target group is those who can be described as Asian Commonwealth pupils – pupils of an Asian ethnicity whose family has a heritage in a former British colonial country: Bangladesh, India and Pakistan. For brevity, this group will be described as Asian. The sample includes 8,420 pupils, of which 1,329 are Asian (15.8 per cent, compared to 15.4 per cent amongst all pupils in the population). The sample covers 6,573 of the 25,053 OAs in London (26.2 per cent) and is therefore (deliberately) limited in the information it provides. The geodemographic classification is the London Output Area Classification (LOAC). It is free to download from https://data.cdrc.ac.uk.

The target group, the pupils of Asian ethnicity, has been chosen as a deliberately non-demanding test of the classification: pupil data are not amongst those used to create the classification but more general and census-based measures of ethnicity are. They are not the same thing but are necessarily correlated (Asian pupils in state schools in 2011 are a subset of the whole 2011 Census population subgroup that is Asian). This will allow for a degree of common-sense groundtruthing. For example, if there is a low prevalence of Asian pupils in the neighbourhood type described by LOAC as 'Settled Asians' then we can reasonably assume there is something wrong, either in the analysis or in the classification itself.

Before turning to the classification, however, we can consider a simpler a way of targeting the neighbourhoods with the highest percentages of Asian pupils. We have the sample of pupils, and we know from their home postcode which small area neighbourhood (which Census OA) they reside in. It is therefore straightforward to calculate the percentage of pupils that is Asian per OA and to use that information to target the places where the percentage is greatest. Unfortunately, to do so encounters two problems. First, recall that only 6,573 of the 25,053 OAs are in the sample. For most OAs no calculation can be made. There is no information about them. Second, even for those that are included the mean number of sampled pupils per neighbourhood is 1.28 but with the first quartile, median and third quartile all at one. The small numbers involved mean that in the majority of neighbourhoods the percentages of pupils that are Asian can take on one of only two values: all (100 per cent) or none (0 per cent). The uncertainty in these estimates is too great to be useful.

The geodemographic classification does better because it allows data to be pooled over a shared neighbourhood type. There are, for example, 1,310 sampled pupils living in LOAC's Super group C, of which 466 (35.6 per cent) are, indeed, Asian. That is not a majority so arguably it is misleading to label the neighbourhoods as Settled Asians. However, geodemographics tends to focus on relative differences – if we were to pick a pupil at random from each of the neighbourhood types, the pupil selected from the Settled Asians type has the greatest probability of being Asian.

The percentage of the sample that is Asian in the Settled Asians neighbourhood type is 2.25 times greater than the percentage of Asian pupils in the sample as a whole, giving an index score of 225. These values are shown in Table 1, which provides the counts, percentages and index values for the neighbourhood types at the two levels of the LOAC hierarchy. LOAC is not unusual in having a choice of analytical scales. At the coarser, Super group scale it has eight distinct neighbourhood types. These sub-divide to form nineteen, more detail Groups. Encouragingly, the Groups named East End Asians (C3) and Bangladeshi enclaves (B2) contain the greatest percentages of Asian pupils within the sample. To reach Asian pupils it is logical to target either Super group C or, for greater spatial precision, Groups C3 and B2.

[TABLE 1 ABOUT HERE]

Placing geodemographics within a modelling framework

The percentages shown in the penultimate column of Table 1, from which the index values are derived, are (when divided by one hundred) equivalent to the sample probability of selecting a pupil who is Asian from each of the neighbourhood types. Those probabilities can also be calculated in a regression framework, using a logit model of the form,

$$\log_e\left(\frac{\hat{p}_{ik}}{1-\hat{p}_{ik}}\right) = -1.674 + \sum\nolimits_{n_k} \hat{\beta}_{.k} X_{.k} + \hat{\varepsilon}_i \qquad [1]$$

For this model, the y-variable is a dummy (categorical) variable coded as either one (if the pupil is Asian) or zero (if the pupil is not). The value $\hat{p}_{ik}$ is the estimated probability that the pupil is Asian, given the neighbourhood type in which they are living, and $\log_e\left(\frac{\hat{p}_{ik}}{1-\hat{p}_{ik}}\right)$ is the log odds of that probability (the log of the odds of a randomly selected pupil being Asian). The neighbourhood types enter the model as a series of dummy variables ($x = 1$ if the pupil lives in the neighbourhood type, else $x = 0$), where the number of dummy variables is equal to the number of types: eight at the Super group level or nineteen at the Group level. The value -1.674 is an offset equal to the overall proportion of the sample that is Asian, measured on the logit scale (i.e. $\log_e \frac{0.158}{1-0.158}$). Including it means we are modelling whether the probability (specifically, the log odds) of selecting an Asian pupil in each neighbourhood type is above or below the (mean) probability for the whole sample. The subscripts in the equation remind us that the model is implicitly hierarchical because the pupils (i) live in a neighbourhood type (k). The residuals (the model errors), $\hat{\varepsilon}_i$, are at the pupil level. Rearranging Equation 1 provides the probability per neighbourhood type of selecting a pupil who is Asian:

$$\hat{p}_{ik} = \frac{e^{-1.674+\hat{\beta}_k}}{e^{1-1.674+\hat{\beta}_k}} \qquad [2]$$

