Journal of the Royal Statistical Society
SERIES A Statistics in Society

*J. R. Statist. Soc.* A (2011)
**174**, *Part* 1, *pp.* 17–30

# Predicting participation in higher education: a comparative evaluation of the performance of geodemographic classifications

Chris Brunsdon,

*University of Leicester, UK*

Paul Longley and Alex Singleton

*University College London, UK*

and David Ashby

*Dr Foster Research, London, UK*

[Received July 2008. Final revision January 2010]

**Summary.** Participation in UK higher education is modelled by using Poisson regression techniques. Models using geodemographic classifications of neighbourhoods of varying levels of detail are compared with those using variables that are directly derived from the census, using a cross-validation approach. Increasing the detail of geodemographic classifiers appears to be justified in general, although the degree of improvement becomes more marginal as the level of detail is increased. The census variable approach performs comparably, although it is argued that this depends heavily on an appropriate choice of predictors. The paper concludes by discussing these results in a broader practice-oriented and pedagogic context.

*Keywords*: Geodemographics; Higher education; Participation; Poisson regression; Postcodes

## 1. Participation in UK higher education

The UK higher education (HE) sector has come under increasing pressure from central government to extend access to prospective students who are identified as coming from under-represented socio-economic groups. Such applicants tend to originate in neighbourhoods with shared characteristics. To meet these participation targets, HE institutions are encouraged to extend access across several performance indicators including a measure of low neighbourhood participation, where neighbourhoods are classified by using a 'geodemographic' classification (the SuperProfiles system: Brown and Batey (1994)). Within the UK, these access targets became of increasing financial importance to English institutions following the 2004 Higher Education Act which granted institutions the right to charge variable tuition fees of up to £3000, as long as the Office for Fair Access considered that the institution was ensuring that its courses of HE were made available to all. In terms of public accountability and also wider issues of social justice it is of core importance for institutions to be seen to be open, and for them to have confidence in their

*Address for correspondence*: Chris Brunsdon, Department of Geography, University of Leicester, University Road, Leicester, LE1 7RH, UK.
E-mail: cb179@le.ac.uk

 0964–1998/11/174017

methods of targeting. Targeting methods that are inaccurate, imprecise or otherwise error prone can have negative effects on the life chances of prospective students and, where neighbourhood profiling techniques are recommended as an effective method, assessment of the relative merits and performances of the different classifications that may be used is essential. In this context, this paper will evaluate a range of neighbourhood classifications as tools to provide institutions with evidence on which to base their selection of targeting methods.

Neighbourhood classification can be achieved by using a range of univariate or composite indicators. Within the UK education domain, for example, the percentage of pupils receiving free school meals provides a simple but robust indicator of relative child poverty, though it is only a blunt-edged indicator of more broadly based inequalities in educational opportunity. UK local area government policy interventions are more usually predicated on the index of multiple deprivation, i.e. a composite indicator representing seven domains of hardship, and which was last updated in 2007. Yet even such multifaceted indicators of physical and social conditions may fail to depict diversity of circumstances that characterize the full spectrum of social, economic and demographic conditions. Within the private sector, the response to this has been that almost every sizable organization with customers has adopted 'geodemographic' classifications to portray the richness and diversity of small area conditions.

As such, 'geodemographics' are often described as the analysis of people by where they live (Sleight, 1997). Geodemographic classifications allocate each locality to one of a set of discrete classes which is deemed to represent the defining characteristics of the people living within the area. Although the origin of these classifications lies in analysis of neighbourhood deprivation in public sector applications (Harris *et al.*, 2005), geodemographics are best known as a commercial method for profiling and small area targeting of potential customers (Birkin, 1995; Birkin *et al.*, 2002). However, there is increasing interest in the application of geodemographic classifications as a method of improving public sector service delivery. This 'renaissance of geodemographics' (Longley, 2005) has taken place, *inter alia*, in applications in policing (see Ashby and Longley (2005) and Ashby (2005)), health (see Aveyard *et al.* (2002) and Farr and Evans (2005)) and education (see Batey *et al.* (1999), Tonks and Farr (2003) and Singleton and Longley (2009)).

Repeat purchasing of commercial geodemographic solutions provides some evidence of their value in private sector applications, yet there has been only limited formal quantitative evaluation of the absolute and relative performance of different classifications (but see Voas and Williamson (2001a, b) and Webber (2005)). This paper sets out to develop a more formal evaluation of two geodemographic classifications, in the important applications domain of participation in HE.

In this paper, the framework for the analysis of the data will be set out in the next section. Following this, issues surrounding the data that are used will be considered. Next some statistical models making use of two geodemographic classification schemes as well as 'raw' census data will be set out, and in Section 5 the performance of the models as predictors of participation will be compared. The paper ends with conclusions and recommendations based on the results.

