UCL CENTRE FOR ADVANCED SPATIAL ANALYSIS

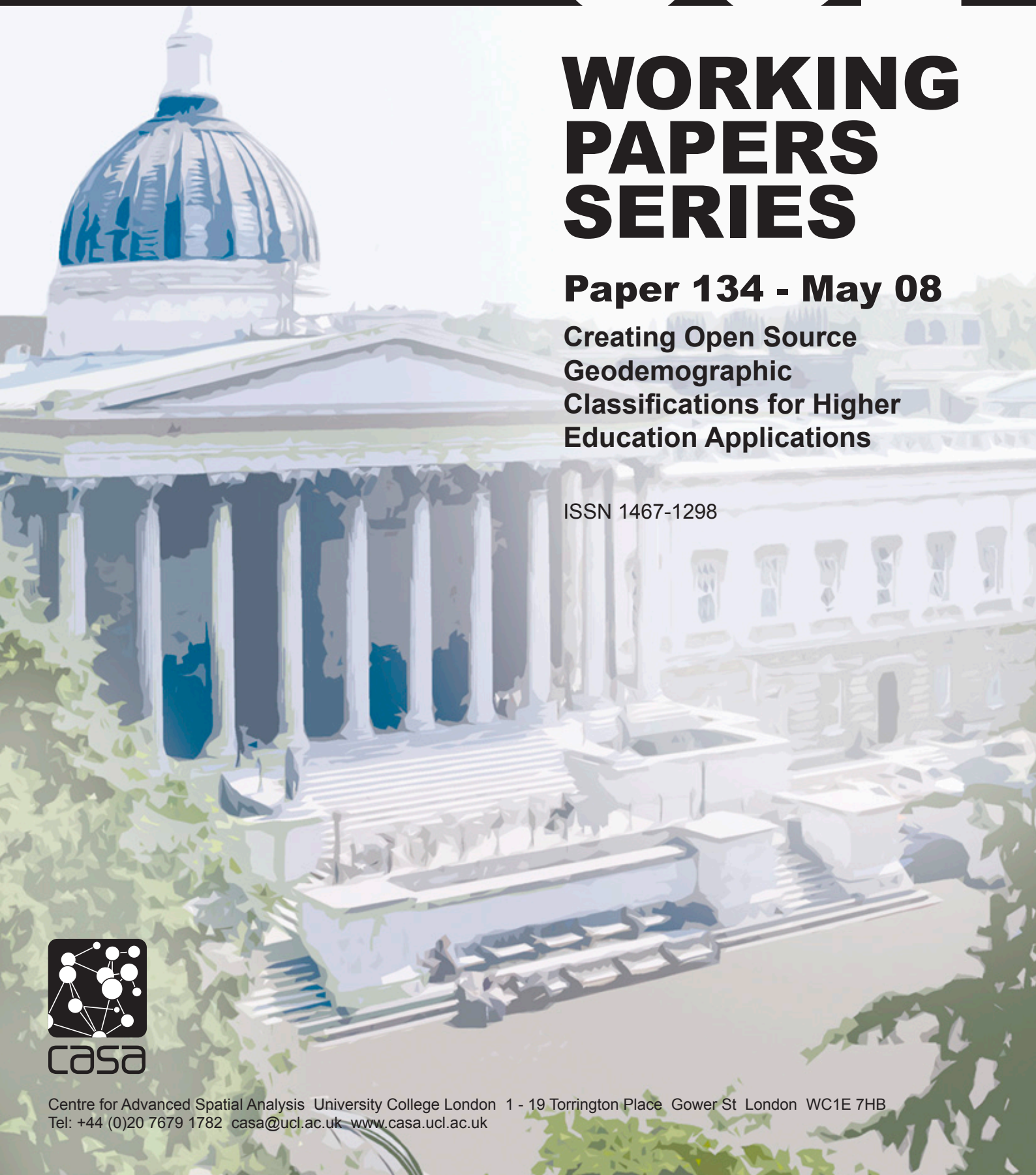
UCL

WORKING PAPERS SERIES

Paper 134 - May 08

Creating Open Source Geodemographic Classifications for Higher Education Applications

ISSN 1467-1298



casa

Centre for Advanced Spatial Analysis University College London 1 - 19 Torrington Place Gower St London WC1E 7HB
Tel: +44 (0)20 7679 1782 casa@ucl.ac.uk www.casa.ucl.ac.uk

Creating Open Source Geodemographic Classifications for Higher Education Applications.

Singleton, A.D., Longley, P.A[⊥].

Abstract

This paper explores the use of geodemographic classifications to investigate the social, economic and spatial dimensions of participation in higher education. Education is a public service that confers very significant and tangible benefits upon receiving individuals: as such, we argue that understanding the geodemography of educational opportunity requires an application-specific classification, that exploits under-used educational data sources. We develop a classification for the UK higher education sector, and apply it to the Gospel Oak area of London. We discuss the wider merits of sector specific applications of geodemographics, with particular reference to issues of public service provision.

Keywords: geodemographics; higher education; classification; Output Area Classification (OAC), GIS

This paper is a draft of a paper in preparation for submission to Papers in Regional Science.

[⊥] A D Singleton, Research Officer, Spatial Literacy in Teaching Project, P A Longley, Professor of Geographic Information science, Department of Geography, and Centre for Advanced Spatial Analysis, University College London, Gower Street, London WC1E 6BT, UK; a.singleton@ucl.ac.uk, p.longley@geog.ucl.ac.uk

1. Introduction

This paper addresses the development and application of geodemographic classifications to better understand participation in UK higher education to higher education. Our motivation originates in the observation that general purpose classification systems (such as those marketed by commercial providers) can claim no particular status in accounting for the consumption of the various services provided by the *public* sector[1], the more so because of the 'black box' nature of the weighting schemes used to derive such classification systems. Moreover, there is clearly a spatial, as well as a socio-economic, dimension to the pattern of participation in higher education (Sa *et al* 2003).

Accordingly, we seek instead to develop a bespoke geodemographic clustering system to account for decision making in relation to prevailing provision of Higher Education (HE), using HE data provided by the Higher Education Statistics Agency (HESA: www.hesa.ac.uk) and Universities and Colleges Admissions Service (UCAS: www.ucas.ac.uk). As is the case with most commercial classifications, Census data account for a substantial part or in the unique case of the People and Places (P2) from Beacon Dodsworth[1] all of the data, but our approach is to supplement these with systematically collected HE domain-specific data rather than the mélange of shopping questionnaires and other sources that are used in developing commercial classifications. While a number of commercial geodemographic systems have been used to account for differences in the uptake of HE between different groups, we believe that our classification is the first to have been specifically designed for this purpose.