The model may be described as a fixed effects model where the probability per neighbourhood type is estimated on a logit scale as the increase or decrease in -1.674 by the amount $\hat{\beta}_k$. Those probabilities can also be estimated as a random intercepts model using an explicitly multilevel framework wherein

$$\log_e\left(\frac{\hat{p}_{ik}}{1-\hat{p}_{ik}}\right) = -1.674 + \hat{v}_k + \hat{\varepsilon}_i \qquad [3]$$

This model replaces the fixed estimates (the $\sum_{n_k} \hat{\beta}_{.k} X_{.k}$ in Equation 1) with a second sources of residual error: that at the neighbourhood level, $\hat{v}_k$, which is net of the error at the pupil level, $\hat{\varepsilon}_i$ (and *vice versa*). The neighbourhood level residuals are assumed to be from a distribution that is Normal, with variance, σ_k. The log odds of selecting an Asian pupil per neighbourhood type is now the increase or decrease in -1.674 by the amount $\hat{v}_k$. The greater the variance (the greater σ_k), the greater the geodemographic classification is able to discriminate between the neighbourhood types.

The multilevel model can be estimated using software such as MLwiN or, as here, the lme4 library for the open source software, R. Introductions to multilevel modelling are provided by Kreft and De Leeuw (1998), Snijders and Bosker (2011), Finch et al. (2014) and Robson and Pevalin (2015), amongst others. The model used here is not at all complicated: it is simply allowing the probability of selecting an Asian pupil to vary randomly from one neighbourhood type to another where random means that the differences between the neighbourhood types is assumed to arise as a random realisation of some population of neighbourhood types.

An advantage of modelling the probabilities within a regression framework is that confidence intervals can be calculated. This applies to both the fixed and random effects models, and they are shown in Figure 1 at a 95 per cent confidence having been converted back into percentage and index values. The estimated percentage of Asian pupils within a neighbourhood type may be regarded as statistically significant above or below the average for the sample when the 'error bar' does not cross the dotted line running horizontally across the charts. Figure 1 confirms the much higher percentage of Asian pupils in the neighbourhood type described as Settled Asians (Super group C), and most especially East End Asians (Group C3) and Bangladeshi enclaves (B2).

[FIGURE 1 ABOUT HERE]

There is little difference in the percentage estimates arising from the fixed or random effects models although the latter tends to be a little more conservative, pulling the estimates very slightly in towards the mean. It is also the more parsimonious: whereas the fixed effects model contains as many variables as there are neighbourhood types (eight or nineteen), the random effects model requires only the variance σ_k to be estimated. Consequently the random effects model outperforms the fixed effects model on a measure of model fit that penalises for model complexity such as the Akaike Information Criterion (AIC).

For either of the two models, the estimates by neighbourhood type allow us to extrapolate beyond the sample and the 6,573 Census neighbourhoods (OAs) it includes, to the remaining 18,480 OAs that are not in the sample but are in the geodemographic classification. We do so by assigning each OA an estimated percentage of Asian pupils that is the percentage for their neighbourhood type. These estimates can then be compared with the actual percentages of Asian pupils per OA for the 22,904 of OAs that contained a pupil in 2011.

To test the sample predictions against the actual values for the whole population would not normally be possible (and if it were, there would be no need to make predictions from the sample). However, here we are using a sample of a known population for which we have the full set of data. The Pearson and Spearman rank correlations between what is predicted using the geodemographic classification and the actual percentages of Asian pupils per OA are $r = 0.458$ and $r_S = 0.433$ at the Super group level, and $r = 0.647$ and $r_S = 0.488$ at the more detailed Group level. The correlations are the same for the fixed and random effects models. In either case the use of the geodemographic classification greatly improves upon a prediction based only on the sample data per OA, which gives much weaker correlations of $r = 0.370$ and $r_S = 0.118$. This highlights the value of using geodemographic classifications as a way to pool information across a relatively small sample and to better target neighbourhoods on that basis.

Extending the model

To this point there has been little to distinguish the multilevel random intercepts model from the more standard fixed effects model. However, the advantages of the former become obvious when the hierarchical structure of the data is considered more fully. The pupils (level *i*) reside in census OAs (level *j*) that are classified into neighbourhood Groups (level *k*) that further aggregate into Super groups (level *l*). In principle we might consider a fixed effects model of the form,

$$\log_e\left(\frac{\hat{p}_{ijkl}}{1-\hat{p}_{ijkl}}\right) = -1.674 + \sum_{n_l}\hat{\beta}_{.l}X_{.l} + \sum_{n_k}\hat{\beta}_{.k}X_{.k} + \sum_{n_j}\hat{\beta}_{.j}X_{.j} + \hat{\varepsilon}_i \qquad [4]$$

to provide probability estimates at each level of the geodemographic hierarchy. However, to do so would require 6,600 dummy variables: 6,573 at the OA level, 19 at the Group level, and 8 at the Super group level. This is too many to be sensible. It also conflates the scales of analysis. In doing so, it ignores the correlations between observations within groups (the errors, $\hat{\varepsilon}_i$, are assumed to be independent when, in fact, the hierarchical structure suggests they will be correlated within groups). In any case, because of the way the various parts of the hierarchy nest exactly into each other at an upper level, providing separate estimates of their effects will not actually be possible (there is an identification problem).