## 2. The analysis

For this analysis HE participation is defined as the proportion of 18–19-year-old students living in an area who are engaged in full-time HE. Two geodemographic classifications are compared for their ability to predict these rates, and these results are also benchmarked against the performance of a standard Poisson regression model, the details of which will be introduced

**Table 1.** Classification levels used in six different UK geodemographic systems (adapted from Vickers and Rees (2006))

| *Classification system* | *Clusters in level 1* | *Clusters in level 2* | *Clusters in level 3* |
|---|---|---|---|
| Mosaic 2001 (Experian, Nottingham) | 11 | — | 61 |
| Cameo (EuroDirect, Leeds) | 10 | — | 58 |
| ACORN (CACI, London) | 5 | 18 | 57 |
| PRiZM (Claritas UK, Middlesex) | — | 16 | 60 |
| SuperProfiles (Batey and Brown, 1994) | 10 | 40 | 160 |
| Output area classification (Vickers and Rees, 2007) | 7 | 21 | 52 |

in Section 4. As implied above, the construction of geodemographic indicators entails multidimensional clustering or profiling of socio-economic data. Solutions are available at different spatial resolutions, but most of the available options share the kind of hierarchical classification structure that is shown for six different geodemographic systems in Table 1. These are principally commercial systems, with the Office for National Statistics output area classification (OAC) being the exception, and have been developed by companies that are based in the UK (sometimes as subsidiaries of US parents). Longley and Goodchild (2008) have described some systems that have been developed for the US market.

In this evaluation the performance of the following geodemographic classifications, and their respective levels of disaggregation (measured in terms of numbers of constituent classes, $n$), will be assessed:

(a) OAC supergroups ($n = 7$);
(b) OAC groups ($n = 21$);
(c) OAC subgroups ($n = 52$);
(d) Mosaic groups ($n = 11$);
(e) Mosaic types ($n = 61$).

The OAC is a public domain, open source, free-to-access classification that was created as part of a collaboration between the Office for National Statistics and the University of Leeds (Vickers and Rees, 2007). It has a wide user base in the university and public sectors, and an active user group (`http://www.areaclassification.org.uk/`). Mosaic has an established pedigree as a private sector geodemographic system and is probably the most popular in the UK in terms of market share. Comparison of classification outcomes at the finest level of spatial granularity is to some extent compromised, in that the OAC is disseminated by using the census of population output area (OA) geography, whereas Mosaic is available for postcode units. An OA typically comprises several postcode units, grouped together by the Office for National Statistics by using algorithms that were designed to maintain within-area homogeneity of social and built environment conditions. For our comparison, geodemographic indicators pertaining to areal units that are coarser than postcode units are referenced to individuals by geocoding postcode units to population-weighted centroids. A population-weighted centroid is essentially a single point that is used to represent a postcode unit, chosen to be close to the locations of residents in that postcode. These centroids are provided, for example, by the Ordnance Survey in their Code-point product, or the *National Statistics Postcode Directory* (Office for National Statistics, 2009) file that is discussed in the next section. Therefore, because OACs are based

on census OAs whereas Mosaic is based on postcode units, the OAC-based analysis participation rates will be computed for OAs and for Mosaic they will be computed for postcode unit areas.

## 3. Data issues

The HE data that are used in this evaluation were provided by the Higher Education Statistics Agency (HESA) and cover all students of English domicile studying within English institutions in the year 2001. This data set is a subset of the data that were gathered from HE institutions by the HESA as part of an 'HESA return' which includes various characteristics of the students studying at various institutions. These data are primarily used by central policy makers in planning and monitoring of the sector and are also used to devise key indicators of performance in widening participation and to access equality. HE institutions initially collect data on their admitted students via the Universities and Colleges Admissions Service application cycle, where electronic records of each applicant are transferred as part of the offer and acceptance process, and this in turn populates internal acceptance databases or student records. Institutions are encouraged by the HESA to update these details further and to amend erroneous, missing or incomplete attributes. There are, of course, limitations to what can be achieved by using these statistics: in particular, they contain no information on parental education, so it is not possible to gain any perspective on intergenerational changes in inequalities of access to HE. For this evaluation only university participation of 18–19-year-old students will be analysed, although the HESA data are an important resource for a range of related analyses. The data source for OA-based counts of 18–19-year-old students is the 2001 UK census. The numbers of 18–19-year-old students in HE may be derived from the HESA file as each entry is supplied with a postcode, and the *National Statistics Postcode Directory* file may be used to associate a postcode with an OA. Thus, to compute the proportions of 18–19-year-old students in HE the numerator is derived from the HESA, whereas the denominator comes from the census. The fact that the two figures come from different sources can cause some problems. In particular, in areas of very high participation it is possible for the numerator to exceed the denominator—apparently giving a participation rate of over 100%! This can be caused in various ways. Firstly, given that the two data sets were not based on exactly the same date, it is possible that new 18–19-year-old students moved into (or out of) an area between surveys. Secondly, census undercounting is possible and, finally, the practice of some students of registering at their term-time address for one survey, but from their parental home for the other, may be another cause.

For the Mosaic analysis, the core spatial unit is the postcode unit, which is typically more geographically precise than the 2001 census OAs. However, the two sets of areal units do not nest exactly, so some postcode units straddle the boundaries of the OAs. Here, we need to impute the denominators, as the census does not release counts at the postcode unit level. This is done with the aid of two variables in the *National Statistics Postcode Directory* file—the OA within which a postcode unit is located, and the residential address count for that postcode unit. For a given OA, let $U_{\mathrm{OA}}$ be the set of all postcodes within that OA, let $N_{\mathrm{OA}}$ be the count of 18–19-year-old students in the OA and let $A_i$ be the residential address count in postcode unit $i$. Then, $N_i$, the estimated count of 18–19-year-old students in postcode unit $i$, is given by

$$N_i = N_{\mathrm{OA}} A_i \Big/ \sum_{j \in U_{\mathrm{OA}}} A_j. \tag{1}$$

Thus, the 18–19-year-olds count for each postcode in an OA is apportioned on the basis of residential address counts. Clearly this is an approximate denominator—if for example all 18–19-year-old students clustered in one or two postcode units within an OA then the estimates would be quite misleading. This compounds the problems that were outlined above—not only are the numerator and denominator drawn from different sources, but also the denominator is a fairly crude estimate. This perhaps suggests that the Mosaic-based approaches have an intrinsic disadvantage in the models that are used here.