Our methodology is dependent upon the National Statistics Output Area Classification (OAC: Vickers and Rees, 2007), which is an open source geodemographic typology built entirely from the 2001 Census data. Unlike commercial solutions, the derivation of this classification is in the public domain, the classification can be reproduced entirely from public data sources and it has the status of a national statistic. OAC divides neighbourhoods into a hierarchy consisting of 7 Supergroups, 21 Groups and 52 Subgroups in a way that is designed to present balanced summary measures of demographic and socio economic conditions. For present purposes, however, it is deficient in data that are directly linked to

[1] One partial exception is Health Acorn (www.caci.co.uk/acorn/healthacorn.asp). This does not directly include health sector data, but does include other data relating to health outcomes, e.g. diet information.

participation in collectively provided education services. In seeking to remedy this, a technical contribution of this paper is to create a further level to the OAC hierarchy, further dividing the Subgroups into 176 “Microgroups” that can be used to classify all UK Output Areas (OAs). HE data are appended to these Microgroups, which are then re-clustered to build a new two level hierarchical classification informed directly by education domain data.

We suggest that this is a valid and useful reworking of the OAC classification prior to its use as part of a bespoke (i.e. application specific) educational geodemographic system, and that it has wider implications for the development of bespoke geodemographic discriminators for which domain specific data can be made available at fine spatial levels of granularity. We suggest that this approach has inherent advantages over attempts to ‘re-badge’ commercial classifications, and that it has wider implications for applications concerning the uptake and use of public goods and services. Moreover, we argue that transparency in the construction and weighting of geodemographic classifications is an important consideration when applications raise issues of social equity in the allocation of public goods and services.

In the remainder of this paper we begin by describing the creation of a bespoke geodemographic classification that combines public domain and HE sector-specific data, using clearly specified techniques and tools. We then develop a pilot application which might be refined and deployed as a service to Higher Education (HE) for a range of applications.

2. University enrolment and HE data

The Universities and Colleges Admissions Service (UCAS) centrally manages the application process for almost every full time undergraduate HE course in the UK. Applicants make an initial selection of six choices (applications) which each identify an institution, course and campus. UCAS is the custodian of these data along with various attributes on the individual applicant. The majority of applicants submit their applications electronically and much of the data processing is automated.

In order for HE funding councils to apportion funds appropriately based on student admissions through UCAS, data are required on the size, and nature of each institution's annual intake. These data are acquired through HESA, which serves as the "central source for collection and dissemination of statistics about publicly funded UK higher education" (HESA, 2006). All publicly-funded UK HE institutions are required to submit an annual "HESA Return", which follows a standard format that details the numbers and characteristics of students within the institution. Various data are collected: however the most important source in terms of volume is derived from UCAS sources supplied at the end of each annual application cycle. Institutions are encouraged to maintain and update these data as they can have bearing on aspects of central government funding in subsequent academic years. Data sourced by HESA through UCAS and other sources make it possible for Higher Education Funding Council for England (HEFCE) to calculate institutional allocations of additional government funding to support widening participation initiatives amongst young participants (defined as aged less than 21): in this case, the measure is derived by students being grouped into participation rate assigned by the ward in which they reside (HEFCE, 2005). The two key datasets for the HE sector are assembled by UCAS and HESA, and it is the former of these that is specifically associated with undergraduate admissions. Both UCAS and HESA made data available for this study, and the variables that are of interest are discussed below.

Having asserted that the nature and consequences of decision-making in HE justifies the development of a bespoke geodemographic system, it is incumbent upon us to outline the range of applications in which such geodemographic discriminators may be useful. HE activities including institutional marketing, extending access, widening participation or subject specific targeting are all candidate applications, and these should therefore be borne in mind when identifying candidate input variables for inclusion within the classification. In the literature on widening participation from which these applications predominantly extend there are a range of discussions on the determinants of access to HE inequalities. Reid (1998) discusses that there are two interpretations of inequality in Higher Education: first, that there is bias in the university selection process; and second, social class has an inhibitor effect on the perceived availability or benefits of Higher Education. The first of these interpretations was publicly highlighted in 2001 with the case of Laura Spence. Her rejection by the University of Oxford on the basis that she "did not show potential" created a media circus that even involved the then

Chancellor of the Exchequer (and now Prime Minister) who declared it “an absolute scandal”. The second of these interpretations relates to how middle class parents ‘invest all kinds of effort, including significant material resources in developing social capital’ (Walker, 2003:172), creating environments where socialisation processes can occur, and creating advantage or disadvantage under certain situations (Bourdieu & Passeron, 1977). Outside of the social, cultural and economic actors on human capital accumulation the 2003 Higher Education White Paper (DfES, 2003:68) accepted that ‘the single most important cause of the social class division in Higher Education participation is differential attainment in schools and colleges’. It is therefore important to select input variables for their ability to stratify both recorded causes of participation disadvantage such as variable attainment, and also attempt to measure those conceptual causes such as human capital accumulation. The data available for this study are for 2001 and cover all students domiciled in England and studying at English institutions. This database contains a variety of suitable variables for inclusion in the cluster analysis and those selected for our analysis are shown in Table 1, and the undergraduate courses to which they relate are shown in Table 2. The variables selected aim to measure the characteristics of participation in terms of their stratification between different groups of people (e.g. ethnic groups, independently educated, course choices, distance travelled), include direct measures of participation (e.g. participation rate) and finally those causes of these inequalities (e.g. Social Class and A-Level scores). The core advantage in informing a classification with application specific data at the build stage is that the groups which result from the clustering procedures should better fit the underlying dimensions which they seek to represent. Thus, for an application in HE, it is sensible to include variables relating to those actual participants in HE, rather than a blend of possibly undisclosed variables which may show some correlation to the social, economic and spatial patterns they are seeking to measure.

Table 1: HE input variables to the cluster analysis.