The better approach is the multilevel one, where

$$\log_e\left(\frac{\hat{p}_{ijkl}}{1-\hat{p}_{ijkl}}\right) = -1.674 + \hat{\omega}_l + \hat{\nu}_k + \hat{\mu}_j + \hat{\varepsilon}_i \qquad [3]$$

and $\widehat{\omega}_l$, $\hat{v}_k$ and $\hat{\mu}_j$ are random intercepts at the three levels of the hierarchy above the pupil. This requires only three estimates of the variance, σ_j, σ_k and σ_l to be made, and from them the variances at each level of the hierarchy can be compared. Of the total variance in the log odds of selecting an Asian pupil for the sample, 17 per cent is at the OA level, 56 per cent is at the Group level, and 27 per cent is at the Super group level. This implies what the earlier correlations had suggested: the best scale at which to undertake the neighbourhood targeting (using geodemographics alone) is the Group level because that is where you get the greatest variation and differentiation.

This is also evident from what are known as caterpillar plots, shown in Figure 2. Net of the differences at the Group and OA levels, the differences between the neighbourhoods at the Super group level are not especially strong: for example, the 95 per cent confidence interval for the Super group with the highest percentage of Asian pupils (Super group C: Settled Asians) overlaps with the estimates for Super groups G (Multi-Ethnic Suburbs) and B (High Density and High Rise Flats). At the Group level, Groups B2 (Bangladeshi enclaves) and C3 (East End Asians) are more distinctly different from the rest. The OA level is not shown but with the least amount of the variation being at this level, we may anticipate that the differences will not be large between OAs with the geodemographic differences having been taken into account.

If a choice were to be made then the analysis supports targeting at the Group level. However, the choice is unnecessary because the multilevel model provides simultaneous estimation of the variations at each level of the hierarchy, which then allows predictions to be made from whatever information is available for each OA: for those in the sample the OA level information can be considered in conjunction with the geodemographic level estimates; for OAs not in the sample only the geodemographic estimates are available. In this example the improvements in the predictions are marginal: $r = 0.661$ and $r_S = 0.498$ (compared to $r = 0.647$ and $r_S = 0.488$ previously, predicting from the Group level only). Nevertheless, having established the framework we can now take better advantage of the locational information available to improve the predictions further.

[FIGURE 2 ABOUT HERE]

Integrating geography into geodemographics

Recall that we have the postcode of each pupil in the sample and it is that which provides the link to the census OA and therefore to the LOAC classification of neighbourhood types. This is useful but it is not especially geographical. In fact, contrary to the geo in the title, geodemographic classifications are largely blind to geography (Harris et al., 2005). What we mean by this is that any geographical patterning of the neighbourhood types is an output of the classification and the data that went into it, not a necessary consequence of the classification algorithm. Neighbourhood types are formed by using census and/or other data to create statistical profiles of neighbourhoods and then by using those profiles to cluster neighbourhoods together on a like-with-like basis. Neighbourhoods that share a boundary may end up belong to the same geodemographic grouping but there is nothing in the process that would guarantee it. The outcome depends on the similarity of their profiles not upon their locations.

It may be argued that this is exactly what is required. Instead of assuming a geography *a priori*, any geographical patterning of the neighbourhoods emerges out from the data. This is a reasonable claim. The only weakness is that within a local area, the differences between geodemographic neighbourhood types may not be as great as the differences between the types imply. Referring back to Tobler's first law of geography, if near things are more related than far things, then we might expect local similarities that span across geodemographic groupings and are due to the local context.

Another way to appreciate the importance of geography and of local context is to recall that geodemographic analysis goes beyond the adage that birds of a feather flock together. It also assumes that similar types of people are found in similar types of neighbourhood, regardless of whether those neighbourhoods are located close to each other or not. That is a reasonable working assumption but also insensitive to spatial context. It would be beneficial to allow for both geodemographic variation and also geographical variation – to allow that the relevance of a neighbourhood type may vary dependent upon where it is located. Really this is a case of 'having our cake and eating it too.' To draw on the subtitle of a recently published book on geodemographics, it allows us to more fully use location in analysis for research and marketing (Leventhal, 2016).