At this stage it is perhaps important to justify the choice of spatial units in the two analyses. For the OAC-based analysis the choice of units is straightforward, as the classification pertains to OAs and it is straightforward to assign the HESA entries to OAs. For Mosaic there are perhaps two options.

(a) Work with OAs and attempt to classify the OA by using a single Mosiac category—perhaps by assigning that category to every postcode unit that is contained in the OA. In this way, both analyses will be carried out using the same set of spatial units.
(b) Work at postcode level, but estimate the residential address count $A_i$.

We have chosen the second option—essentially because we feel that, when one adopts a neighbourhood classification system, one adopts that system's geographical demarcation of neighbourhoods as well as the typology. Thus, the OAC works on the basis of data aggregated to OAs, whereas Mosaic assumes that postcode units are the most appropriate elemental neighbourhood units. Our analyses will remain consistent with these respective frameworks.

## 4. Statistical models

The counts of 18–19-year-old students in full-time HE will be modelled here as Poisson distributions. For spatial unit $i$, the mean of the Poisson distribution will be $m_i$, modelled as

$$m_i = N_i p_c \tag{2}$$

where $N_i$ is the count of 18–19-year-old students in spatial unit $i$ and $p_c$ is the rate of participation for geodemographic class $c$, where spatial unit $i$ is classified as class $c$. For the three OAC-based analyses the set of $c$ will be either supergroups, groups or subgroups, and the spatial units are OAs. For Mosaic, the set of $c$ will either be the Mosaic group or type, and the spatial units will be postcode units. From equation (2), and supposing that the count of 18–19-year-old students in HE in spatial unit $i$ is $u_i$, the log-likelihood of the observed participation counts is

$$\sum_i -N_i p_c + u_i \log(p_c) + \text{constant} \tag{3}$$

where the constant term does not depend on any of the $p_c$s. It is straightforward to show that this is maximized when

$$\hat{p}_c = \sum_{i \in G_c} U_i \Big/ \sum_{i \in G_c} N_i \tag{4}$$

where $G_c$ is the set of $i$s indexing the spatial units that are classified in class $c$. In this and what follows, the 'hat' notation appearing on the $\hat{p}_c$s indicates that these are maximum likelihood estimates of true values.

Plotting the $\hat{p}_c$s against their respective classifications allows model interpretation—see Figs 1, 2 and 3 for OAC supergroups, groups and subgroups respectively. Here the rectangles indicate rates of participation arranged from high to low and for each the shading shows the OAC supergroup to which lower level grouping belongs. As can be seen, the highest participation occurs in neighbourhoods from supergroups 3 ('countryside') and 4 ('prospering suburbs')—the lowest