Variable	***Numerator***	***Denominator***
Young participation rates	First year students aged 18-19.	Census 2001 18-19
Average distance from student's home to institution	N/A	N/A
Average A-Level Score of students	N/A	N/A
Proportion of students from low social class groups	Undergraduate degree students from the three lowest social classes (IIIM, IV, V)	All undergraduate degree students
Proportions participating in particular degree course groupings*	Students studying undergraduate degree courses within groupings (A-X).*	All undergraduate degree students
Proportion from ethnic minority Groups*	Undergraduate students from ethnic minority groups.	All undergraduate degree students
Proportion of students previously educated in Independent Schools in Years 12 & 13	Undergraduate students who previously attended independent schools.	All undergraduate degree students

* = Course and Ethnic Groups are defined in Table 2.

Table 2: Course and ethnicity groupings.

Course Groups	*Short Code*	*Ethnicity Groups*
Medicine & Dentistry	A	White
Subjects allied to Medicine	B	Black or Black British – Caribbean
Biological Sciences	C	Black or Black British – African
Veterinary Science, Agriculture & Related.	D	Other Black background
Physical Sciences	F	Asian or Asian British – Indian
Mathematical & Comp Sciences	G	Asian or Asian British – Pakistani
Engineering	H	Asian or Asian British – Bangladeshi
Technologies	J	Chinese or Other Ethnic background – Chinese
Architecture, Build & Plan	K	Other Asian background
Social Studies	L	
Law	M	
Business & Administration Studies	N	
Mass Communications and Documentation	P	
Linguistics, Classics & related	Q	
European Languages, Literature & Related	R	
Non-European Languages and Related	T	
Historical & Philosophical Studies	V	
Creative Arts & Design	W	
Education	X	

In a traditional statistical model it would not be appropriate to include use the same data as both dependant and independent variables, and as such the variables selected above could be critised for a

degree of circularity. For example, a core aim of the classification is to stratify participation, however participation is included as an input variable. In defence of this decision it should be noted that cluster analysis does not suffer those same limitations of regression based statistical models as the algorithm seeks similarity rather than explanation. Furthermore, to provide further reassurance that the classification would be safe to use, it might only be used for profiling data from a separate sampling frame, e.g. a separate year of HE data.

Although young participants as defined by HEFCE are applicants accepted by an HE institution who are aged 21 years or younger, in practice the majority that were accepted through UCAS during the period 2000-2004 were aged 18-19. For purposes of estimating participation rates, a base population of 18-19 year olds was extracted from the 2001 Census and compared with the average number of same age band students from the HESA data known to be attending HE. If a 2001 Census base count of all residents aged 21 or less had been used this would of course produce far lower 'participation' rate figures, and might be biased – for example, in underestimating participation rates in new estates with young families whose offspring are only just entering the 18-21 cohort.

The use of distance travelled to accept a degree place provides a useful proxy for the geographic constraints upon choice (Sa *et al* 2003) that are particularly incumbent upon some applicants from lower socio-economic groups, either because of the financial cost of travel to a distant institution or the social networks which may bind them to their local communities (Reay *et al,* 2005). Straight line ('crow fly') distance between the accepting institution and the student's home is used in this analysis, on grounds of simplicity and ease of calculation. (It is, of course, the case that mobility may be more prevalent in and around metropolitan areas which have public transport hubs.) The co-ordinates for student home locations (*i*) and chosen HE destinations (*j*) were taken from the 2001 All Fields Postcode Directory. Including distance in the classification is useful to identify areas where students are less likely to travel, particularly if they reside at home, a factor which often indicates limited financial means.

The 2001 HESA data measure UK A-Levels attainment on a points scale, ranging from 10 points for an A grade to 2 points for an E, and summed across all subjects of study. Prior attainment, particularly with regard to traditional academic qualifications such as A-Level have been seen as "key to the

reaffirmation of middle class privilege in education and employment" (Leathwood and Hutchings, 2003: 153) and as such is likely to provide a good discriminator of neighbourhood inequality of outcome. Where these scores are not recorded in the HESA data, applicants will usually have qualified for HE through a non A-Level route; for purposes of this analysis these are recorded separately as a binary non-A level variable.

In 2001, HESA data on social class were recorded using the Registrar General's Social Scale, which groups occupations into 7 different categories. Low rates of participation by those from households with earners in skilled manual, partly skilled and unskilled occupations have been documented ever since the Robbins Report (Robbins, 1963), and the extent to which these social barriers have been successfully addressed is debatable. In order for the classification to discriminate between the higher and lower social echelons, a variable was created to record the frequency of students from these three social groups.

The 2001 HESA data used the Standard Classification of Academic Subjects (SCAS) to aggregate individual courses into subject groupings. The extent to which different neighbourhood types participate across these subjects is critical for both marketing and widening participation initiatives. The inclusion of the proportion of students within each subject grouping is intended to improve the ability of the classification to discriminate according to subject of study. Information on student ethnicity was included because membership of some ethnic minority groups has been observed to be associated with low participation in some subjects and markedly higher participation with others (Gilchrist *et al*, 2003).

An index score can be used to show the overrepresentation of a target group by a discrete classification when compared to its proportions in a base population – as in Equation (1) where index scores I are calculated by comparing the proportion of a variable n within a target population t relative to a base population b.

$$I_n = \frac{\frac{t_n}{\sum_{n=1}^{n} t_n}}{\frac{b_n}{\sum_{n=1}^{n} b_n}} \times 100 \qquad (1)$$

Table 3 shows index scores created using the 2005 UCAS acceptance data that illustrate the differing propensities of ethnic groups to participate in different courses. It can be seen, for example, that students of Asian ethnicity are almost 2.5 times more likely to study Medicine or Dentistry than the student population as a whole.

Table 3: Indexed participation rates in different subject groupings according to ethnicity and subject of study (source: 2005 UCAS acceptances).