A way to do this within the multilevel framework is to note that in addition to the geodemographic hierarchy (pupils into OAs into geodemographic types) there is also a census hierarchy (OAs into aggregations of OAs into the London boroughs). A cross-classified, multilevel model can consider both simultaneously and allow for the possibility that a greater percentage of the sampled pupils are Asian in the 'Bangladeshi enclaves' (B2) of the borough of Tower Hamlets than in the City of Westminster, for example. That proposition is, in fact, true and can be confirmed by a simple cross-tabulation by ethnic group and neighbourhood type: 201 of the 268 sampled pupils in the B2 neighbourhoods of Tower Hamlets are Asian (75 per cent), as opposed to the 1 of 10 in Westminster (10 percent). An issue with the cross-tabulation is that the small numbers problem reoccurs: the highest percentage is for the City of London, where the one pupil is Asian.

The advantages of what may be described as a multiscale, multilevel model are two-fold. First, it permits identification of which are the most important geodemographic and geographical scales to find differences between neighbourhoods. For the sample data, of the total variation above the pupil level, almost none of it is at the OA scale net of the other levels, 1.2 per cent is at the Lower Level OA scale (LLOAs, an aggregation of OAs), 5.2 per cent is at the Middle Level OA scale (MLOAs, an aggregation of LLOAs), 41.5 per cent is at the borough scale, 15.1 per cent is at the Group Scale, and 37.0 is at the Super group scale. The implication is that we should target at the Super group Scale but also recognise the differences between boroughs.

Second, the uncertainty in the estimates can be recognised. For example, in the upper caterpillar plot of Figure 3, the confidence interval for the City of London is very wide (and, in addition to this, the estimate for the City of London has been pulled towards the mean in recognition of the small sample size). In contrast, the borough of Tower Hamlets does appear to have a higher percentage of Asian pupils that is substantively different from other boroughs. In the lower plots, there is some suggestion that Super group C (Settled Asians) is different from the rest but differences between the Groups are now less evident, although they remain at the extremes (e.g. comparing Group B2 with B1).

Using the predictions at the various levels of the model, the correlations with the actual OA percentages of Asian pupils are now $r = 0.743$ and $r_S = 0.564$, increased from $r = 0.661$ and $r_S = 0.498$. This implies a gain in the efficiency of the targeting and it comes at little or any additional cost. All we have done is inserted a little geography into geodemographics.

[FIGURE 3 ABOUT HERE]

Summary and Discussion

Table 2 summaries the correlations reported through this paper. They are measure of fit, of the ability of the various models to predict the actual percentage of Asian pupils in London's neighbourhoods. To these, some additional measures have been added. The first is a weighted Pearson correlation. The logic of this is that the 'true' percentages per OA also contain uncertainty and this increases where the total number of pupils living in the OA is small. The correlation weights by the square root of the total pupil count.

The next column measures the gain in the targeting from using each model, measured as the percentage increase in the weighted correlation when compared to using the sample estimates alone. Using the geodemographic classification at the Group level returns a 67.4 per cent increase but the greater gain is when the cross-classified, multiscale model is used, at 92.4 per cent.

The next two columns ask 'if we wanted to target the top 25 per cent (the top quarter) of OAs with the highest percentages of Asian pupils, what would the error be?' There are two sources of error. The first is an error of omission, the percentage of OAs in the actual top quarter that would be missed if the top quarter from the model predictions were used. The second is an error of commission, the percentage of OAs from the model predictions that are erroneously taken to be amongst the top quarter. To avoid the small numbers problem, OAs with a total population count of ten or less are omitted. The best predictions are where both the errors of omission and commission are lowest, and that is for the multiscale model. The same is true if we wanted to target the top 10 per cent (top decile) or the top 5 per cent.

[TABLE 2 ABOUT HERE]

In summary, the table shows strong evidence to support the use of a geographically informed approach to geodemographic analysis. Of course, the results are specific to the case study and there is no necessary guarantee that the same finding would emerge with a different set of data. However, that does not change the arguments in favour of using a multilevel approach. The point is that the approach allows the variations at the various geodemographic and geographical scales to be explored and queried, and for the statistical significance of the neighbourhood types to be assessed. In our example there are geographical differences between localities as well as between neighbourhood types. In other examples that may not be so but in any case it is worth checking for the potential gains in targeting that may arise from doing so.

Conclusion

In this paper we have discussed multilevel modelling as a statistical framework for geodemographic analysis. We have shown how it can be used to produce standard the index values familiar to users of geodemographics, to model at multiple levels of a geodemographic and geographic hierarchy simultaneously, to address issues of statistical confidence and to handle some of the uncertainty associated with the data. There are parallels in this work to research looking at patterns of segregation at multiple scales (Johnston et al., 2005; Manley et al., 2015).

The models we have used here are comparatively simple. They are multilevel models with random intercepts. The models may be extended to include predictor variables at any one or more of the model's levels, the effects of which may be either fixed or allowed to vary (as random slopes) between neighbourhood types or from place-to-place. For online training in multilevel modelling, see http://www.bristol.ac.uk/cmm/learning/online-course/.

Our key argument is that the hierarchical structure of geodemographic classifications makes them a natural choice for multilevel analysis and that doing so provides opportunity to insert more geography into geodemographics, with improvements in the accuracy of the neighbourhood targeting as a potential result.