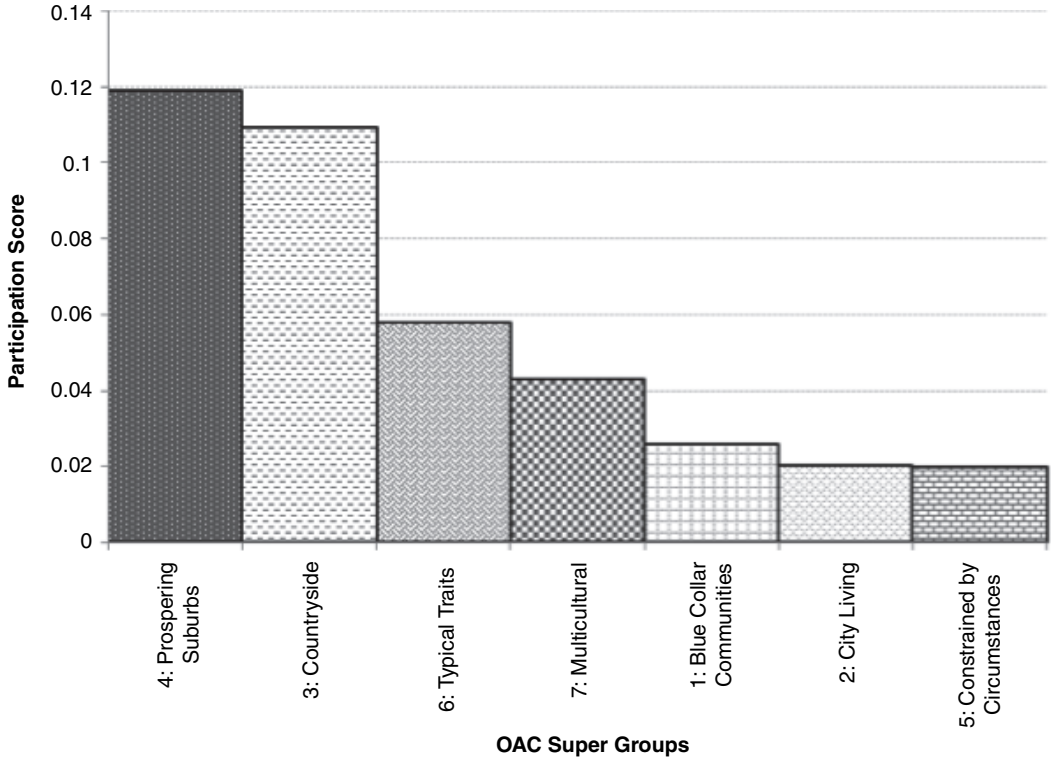


**Fig. 1.** Participation index by OAC supergroup

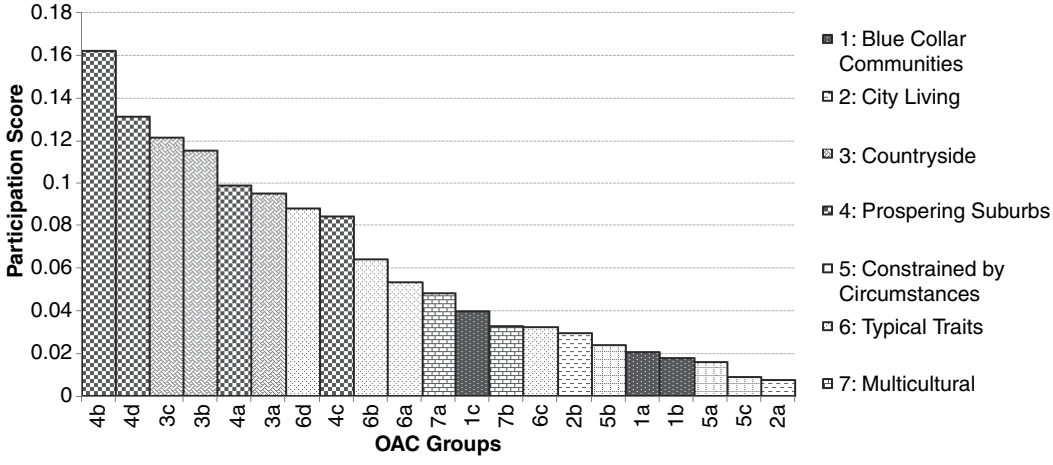


**Fig. 2.** Participation index by OAC group

in supergroup 5 ('constrained by circumstances'), suggesting that concerns that participation in HE is lower in less affluent and urban neighbourhoods are well founded. Fitting Poisson models by using maximum likelihood estimation allows geodemographic-based analyses to be statistically evaluated: in particular, testing the fit of subgroups, groups and supergroups is achievable by using *analysis of deviance*. This is possible as each increase in level of detail (i.e. supergroups

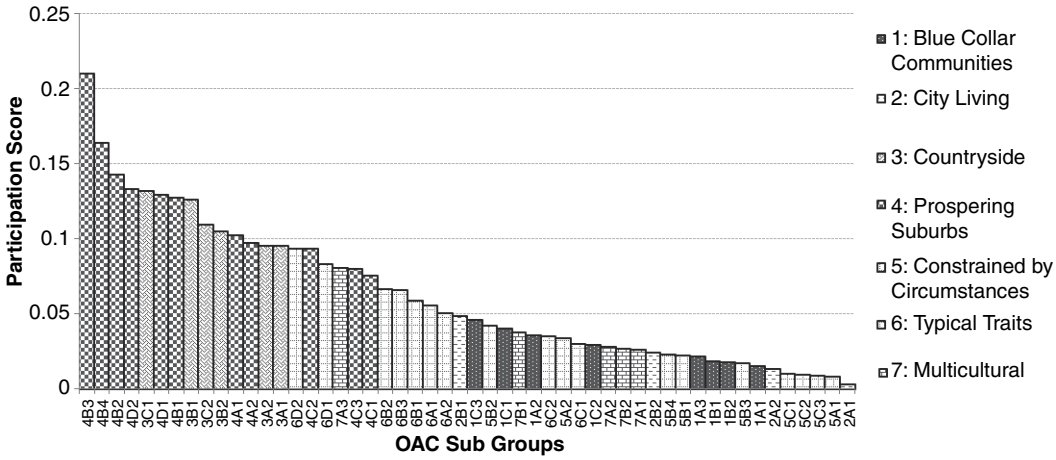


**Fig. 3.** Participation index by OAC subgroup

**Table 2.** Analysis of deviance for the OAC models

| *OAC model* | *Residual degrees of freedom* | *Deviance from null model* | *Model degrees of freedom* | *Deviance* | *p-value* |
|---|---|---|---|---|---|
| Supergroup | 152659 | 200423 | | | |
| Group | 152645 | 194875 | 14 | 5549 | $< 0.001$ |
| Subgroup | 152614 | 191777 | 31 | 3098 | $< 0.001$ |

to groups and groups to subgroups) is linked to a nested hypothesis test—each level of detail is a hierarchical subdivision of the previous level. Here we see that there is a justification for using subgroups—an analysis of deviance is given in Table 2. The second column is equal to the number of OA observations minus the number of parameters fitted (the number of $p_c$s for each level of classification). The subtracted term is the value of the third column, and the fifth column is the difference in deviance between the null model with no predictors except an overall mean level for $u_I$ and the fitted model corresponding to the appropriate classification level. This value is the log-likelihood ratio between the null model and the latter model, which has an asymptotic $\chi^2$-distribution with degrees of freedom as stated in the fourth column (see for example Cox and Hinkley (1974)).

In each case the decrease in model deviance when moving to a more detailed classification is statistically significant.