	Asian	Black	Mixed	Other	White
A Medicine & Dentistry	246	63	124	188	88
B Subjects allied to Medicine	152	135	76	116	93
C Biological Sciences	68	76	100	83	107
D Veterinary Science ,Agriculture & Related	26	16	36	27	117
F Physical Sciences	54	36	75	41	113
G Mathematical & Computer Sciences	190	137	93	137	85
H Engineering	123	137	96	130	94
J Technologies	72	64	73	100	106
K Architecture, Building & Planning	97	85	88	101	101
L Social Studies	92	158	110	99	97
M Law	178	147	108	140	88
N Business & Administrative Studies	165	159	94	126	87
P Mass Communications and Documentation	57	111	129	104	104
Q Linguistics, Classics & Related	41	38	112	73	112
R European Languages, Literature & Related	24	30	121	76	116
T Non-European Languages and Related	42	31	195	116	111
V Historical & Philosophical studies	28	20	90	70	116
W Creative Arts & Design	40	67	118	92	107
X Education	46	50	53	51	112

The inclusion of a variable identifying whether a student previously attended independent school is designed to improve the ability of the classification to identify those neighbourhoods which supply disproportionate numbers of students previously educated outside the state sector. One of the HEFCE widening participation performance indicators is based on proportion of students coming from state schools, so the inclusion of this variable is in line with this performance measure.

The full range of additional variables included in the cluster analysis is detailed in Table 4, along with the short codes that are used in the presentation of results.

Table 4: Variables included in the cluster analysis and their short code references.

Variables	
A-Level points	Social studies
Distance travelled to attend institution	Politics & Law
Lower Social Class	Business and Administrative Studies
Black Caribbean	Mass communications and Documentation
Black African	Linguistics, Classics and related subjects
Other Black	European Languages, Literature and related subjects
Asian Indian	Eastern, Asiatic, African, American and Australasian Languages.
Asian Pakistani	Humanities
Asian Bangladeshi	Creative arts
Chinese	Education
Other Ethnicity	Combined and general courses not otherwise classified
18-19 Young Participation Rates	No A-Level Points (i.e. non A-Level Qualifications)
Medicine and Dentistry	
Subjects allied to medicine	
Biological Sciences	
Agriculture and related subjects	
Physical Sciences	
Mathematical sciences and Informatics	
Engineering	
Technology	
Architecture, Building and Planning	

3. Creating the building blocks of an open source geodemographic classification

In recent years, the more proactive stance of UK government departments towards the dissemination of public statistics, including the Census of Population, has made it possible for a greatly broadened constituency of interested parties to develop their own classifications of neighbourhood characteristics. However, the construction of geodemographic classifications is a skilled task, and this new freedom inevitably raises issues surrounding the inter-correlation of census measures, as well as considerations of how compound indicators might be construed as representing a single or multiple construct of reality. Many of these issues have already received detailed investigation in the creation of successful general purpose classifications, and thus there is merit in building upon the achievements of such classifications as a base upon which to add bespoke elements. The UK Office of National Statistics' (ONS) Output Area Classification (OAC: Vickers and Rees 2007) was created from the 2001 Census using 41 variables common across all of the UK, and describes the demographic, household composition, socio-economic and employment characteristics of each Census Output Area (OA) in England, Scotland, Wales and Northern Ireland. Vickers and Rees used *K*-means clustering to create a

high order classification comprising seven Supergroups. The input data pertaining to the OAs that had been classified into these seven clusters were then subdivided, and each re-clustered to create a second tier comprising 21 groups. This process was then repeated a third tier of 52 Subgroups (see Figure 1). OAC has been ratified it as a UK national statistic by ONS, and the classification can be downloaded from its user group website (www.areaclassification.org.uk). The classification has recently been appended to a series of large national datasets including the National Statistics Postcode Directory and the 2001 Census of Population.

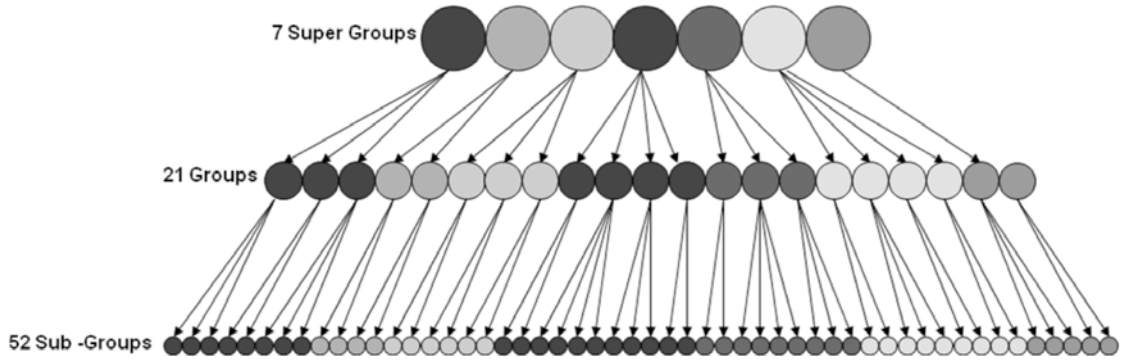


Figure 1: The National Statistics Output Area Classification hierarchy.

The variables included in OAC were selected to "represent the main dimensions of the 2001 Census" (Vickers and Rees, 2007: 383), and although these do include a single variable on HE attainment, they do not incorporate direct measures of area HE participation rates in either aggregate form, or broken down by subject preference. It is the aim of our classification to infuse spatial variation in socioeconomic characteristics pertinent to HE participation into the OAC classification hierarchy. Vickers and Rees (2007: 381) take the established line that when clustering data "there is no right or wrong answer", just a range of different combinations leading to an "infinite number of parallel classifications". This view is also presented by Gordon (1981), who contends that the statistical process of "clustering" of attribute space can better be considered to be a process of dissection, wherein clusters should not be conceived as discrete objects existing within a multidimensional space, but rather as the outcome of dissecting more subjective and fluid categories, the boundaries of which can be repositioned to create alternative representations.

In the context of these arguments, there are two broad methods which might be used to build upon the experience of the OAC classification in order to construct an educationally weighted geodemographic classification:

1. Re-cluster a new classification from OA level upwards, based upon the documented experience of creating OAC, but including educational data alongside the original OA data in the 'bottom up' classification.
2. Adapt the existing OA classification, by re-clustering it from a finer scale created as a disaggregation of Subgroups, and after adding sector specific data.

There are a number of problems associated with the former option. First, the classification would need to be recreated from first principles, and the valuable qualitative experience of creating the OAC would need to be re-learnt, for example, with respect to comprehensive evaluation and normalization of input variables. Second, while the variables used in the OAC classification are generally quite highly variable interval scale counts derived from decennial census data, participation rates in HE tend inevitably to fluctuate at fine geographical scales between years (Corver, 2005). This is likely to create an uneven geographical coverage which would either increase the prevalence of outlier values or require many structural zeroes to be accommodated in the geographic matrix, with deleterious consequences for the classification procedure. This in turn would lead to a need for standardisation and careful population weighting. For these reasons, we chose to pursue the second option.