Acknowledgments

An initial version of this paper was presented at a seminar in Ritsumeikan University, Kyoto, and at the Mosaic users conference, Tokyo, in February 2016. RH is especially grateful to Professors Yano and Nakaya for their invitation and generous hospitality.

References

Birkin, M. (1995) Customer targeting, geodemographics and lifestyle approaches. In: P. Longley and G. Clarke (eds.) *GIS for Business and Service Planning*. Cambridge: GeoInformation International, pp. 104–49.

Booth, C. (1888) Condition and occupation of the people of East London and Hackney, 1887. *Journal of the Royal Statistical Society* 51(2), 276–339.

Booth, C. (1902–3) *Life and Labour of the People in London* (3rd edition). London: Macmillan.

Burrows, R., Ellison, D. and Woods, B. (2005) *Neighbourhoods on the Net: The Nature and Impact of Internet-based Neighbourhood Information Systems*. Bristol: Policy Press.

Charlton, M., Openshaw, S. and Wymer, C. (1985) Some new classifications of census enumeration districts in Britain: a poor man's ACORN. *Journal of Economic and Social Measurement*, 13(1), 69–96.

- See more at: http://www.popline.org/node/419597#sthash.O3SGT8Pz.dpuf

Curry, M. (1998) *Digital Places: Living with Geographic Information Technologies*. London: Routledge.

Dalton, C.M. and Thatcher, J. (2015) Inflated granularity: Spatial "Big Data" and geodemographics. *Big Data & Society*, July–December 2015, 1–15

Furness, P. (2008) Real time geodemographics: New services and business opportunities (and risks) from analysing people in time and space. Journal of Direct Data and Digital Marketing Practice, 10(2), 104–115.

Goodchild, M.F. (2004) The Validity and Usefulness of Laws in Geographic Information Science and Geography. *Annals of the Association of American Geographers,* 94(2), 300–303.

Goss, J. (1995) Marketing the New Marketing: The Strategic Discourse of Geodemographic Information Systems. In: Pickles, J. (ed.) *Ground Truth: the Social Implications of Geographic Information Systems*. New York: The Guilford Press, pp. 130–170.

Harris, R. (2002) The Diversity of Diversity: Is There Still a Place for Small Area Classifications? *Area*, 33(3), 329–336.

Harris, R. (2017) Local statistics and place-based analysis. In: Richardson, D., Castree, N., Goodchild, M. M., Kobayashi, A., Liu, W. and Marston, R. A. (eds.) *The*

International Encyclopedia of Geography. New York: John Wiley & Sons, forthcoming.

Finch, W.F., Bolin, J.E., Kelley, K. (2014) Multilevel Modeling Using R. Boca Raton, FL: CRC Press.

Harris, R., Johnston, R. and Burgess, S. (2007) Neighborhoods, Ethnicity and School Choice: Developing a Statistical Framework for Geodemographic Analysis. *Population Research and Policy Review*, 26(5), 553–579.

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

Kestle, J. (2011) How Geodemographics Has Changed Over the Past 40 Years: Part One - A History of Micromarketing. *Directions Magazine*, Aug 15, 2011. http://www.directionsmag.com/entry/how-geodemographics-has-changed-over-the-past-40-years-part-one-a-hist/194441 [Accessed April 5, 2016].

Johnston, R., Forrest, J., Jones, K. and Manley, D., 2016, The scale of segregation: ancestral groups in Sydney, 2011. *Urban Geography*, in press.

Kreft, I. and De Leeuw, J. (1998) Introducing Multilevel Modeling. London: Sage.

Leventhal, B. (1993) Birds of a feather? Or, geodemographics – an endangered species? *Proceedings of the Market Research Society Conference*, pp. 223–39.

Leventhal, B. (2016) Geodemographics for Marketers: Using Location Analysis for Research and Marketing. London: Kogan Page.

Longley, P. (2005) Geographical Information Systems: a renaissance of geodemographics for public service delivery. *Progress in Human Geography*, 29(1), 57–63.

Manley, D., Johnston, R., Jones, K. and Owen, D., 2015, Occupational segregation in London: A multilevel framework for modelling segregation. In: Socio-

Economic Segregation in European Capital Cities: East Meets West. London: Taylor and Francis, pp. 30–54.

Nnoaham. K.E., Frater, A., Roderick, P., Moon, G. and Halloran, S., 2010. Do geodemographic typologies explain variations in uptake in colorectal cancer screening? An assessment using routine screening data in the south of England. *Journal of Public Health*, 32(4), 572–581.

Robson, K. and Pevalin (2015) *Multilevel Modeling in Plain Language*. London: Sage.

Singleton, A., Pavlis, M. and Longley, P.A. (2016) The stability of geodemographic cluster assignments over an intercensal period, in press.

Singleton, A.D. and Spielman, S.E. (2014) The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom. *The Professional Geographer*, 66(4), 558–567.

Siu, D.Z. (2004) Tobler's First Law of Geography: A Big Idea for a Small World? *Annals of the Association of American Geographers*, 94(2), 269–277.

Sleight, P. (2004) *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business*. Henley-on-Thames: World Advertising Research Center.