A further approach to assessing the predictive ability of each of the models is to use *holdback cross-validation*. In this approach a random subset (say 10%) of the $u_i$s is 'held back', and the remaining 90% of observations are used to calibrate the model. Using the calibrated model, predicted values for the holdback samples are computed by using

$$\hat{u}_i = \hat{p}_c N_i \tag{5}$$

where the $c$ of $\hat{p}_c$ refers to the classification $c$ of area $i$. To assess the predictive ability of the model, the predicted $u_i$s for the holdback sample are compared with the actual values, by using the *mean absolute deviation* MAD:

**Table 3.** Holdback cross-validation results (OAC)

| *Model* | *MAD* |
|---|---|
| Simple | 0.678 |
| Supergroup | 0.626 |
| Group | 0.618 |
| Subgroup | 0.612 |

$$n_H^{-1} \sum_{i \in H} |\hat{u}_i - u_i| \tag{6}$$

where $H$ is the set of $i$s indexing the holdback sample and $n_H$ is the size of this sample. This measures the mean absolute difference between the predicted and actual numbers of university entrants in each OA. This allows the predictive performance of the model to be assessed independently of the model calibration, and so problems of overfitting are greatly reduced.

## 5. Results of analysis

The results of the analysis that was described above are set out in the following Sections 5.1–5.3.

### 5.1. Cross-validation results

For the OAC classification the holdback cross-validation results for each level are shown in Table 3. The MAD figures for subgroups, groups and supergroups are listed together with the figure for the 'simple' model, where all OAs are assigned to the same class—in other words for the simple model we ignore geodemographics and predict everywhere by using the national average participation rate.

From these results it is clear that the findings from the nested hypothesis tests are borne out. It is still the case that going to the most detailed classification is justified. However, an extra insight is gained from this analysis—although we can see that increasing the detail of the classifier improves performance, the marginal pay-off reduces with the sophistication of the classifier. The biggest improvement occurs when moving from the simple model to the supergroup classification—and the smallest when going from groups to subgroups. This is consistent with Figs 1–3, where the shading of the bars shows that, for the group and subgroup analyses, it is still the case that there is a clear separation of the supergroups—with 3 and 4 dominating the high participation categories, and 5 at the low end of the scale.

### 5.2. Mosaic results

The model that was outlined in equation (2) was also fitted to the Mosaic neighbourhood classification, although in this case the spatial units were postcode units, and the $N_i$s were *estimated* counts of 18–19-year-old students as set out above. The estimated participation rates for Mosaic 'groups' (11 classes) and 'types' (61 classes) are graphed in Fig. 4 and Fig. 5 respectively.

These results tell a similar story to the OAC results—particularly for the groups (which are similar in scope to the OAC supergroups). From these, the two groups with the highest classi-