The first stage in developing the bespoke educational classification entailed the creation of a new finer tier in the OAC typology. This proceeded in a way analogous (albeit different procedurally) to Experian's (Nottingham, UK) Mosaic Segments product, which provides a 243 cluster disaggregation of its 61 Mosaic Types. Previous work using the Mosaic classification (Singleton and Farr, 2004) has suggested that this fine level of disaggregation is effective for re-clustering of education data.

Commercial classification builders tend to cluster at the finest level first and then aggregate these fine segments into successively larger groups. The OAC methodology created the classification in the opposite way, disaggregating first at the highest level and then dividing these groups into the smaller clusters which form the lower two tiers. OAC was created using the *k-means* clustering method (MacQueen, 1967). This is an iterative relocation algorithm that assigns each data point into one of *k* clusters based on a standardised Euclidean minimum distance metric. The algorithm seeds the initial

locations of the *k* cluster centroids as random data points within this data matrix. The distance of the data points to each cluster centroid is then calculated, and each data point is provisionally assigned to its nearest cluster centre. A clustering criterion statistic is then applied to measure the homogeneity within these temporary cluster allocations. After the first iteration of the model, the *k-means* algorithm attempts to find a local optimum through an objective function that reallocates data points iteratively from their initial assignments. Each data point is considered for reallocation to other clusters, and after each test the model objective function is recalculated. Where reassignment of data points does occur, the cluster centroid values for the gaining and losing clusters are recalculated. Once the objective function is minimised, or the user-specified maximum number of iterations is reached, no further reallocation of data points takes place. However, writing in the 1970s, Everitt (1974: 26) observed that "there is no way of knowing whether or not the maximum of the criterion has been reached". This is because in a single *k-means* model there are likely to be multiple local optima, since the random placement of the initial cluster seed centroid means that multiple locally optimised models are possible. Using the data from the OAC classification, Figure 2 illustrates the problem of running models starting from two different initial values to convergence, with $k = 9$ and two variables (Black versus Asian - Indian, Pakistani or Bangladeshi).

Model Run 1 **Model Run 2**

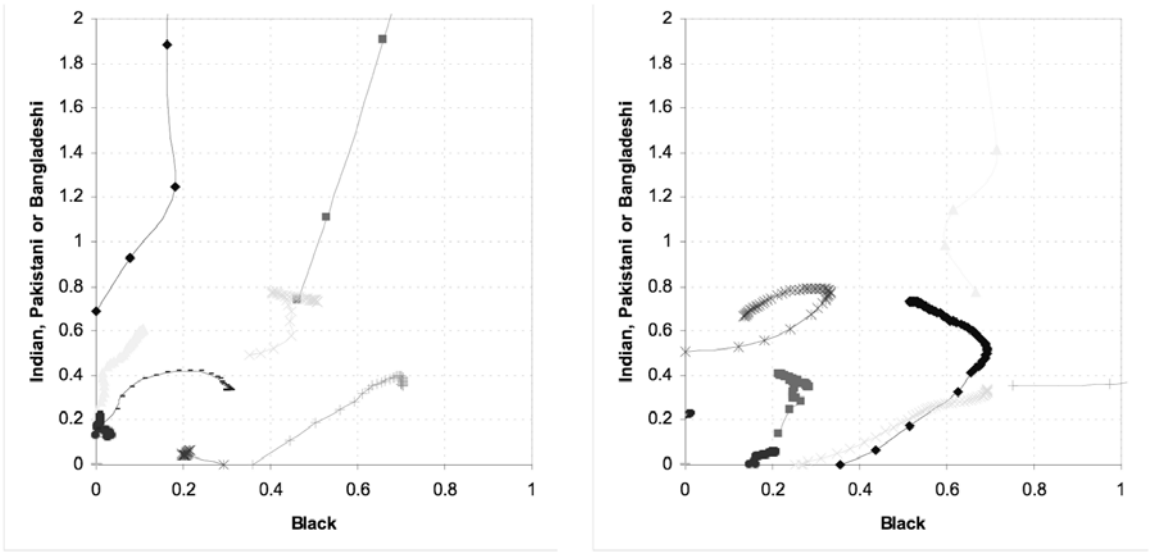


Figure 2: An illustration of the effect of random initial seed locations upon final model outcome.

These graphs show how the path of the cluster centroid can converge upon entirely different locations, depending upon the random placing of the initial seed. Furthermore, as each iteration of the model reallocates data points to cluster centroids, "making the 'best' decision at each particular step does not

necessarily lead to an optimal solution over-all" (Harris et al, 2005: 162). While the partitioning of the input data in any given cluster model is globally optimised, this outcome may be critically dependent upon initial conditions – specifically the random placing of the cluster seeds – and there is no benchmark of global model performance for an individual data set. However, recent experimentation with multiple seeding algorithms (Brunsdon and Charlton 2006) suggests that, given sufficient computational power, a globally optimised local model can be obtained by running k-means multiple times to convergence, comparing the results from each cluster analysis and saving the best performing classification. Figure 3 shows the results from the same *k*=9 model which was run with a random seed allocation 150 times: for each model an R-squared statistic was generated in order to estimate the quality of the model discrimination. This graph highlights the variability in overall model performance arising from placement of the initial seeds.

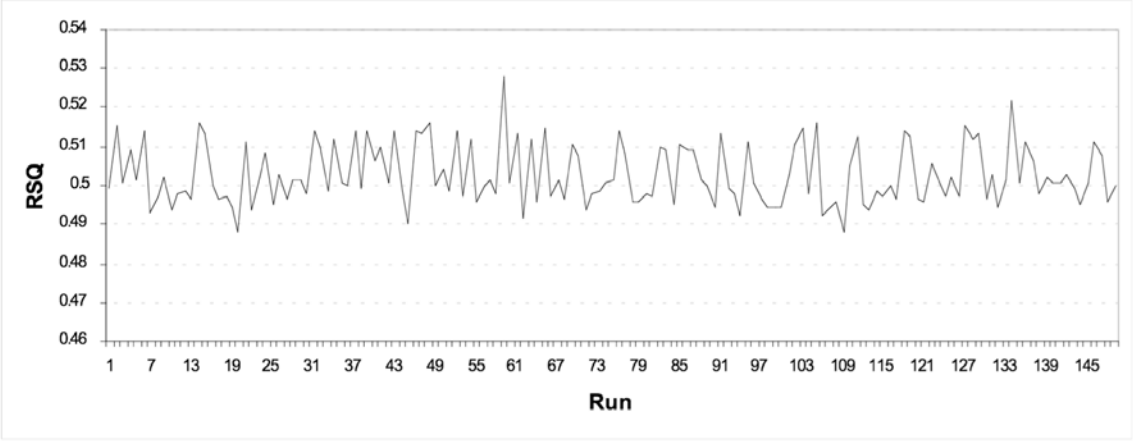


Figure 3: R-squared results from repeated runs of a nine cluster model.