Snijders, T.A.B. and Bosker, R.J. (2011). *Multilevel Analysis: An Introduction To Basic And Advanced Multilevel Modeling* (2nd edition). London: Sage.

Tobler, W. (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46 (supplement, June), 234–240.

Troy, A. (2008) Geodemographic Segmentation. In: Shekhar, S. and Xiong, H. (eds.) Encylopedia of GIS. New York: Springer, pp. 347–355.

Vickers, D.W. and Rees, P.H. (2006) Introducing the National Classification of Census Output Areas, *Population Trends*, 125.

Vickers, D.W. 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.

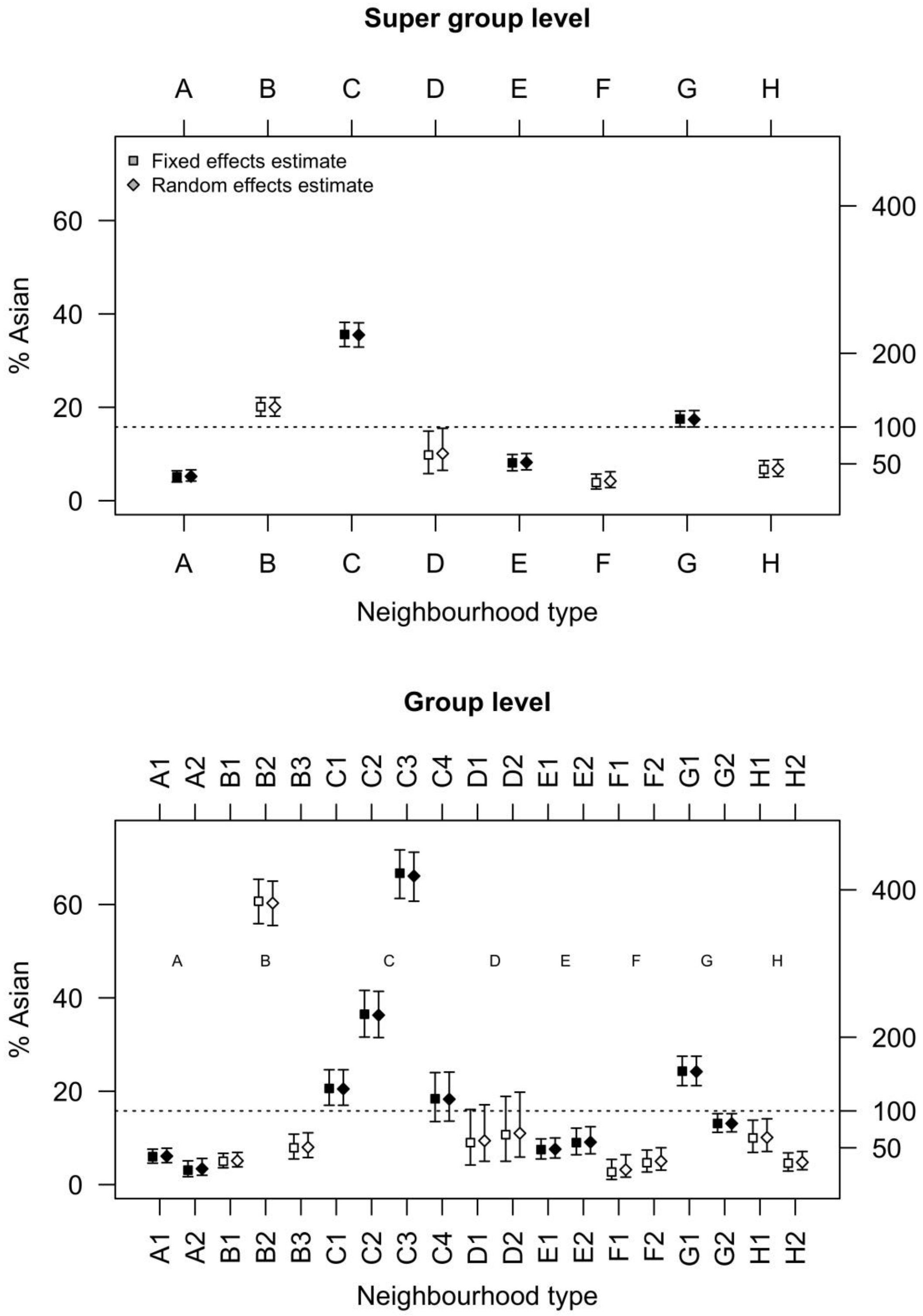
Voas D. and Williamson, P. (2002) The diversity of diversity: a critique of geodemographic classification. *Area* 33(1), 63–76.