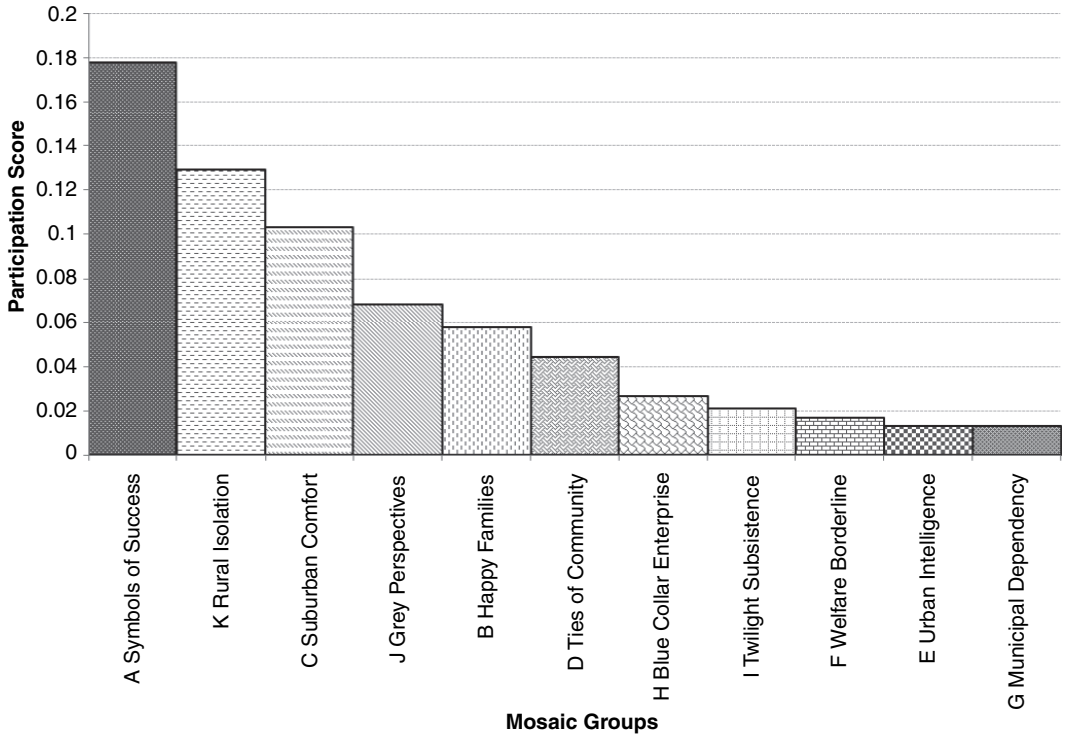


**Fig. 4.** Participation index by Mosaic group

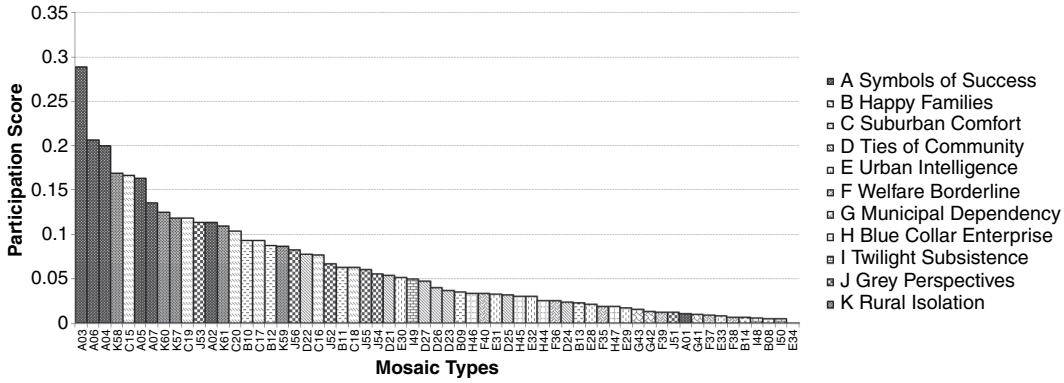


**Fig. 5.** Participation index by Mosaic type

fication are A ('Career professionals living in sought after locations') and K ('People living in rural areas far from urbanisation'). These categories are quite similar in description to the two highest participation OAC supergroups. The lowest group is G ('Low income families living in estate based social housing') and this result is also similar to the OAC situation. As before, it is possible to assess the predictive ability of these models by using cross-validation—although with the Mosaic approach it is postcode units rather than census OAs that are 'held back', and the mean absolute deviation is computed in terms of predictions at the postcode unit level. Results are shown in Table 4. Since postcode units are generally smaller than OAs, predicted

**Table 4.** Holdback cross-validation results (Mosaic)

| *Model* | *MAD* |
|---|---|
| Simple | 0.116 |
| Group | 0.111 |
| Type | 0.109 |

and actual values are lower and so the MADs are correspondingly reduced. Thus, unfortunately these numbers cannot be directly compared with those for the OACs.

As before, it can be seen that, although increasing the detail of the classifier does improve predicted performance, once again the marginal improvement decreases as detail increases. Also, the relative improvement on the simple model based on national average participation is smaller than with the OACs. This is perhaps because the estimation of base 18–19-year-old populations introduces an extra element of certainty in the prediction process, which to some extent offsets the predictive benefit of Mosaic groups or types.

Of note in Figs 4 and 5 are the relatively low participation rates in group E ('Urban intelligence') and within this category Mosaic types E33 ('Town Gown transition') and E34 ('University Challenge'). As the labels suggest, these neighbourhood categories characteristically contain a very high proportion of students and are typically student residential areas which are very close to universities. Although it may initially appear counterintuitive to see very low participation rates in those neighbourhoods which are dominated by students, this is logically explained by the data and approach that is used here: although these areas have a very high dependence on and active participation at local universities, the predominant characteristics of the population dictate that very few residents will apply to university courses and hence register on the HESA statistics from these addresses. Similarly, relatively low participation rates can be found within the type A01 ('Global connections') (Fig. 5), which is inconsistent with the overarching trend for the aggregate group A (Fig. 4). Again, such trends are explained by the underlying population composition and typical traits—global connection areas predominantly contain well-educated, extremely wealthy individuals largely living in central London. There is an over-representation of single-person households in these areas, with such individuals being at a life stage that is unlikely to be conducive to university applications.

### 5.3. Comparison with Poisson regression

Finally, the OAC results at the OA level were compared with a Poisson regression model. Here, the participation rate (at OA level) is modelled as a function of a number of variables that were derived from the 2001 census, also at OA level. A Poisson model takes the form

$$\mu_i = N_i \exp(\beta_0 + \beta_1 v_{1i} + \ldots + \beta_m v_{mi}) \tag{7}$$

where $\mu_i$ is the number of participants in HE for area $i$, $N_i$ is the underlying population of 18–19-year-old students in area $i$, $m$ is the number of predictor variables $\{v_{1i} \ldots v_{mi}\}$ for OA $i$, $\{\beta_0, \ldots, \beta_m\}$ are a number of regression coefficients to be estimated and $N_i$ is the population at risk in OA $i$. A model of this form can be calibrated by using the `glm` function in R (R Development Core Team, 2008). Thus, in this model, the simple geodemographic class-based participation rate of equation (2) is replaced by a mathematical function of the census-based predictor variables. Here, six variables were used—these are listed in Table 5.

**Table 5.** Variables in the Poisson regression model

| *Coefficient* | *Name* | *Description* |
|---|---|---|
| $b_1$ | DENSITY | Population density (people per square metre) |
| $b_2$ | OVER65 | Proportion of population aged 65 years or over |
| $b_3$ | UNDER30 | Proportion of population aged below 30 years |
| $b_4$ | OWNOCC | Proportion of households that are owner occupied |
| $b_5$ | UNEMP | Proportion of economically active population who are unemployed |
| $b_6$ | EDUC | Proportion of population aged 16–74 years with highest level of educational qualification |

It is perhaps only reasonable to declare that the choice of these variables was guided by the outcome of the earlier geodemographic analysis. In particular OWNOCC and UNEMP were chosen to represent the affluence–deprivation dimension, and DENSITY was chosen to represent the urban–rural dimension that was highlighted by both the OAC and the Mosaic analyses. The coefficient estimates for the model are given in Table 6.