A further issue with the *k*-means algorithm is the *a priori* decision to define an "appropriate" number of clusters. In describing the construction of OAC, Vickers and Rees (2007) cite the prevailing views of geodemographic practitioners about appropriate cluster frequency, based on understanding of what has been deemed successful practice in commercial products. This leads them to advocate a three tier partitioning of 2001 Census data, first into 7, then 21 and then 52 partitions of their OA data. A further method demonstrated in a geographical context by Debenham (2001) is to calculate the average distance between the data points and their assigned cluster centroid at model convergence for a range of different *k* values. A judgement can then be made on an appropriate number of clusters, weighing up the relative merits of a cluster taxonomy which can be readily interpreted by end users, and a level of detail which yields reasonably homogenous within-cluster characteristics.

The non-educational input data used to create the finer level classification consisted of the same set of standardised 2001 Census variables at OA level that was used to construct OAC. This dataset was split into 52 separate groupings of Output Areas mirroring their assignments in the OAC subgroup classification and each of the resulting 52 datasets was separately re-clustered using the *k*-means algorithm of the SAS (www.sas.com) statistical software.

Different numbers of OAs are assigned to each OAC Subgroup and it is necessary to take this into account in any further partitioning of Subgroups in order to maintain a balanced degree of uniformity between the newly created Microgroups at OA level. The alternative of dividing each of the 52 datasets by the same *k* value would create clusters of quite starkly varying sizes because the totals of the 18-19 year old population contained within each of them would differ. The outline objective was to create a total of around 264 clusters (Microgroups) of approximately similar size across all 52 separate datasets, in order to create a classification that was comparable with commercial offerings. The population distribution shown in Figure 4 was used to estimate the initial frequency of the divisions required to create the Microgroup classification. The *x* axis records the 52 OAC Subgroups (ordered by ascending population size), and the y axis denotes the total percentage of 18-19 year olds within each OAC subgroup as measured in the 2001 Census. The initial estimated divisions (*k* values) that would be required to create the Microgroup classification with even population between clusters are denoted on the bars. These estimates are calculated by apportioning the total Microgroups required (264) within each Subgroup using percentage figures of the 18-19 year old population. Some measure of accommodating variable target population size within OAC Subgroups is necessary, since partitioning all of the OAC Subgroups evenly (e.g. 264 / 52 = ~5) would create Microgroups with a very uneven distribution in the 18-19 population cohort. It was found that if uneven clusters such as these were used as the basis upon which to build a bespoke educational classification, they caused the formation of a new classification with very uneven population size and as such with limited applicability.

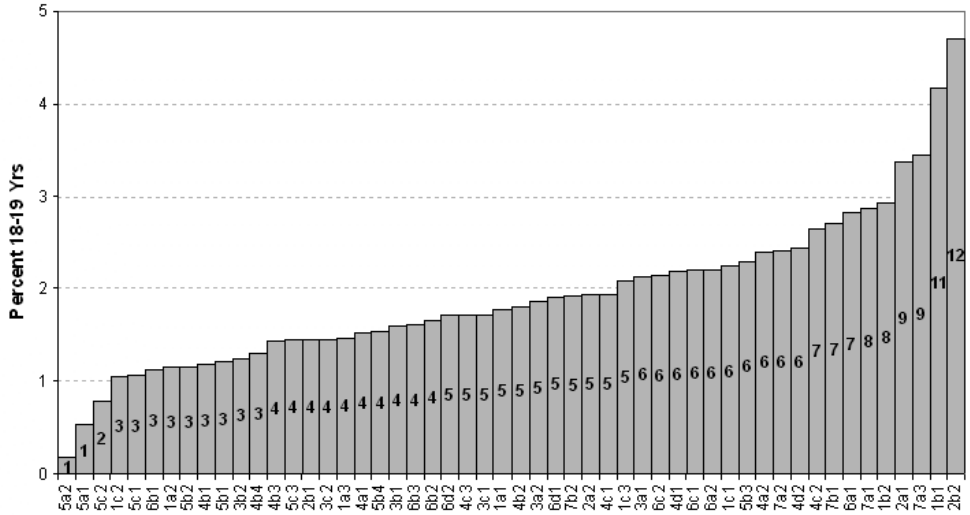


Figure 4: Percentages of all 18-19 year olds falling into each OAC Sub Group and their assigned *k* values.

Even after apportioning initial clustering values using 18 – 19 year old population size, the clustering algorithm still created a number of outlier clusters. In order to minimise the number of small population counts within clusters, thus improving the uniformity of between Microgroup population, a number of the *k* values were therefore manually assigned in order to create more evenly distributed clusters. Although this is not desirable in a classification designed with a transparent and defendable build process, it was deemed necessary to prevent very small outliers being created, and as such having a negative effect in the final classification. Where the initial division of Subgroup datasets had created outlier clusters (of small population size), different values of *k* were re-assigned and then tested in order to assess whether they might create more uniform cluster populations. The final outcome was the creation of the Microgroup classification which divided all Output Areas into 176 Microgroups. This was fewer than the initial aim (264), but resulted in a classification with a reasonably even population distribution of 18-19 year olds. This Microgroup classification provided the classification onto which the education data were appended.