Super group	Name	Asians	All pupils	% Asian	Index score
A	Intermediate Lifestyles	68	1326	5.1	32
B	High Density and High Rise Flats	314	1566	20.1	127
C	Settled Asians	466	1310	35.6	225
D	Urban Elites	16	164	9.8	62
E	City Vibe	76	942	8.1	51
F	London Life-Cycle	21	541	3.9	25
G	Multi-Ethnic Suburbs	318	1822	17.5	111
H	Ageing City Fringe	50	749	6.7	42
	(All)	1329	8420	15.8	100
Group					
A1	Struggling suburbs	56	937	6.0	38
A2	Suburban localities	12	389	3.1	20
B1	Disadvantaged diaspora	38	759	5.0	32
B2	Bangladeshi enclaves	244	402	60.7	385
B3	Students and minority mix	32	405	7.9	50
C1	Asian owner occupiers	89	432	20.6	131
C2	Transport service workers	131	359	36.5	231
C3	East End Asians	208	312	66.7	422
C4	Elderly Asians	38	207	18.4	116
D1	Educational advantage	8	89	9.0	57
D2	City central	8	75	10.7	68
E1	City and student fringe	42	563	7.5	47
E2	Graduation occupation	34	379	9.0	57
F1	City enclaves	6	222	2.7	17
F2	Affluent suburbs	15	319	4.7	30
G1	Affordable transitions	172	708	24.3	154
G2	Public sector and service employees	146	1114	13.1	83
H1	Detached retirement	29	291	10.0	63
H2	Not quite Home Counties	21	458	4.6	29
	(All)	1329	8420	15.8	100

Table 1. The percentage of the sample that is Asian in each neighbourhood type and the resulting index scores.

	Correlations				Target OAs: Top 25%		Target OAs: Top 10%		Target OAs: Top 5%	
Model	Spearman's	Pearson	Weighted	Gain (%)	Omission	Commission	Omission	Commission	Omission	Commission
Sample estimates	0.118	0.370	0.420	-	20.5	71.3	12.0	87.3	62.8	63.7
Fixed effects model - Super group level	0.433	0.458	0.479	14.0	42.4	54.3	50.9	68.5	52.7	84.8
Fixed effects model - Group level	0.488	0.647	0.703	67.4	35.1	35.8	38.1	40.2	39.8	50.8
Random intercepts model - Super group level	0.433	0.458	0.479	14.0	42.4	54.3	50.9	68.5	52.7	84.8
Random intercepts modes model - Group level	0.488	0.647	0.703	67.4	35.1	35.8	38.1	60.2	39.8	50.8
Random intercepts - all geodemographic levels	0.498	0.661	0.718	71.0	34.6	35.0	37.8	39.2	44.8	47.4
Random intercepts - cross-classified, multiscale model	0.564	0.743	0.808	92.4	27.6	27.5	28.2	28.1	39.2	39.2

Table 2. Measures of fit comparing the predicted values with the actual percentages of Asian pupils per London neighbourhood. See text for detail.



Super group level

Group level

Figure 1. Calculating the percentage of the sample that is Asian in each neighbourhood type using fixed and random effects regression models. A 95 per

cent confidence interval is shown around each estimate and the equivalent index values are shown on the right-hand axis of the charts.