All these are negative (suggesting that an increase in the predictor variable reduces participation) except for OWNOCC and EDUC—an increase in owner occupation suggests a greater rate of participation, as does a higher local level of educational attainment in the OA.

Since this model is applied at the OA level, it is possible to compare its predictive ability with the three OAC-based models using the cross-validation approach. Using the same 10% holdback sample as was used for the OAC-based models, the Poisson regression achieved an MAD-score of 0.594, thus marginally outperforming the best of the OAC-based models.

This is perhaps no great surprise—any geodemographic approach groups spatial units (or individuals) into broad categories—and within any category there will be some spatial units that are more peripheral than others, but no indication is given about which spatial units have this characteristic. However, this information is at least partially encapsulated in the individual variables of a regression model, which may explain the improvement in performance. It is also interesting that this improvement in prediction is not spread evenly across all geodemographic groupings. This is illustrated in Table 7, where the MAD-values are shown for each of the OAC supergroups, for the three OAC-based models and the Poisson regression model. It can be seen that, in groups 1, 2, 5 and 7 (respectively 'Blue collar workers', 'City living', 'Constrained by circumstances' and 'Multicultural') the geodemographic-based predictors are most successful. However, in the remaining groups ('Countryside', 'Prospering suburbs' and 'Typical traits') the

**Table 6.** Results of the Poisson regression model

| *Name* | *Estimate* | *Standard error* | *p-value* |
|---|---|---|---|
| INTERCEPT | −0.675 | 0.0517 | |
| DENSITY | −4.30 | 0.373 | <0.01 |
| OVER65 | −1.43 | 0.046 | <0.01 |
| UNDER30 | −3.98 | 0.046 | <0.01 |
| OWNOCC | 0.188 | 0.015 | <0.01 |
| UNEMP | −0.295 | 0.070 | <0.01 |
| EDUC | 1.76 | 0.021 | <0.01 |

**Table 7.** Cross-validation scores by OAC supergroup

| | *Results for the following OAC supergroups* | | | | | | |
|---|---|---|---|---|---|---|---|
| | *1, blue collar workers* | *2, city living* | *3, countryside* | *4, prospering suburbs* | *5, constrained by circumstances* | *6, typical traits* | *7, multicultural* |
| OAC supergroup | 0.323 | 0.449 | 0.842 | 1.019 | 0.225 | 0.576 | 0.583 |
| OAC group | 0.317 | 0.442 | 0.838 | 1.005 | 0.223 | 0.564 | 0.578 |
| OAC subgroup | 0.316 | 0.438 | 0.836 | 1.000 | 0.221 | 0.563 | 0.553 |
| Poisson regression | 0.323 | 0.516 | 0.798 | 0.940 | 0.229 | 0.577 | 0.491 |

improvement in performance of the Poisson regression is quite marked, suggesting perhaps that more subtle characteristics of these neighbourhoods are reflected in the choice of regression variables.

## 6. Conclusions and discussion

Comparing all results, it is worth noting that the improvement in performance for the regression model is fairly marginal—recall that, when considered as a variable, even the OAC subgroup can take only 22 values—but the regression variables can take unique values for every OA, and also that the improvement is not evenly spread across all geodemographic groups. Thus, the OAC analysis gives a creditable performance. Also recall that the choice of regression model variables was based on the results of the OAC supergroup analysis. There are thousands of census variables that could have been used in the Poisson regression, and success in prediction depends strongly on a good choice of predictors. Without prior knowledge of the geodemographic analyses, it is less likely that a good choice of variables would have been made, and in consequence it is unlikely that the model would have performed as well. Thus, one use of the geodemographic analysis might be as a first iteration, before variable selection in some other kind of model.

It is also worth noting that, although it is not discussed in detail here, the calibration of models like those in equation (2) is considerably simpler computationally than models of the form of equation (7). In fact the estimator of equation (4) is essentially just an average rate for each given geodemographic category—and, although it was introduced in an algebraic format here, could be easily explained without formal notation. Thus, not only is the computation simpler, but the calibration itself is also much clearer to users without advanced knowledge of statistics. Finally, the output of the geodemographic approaches is essentially a rate of participation for each category, rather than a mathematical function linking predictors to participation rate. Again, the former is more accessible to users without a strong mathematical background. Thus, at only a marginal cost in performance, the geodemographic classifier approach provides a more accessible and more efficient analysis.

There are, however, some broader issues that this analysis has only begun to broach. Descriptors of 'what is?' are undoubtedly useful in understanding the detailed geography of participation in HE, yet they provide at best a partial basis for understanding 'what if?' scenarios of policy change. There is an emerging view in parts of the geodemographics industry that predictive analysis of change may be best accomplished by using geodemographic classifications that have been

engineered for use in particular applications domains, such as education, health or policing. It may also be that the kinds of predictive models that were developed in equation (7) are best for scenario generation, notwithstanding the problems that we note as regards interpretability. The treatment of geographic scale is also important: although one of the issues that was broached here has been the importance of availability of data for postcode units as opposed to census OAs, there is evidence from education applications that discriminators should be engineered to deal with coarser zonal schemes (see, for example, Harris *et al.* (2007)). This also raises the issue of whether, and if so to what extent, the applicability of the results of this study are specific to the HE domain in the UK alone: Longley and Goodchild (2009) addressed this in the context of the application of geodemographic techniques in a range of public sector settings.