4. Building the bespoke HE geodemographic classification

Individual student records from HESA data were georeferenced to home unit postcodes and linked to Output Areas using the All Fields Postcode Directory[2]. These output areas were then joined to the Microgroup classification, thereby assigning each student from the HESA database to one of the 176 clusters. A series of binary scores (e.g. to represent subject of study) was created for a number of

[2] This is now called the National Statistics Postcode Directory and is available from the ONS website: http://www.statistics.gov.uk/geography/nspd.asp

categorical attributes about the individuals contained within in the database. Additionally a number of continuous variables (see Table 1) were also created for each student (e.g. distance travelled from home to study at university). Using Microgroups as an aggregating field, the HESA data were grouped using Structured Query Language (SQL). Binary variables were summed to create total frequencies of students according to Microgroup for a range of attributes, and the median values of continuously scored student variables were created for each Microgroup. Thus, the output dataset consisted of 176 rows for the Microgroups and a series of columns for both *frequency* counts or *average* scores for a range of variables relating to those students classified by the Microgroups.

The frequency counts for each Microgroup were converted to index scores (where 100 denoted average incidence, 200 double the incidence, and so forth) derived from a base distribution of the total frequency of students recorded in the HESA database. The "young participation" variable, taken from the 2001 Census, used a base score of the total number of 18-19 year olds. The continuous variables (e.g. distance) that were averaged by Microgroup were not converted to index scores, however, because the clustering model required that all input data are measured on the same scale both the frequency and the average variables were converted to *z*-scores. The clustering algorithm treats all variables as continuous and as such the scale must be comparable between variables, otherwise those variables with a larger range will have adverse affect on the final assignment of clusters, skewing results towards their extreme values. The process of conversion to z-scores was used to control for the different scales used to measure the input variables, representing all variables using a standard deviation unit of measurement. Before clustering, the Microgroups with their standardised input variables were weighted by the total population within each cluster, thereby reducing the influence of those Microgroups with smaller population sizes. Unweighted *k*-means is often used for outlier detection in multidimensional datasets, but in geodemographic applications, very low population counts in some clustering units can reduce the efficiency with which it is possible to both describe and discriminate between clusters.

Additionally, before performing cluster analysis, the data were explored to examine the correlations between the variables. It is the case that high correlations amongst the raw variables included within a cluster analysis result in data redundancy and can have undesirable effects in the final assignments to clusters (Vickers and Rees, 2007). Harris *et al* (2005) also emphasise the importance of including

variables that add new information rather than repeat what is already known. It is claimed, for example, that the methodology employed by Experian in the construction of Mosaic allows correlations to be accommodated through the use of weighting schemes, albeit at the expense of introducing subjective value judgments to the classification procedure. Insofar as such weights are not made public, such weighting also renders classifications opaque and non-reproducible by other researchers. For these reasons our own interim view is that, for clustering applications in the public sector, weighting schemes are difficult to justify if they are not empirically grounded and are potentially influenced by the predilections and experiences of the clustering solution creator. This essentially inductive view is that the disbenefit of noise and uncertainty generated by data led generalisation are outweighed by the greater risk of straightjacketing a classification to realise pre-ordained outcomes. Only recently have studies been conducted into how weighting schemes can be automated through an adaptation of the *k* means algorithm (Huang *et al*, 2005): such algorithms remain relatively untested and have not as yet been implemented in commercially available statistical software. Other applications of cluster analysis have side-stepped the complexities of including multiple variables with their related correlations through the reductive technique of Principal Component Analysis (PCA), where "each component represents a weighted combination of the original variables" (Voas and Williamson, 2001: 65). Although some view PCA as useful to filter variables that may be redundant or have negative effects upon classification outcomes (Debenham *et al*, 2002), a contrary view is that the technique results in undesirable information loss and creates complexity in results which are difficult to interpret (Harris *et al*, 2005).

Correlations between input variables have the effect of adding extra weight to one or more dimensions of the classification, thereby creating a very similar effect to manual weighting of raw variable scores. In the absence of manual weighting of correlated variables, the only way to avoid such weighting effects would be to disregard highly correlated variables, raising the question of which of the correlated variables to remove. The analysis of correlation effects and identification of the thresholds at which they should be deemed undesirable presents difficult decisions in practice.

Geodemographic classification through cluster analysis therefore inevitably presents a dilemma between seeking to let the data speak for themselves (Everitt and Dunn 1983) and using manual

intervention to create a classification that is intuitive, fit for purpose and defensible. Unlike commercial classifications, the clustering procedures which created OAC were performed without weighting, and although these weighting schema could have improved overall classification performance and possibly aid discrimination between areas, the argument of Vickers and Rees (2007) is that it would have added potential bias and introduce subjectivity into the composition of output clusters. We are persuaded by these arguments in the creation of our own classification for what is essentially a range of public sector applications.

A key purpose of the HE classification under development here is discrimination between areas that are characterized by abnormally high or low participation rates. One would expect the various factors which lead to these patterns of inequality to be highly correlated – for example A-Level points score with social class as it is well documented in the literature that there are relationships between these variables (Reid, 1998). Both of these variables contribute towards low participation and as such should be included in a classification wishing to measure this dimension. Their potential correlation re-enforces an important dimension of the classification and as such should be allowed to manifest in the final cluster assignment.

In order to further understand the pattern of correlation amongst the Microgroups within the data, a correlation matrix of all input variables was created using a Pearson correlation coefficient (though not reproduced here for reasons of space). High intercorrelation was apparent between A Level points score, education at independent schools, distance travelled to accept a place and HE participation rates of those aged 18-19. As one would expect, each of these variables exhibits a strong negative correlation with low social class and entry through routes other than A-levels. These patterns are unsurprising and the variables are core to the ability of the classification to discriminate between areas of high and low participation. One would also expect the values of these variables to correspond with subject of study, given that entry grades vary between subject groups, subjects appeal to different types of people and subjects are not evenly distributed across HE institutions. For example there was a high negative correlation between low social class and participation in Medicine and Dentistry. The variables chosen for the bespoke educational classification developed here includes a range of variables that are each

directly relevant to educational outcomes. The correlation matrix revealed some of these were correlated, but the decision was made not to use variable weighting for the reasons discussed above.

The *k*-means algorithm clusters the input data matrix into the *k* groups specified by the researcher. Unless the number of groups that should emerge from the dataset is known *a priori*, a method of selecting an appropriate cluster frequency is required. One method of doing this has been demonstrated by Debenham (2001), and entails running the *k*-means algorithm for multiple iterations of *k* and plotting the average distance between each data point and its closest cluster centroid. Charts may be used to identify the homogeneity of each solution for a range of cluster frequencies. In general, the higher the number of clusters, the smaller the mean distances between each data point and its nearest cluster centroid. The charts constructed by this method thus illustrate the trade-off between mean distance and classification complexity.