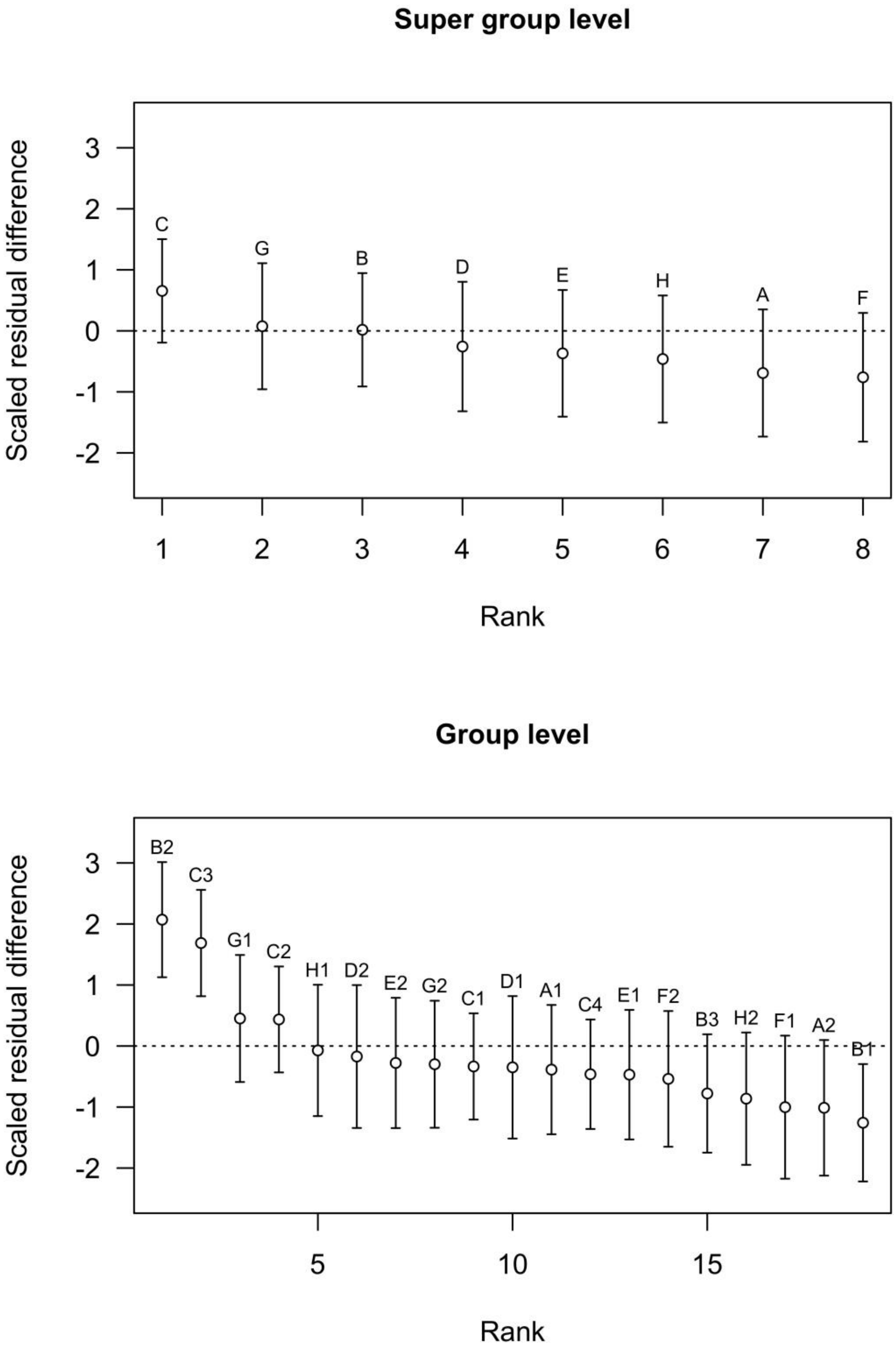


Figure 2. Caterpillar plots indicating the differences between neighbourhood types at each level of the geodemographic hierarchy net of the differences due to other levels of the hierarchy.

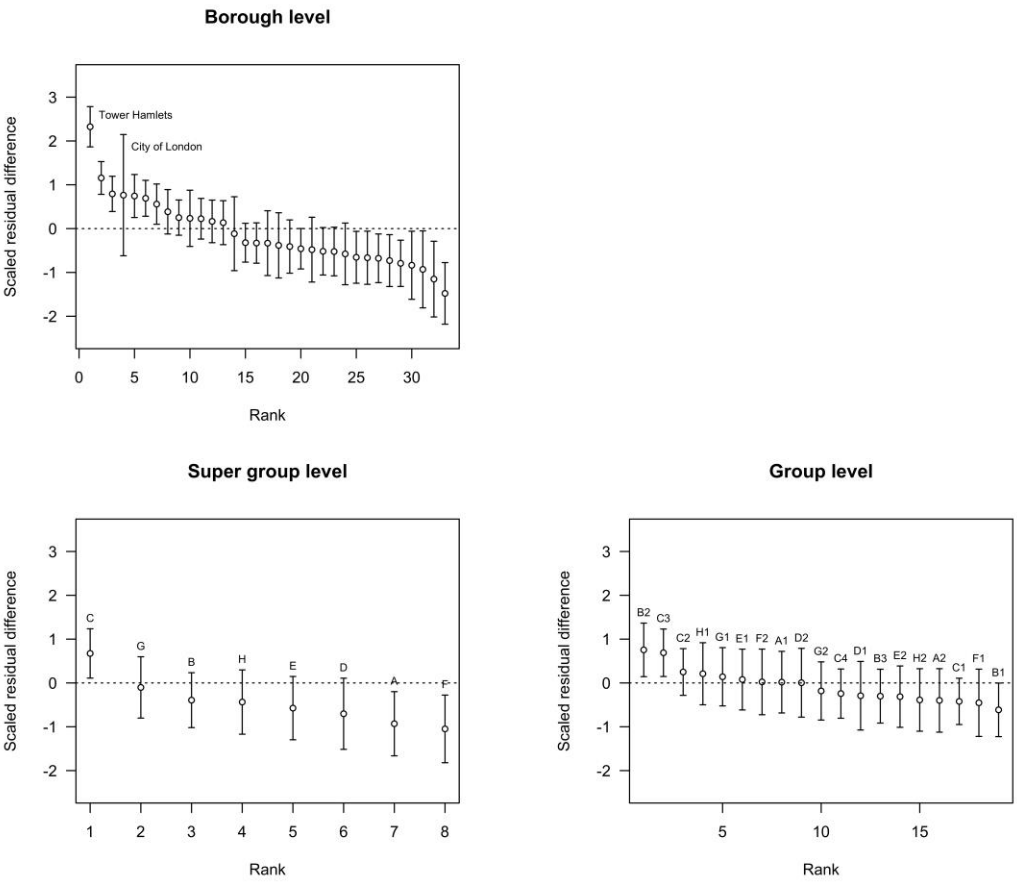


Figure 3. Caterpillar plots indicating the differences between London boroughs and the geodemographic neighbourhood types arising from the multiscale model.