Finally, we suggest that this work has several implications for spatial literacy—the use of geodemographic analyses gives a clear picture of the variation between geographical socio-economic predictors and participation in HE. Since the approach is highly accessible in comparison with more advanced statistical approaches, this suggests that it is a good basis for 'demonstrator' projects to promote spatial thinking in experts in other fields. It is perhaps also a good pedagogical tool for geography undergraduates. In this spirit, Longley *et al.* (2005), pages 483–484, have undertaken a preliminary assessment of the 'geodemography of Geography', i.e. the Mosaic profile of students who find themselves studying geography in UK HE. They linked this discussion to a consideration of the discipline of geography as a socially constructed activity and identified some of the potential social divisions that characterize its future practitioners. In the UK such students have a wide variety of mathematical ability, and an approach with relatively straightforward calculations provides a strong entry level experience into the ideas underlying spatial data analysis and evidence-based policy analysis.

## Acknowledgements

This work was supported by Economic and Social Research Council grants RES-061-25-0303 and RES-149-25-1041 and the Higher Education Funding Council for England Centre of Excellence in Teaching and Learning ('Spatial literacy in teaching') at Leicester University and University College London.

## References

Ashby, D. I. (2005) Policing neighbourhoods: exploring the geographies of crime, policing and performance assessment. *Polcng Soc.*, **15**, 413–447.

Ashby, D. I. and Longley, P. A. (2005) Geocomputation, geodemographics and resource allocation for local policing. *Trans. GIS*, **9**, 53–72.

Aveyard, P., Manaseki, S. and Chambers, J. (2002) The relationship between mean birth weight and poverty using the Townsend deprivation score and the Super Profile classification system. *Publ. Hlth*, **116**, 308–314.

Batey, P. and Brown, P. J., (1994) Design and construction of geodemographic targeting systems. *J. Targetng Measmnt Anal. Marktng*, **3**, 105–115.

Batey, P., Brown, P. J. B. and Corver, M. (1999) Participation in higher education: a geodemographic perspective on the potential for further expansion in student numbers. *J. Geogr. Syst.*, **1**, 277–303.

Birkin, M. (1995) Customer targeting, geodemographic and lifestyles approaches. In *GIS for Business and Service Planning* (eds P. A. Longley and G. P. Clarke). Cambridge: GeoInformation.

Birkin, M., Clarke, G. and Clarke, M. (2002) *Retail Geography and Intelligent Network Planning*. Chichester: Wiley.

Brown, P. and Batey, P. (1994) Design and construction of a geodemographic targeting system: Super Profiles 1994. *Report WP-40*. Urban Research and Policy Evaluation Regional Research Laboratory, Department of Civic Design, University of Liverpool, Liverpool.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*, p. 92. London: Chapman and Hall.

Farr, M. and Evans, A. (2005) Identifying 'unknown diabetics' using geodemographics and social marketing. *Interact. Marktng*, **7**, 47–58.

Harris, R., Johnston, R. and Burgess, S. (2007) Neighborhoods, ethnicity and school choice: developing a statistical framework for geodemographic analysis. *Popln Res. Poly Rev.*, **26**, 553–579.
Harris, R., Sleight, P. and Webber, R. (2005) *Geodemographics, GIS and Neighbourhood Targeting*. Chichester: Wiley.
Longley, P. (2005) Geographical Information Systems: a renaissance of geodemographics for public service delivery. *Prog. Hum. Geogr.*, **29**, 57–63.
Longley, P. A. and Goodchild, M. F. (2008) The use of geodemographics to improve public service delivery. In *Managing to Improve Public Services* (eds J. Hartley, C. Donaldson, C. Skelcher and M. Wallace), ch. 6. Cambridge: Cambridge University Press. To be published.
Longley, P. A., Goodchild, M. F., Maguire, D. J. and Rhind, D. W. (2005) *Geographic Information Systems and Science*, 2nd edn. Chichester: Wiley.
Office for National Statistics (2009) *National Statistics Postcode Directory*. Newport: Office for National Statistics. (Available from `http://www.ons.gov.uk/about-statistics/geography/products/geog-products-postcode/nspd/index.html`.)
R Development Core Team (2008) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Singleton, A. and Longley, P. (2009) Creating open source geodemographics—refining a national classification of census output areas for applications in higher education. *Pap. Regl Sci.*, **88**, 643–666.
Sleight, P. (1997) *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business*. Henley-on-Thames: NTC.
Tonks, D. and Farr, M. (2003) Widening access and participation in UK higher education. *Int. J. Educ. Mangmnt*, **17**, 26–36.
Vickers, D. and Rees, P. (2006) Introducing the area classsification of output areas. *Popln Trends*, **125**, 15–29.
Vickers, D. and Rees, P. (2007) Creating the UK National Statistics 2001 output area classification. *J. R. Statist. Soc.* A, **170**, 379–403.
Voas, D. and Williamson, P. (2001a) The diversity of diversity: a critique of geodemographic classification. *Area*, **33**, 63–76.
Voas, D. and Williamson, P. (2001b) Response (The diversity of diversity). *Area*, **33**, 335–336.
Webber, R. (2005) Classifying pupils by where they live: how well does this predict variations in their GCSE results? *Working Paper 99*. Centre for Advanced Spatial Analysis, University College London, London.