Debenham (2001) conducts his analysis by running a single cluster analysis for each *k* value. This has the disadvantage, described earlier, that the *k*-means algorithm is sensitive to the location of initial seeds – a problem that can be largely circumvented through repeated analysis using multiple initial seed values. Debenham (2001) selects a final *k* value based on interpretation of apparent breakpoints in the plot of cluster homogeneity against number of clusters. However, unless the cluster analysis routine is repeated multiple times, these observations may be anomalous because of inappropriately selection of initial random seeds. Thus although this method is useful in principle, it needs to be adapted in order to provide more robust results.

The method adopted in this study builds on Debenham (2001) and runs the algorithm for k_{n-2} models where *n* is the total number (176) of Micro-Groups within the dataset. However, in order to improve the confidence with which the trade off between cluster homogeneity the k initial seeds were randomly repositioned 10,000 times. The median, minimum and maximum distances were averaged over the 10,000 runs for each *k* value and are graphed in Figure 5. In this Figure, the dark line represents the median and the whiskers the minimum and maximum values.

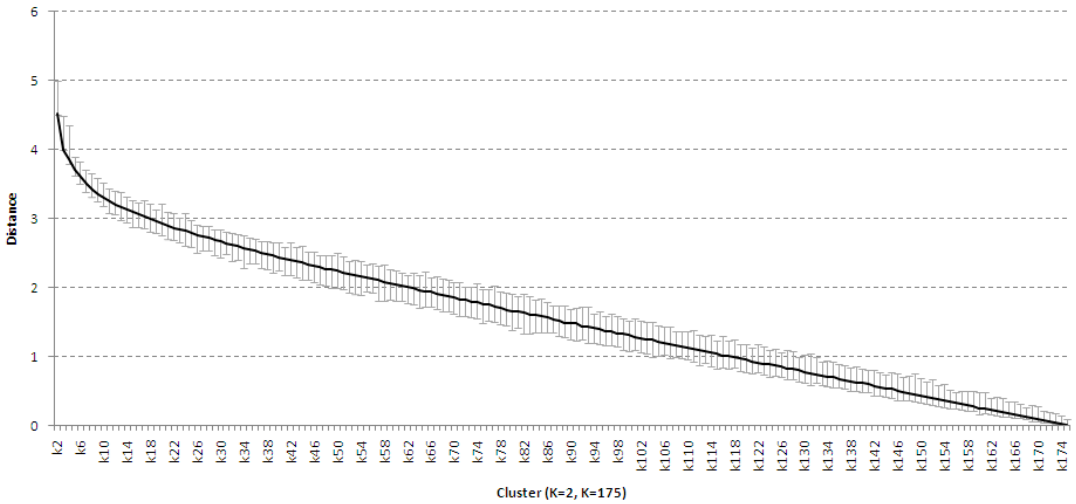


Figure 5: Average distances between cluster data points and closest cluster centroid (k = 2 – 175).

Either side of the median trend line there is a large amount of variability between the most (lower error bar) and least (upper error bar) homogenous solution for each *k* value. This makes it likely that the step functions identified by Debenham (2001) arise because of a small number of particularly good or poor models rather than success in identifying an appropriate cluster frequency through induction. An alternative method is to use *R*-squared statistics that can be calculated from the clustering output by regressing the cluster mean centroid from within the input data matrix against each variable in the input dataset[2]. Using a similar presentation method to the distance chart in Figure 5, the median, minimum and maximum *R*-squared scores are presented for each *k* value in Figure 6. This graph shows that the *R*-squared statistic increases with the number of clusters specified, although not in a linear fashion. Furthermore, as *k* decreases so the variability of the *R*-squared statistic increases, providing further justification of the need for multiple model runs to attain robust information, particularly at lower values of *k*. The increased variability in R-squared at most of the lower *k* values is caused by the grouping of the data points into a smaller number of clusters and this indicates a greater volatility in the assignment of final case allocations between clusters, since this increases the variability of the final classification performance.

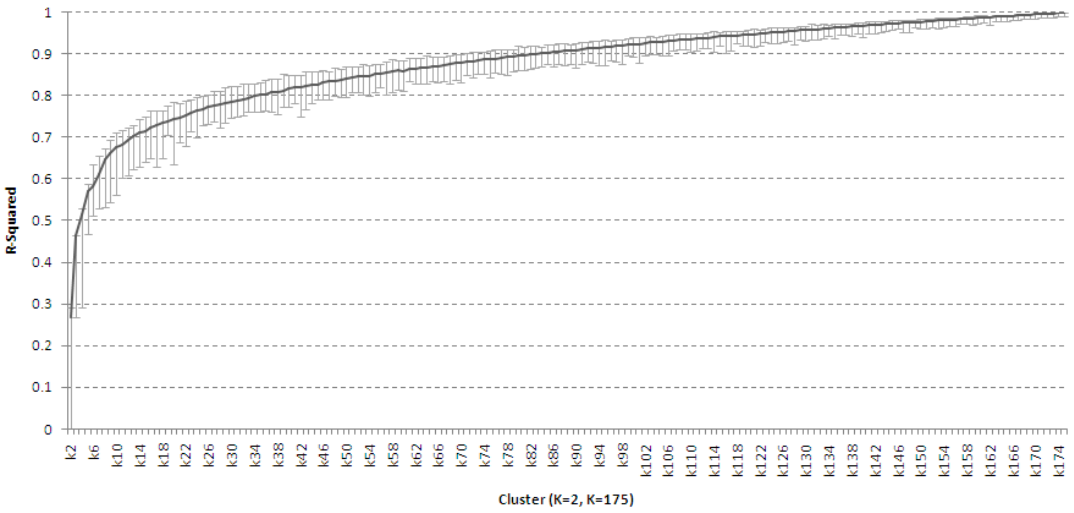


Figure 6: Cluster performance measured by R-squared scores (k = 2 – 175).

Furthermore, these results illustrate how further increases in *k* result in successively smaller improvements in the *R*-squared statistic and how, at the crudest aggregations, much information is lost. The *R*-squared plots are useful for selecting an appropriate cluster number for the dataset, as the loss in performance of the classification can be assessed and compared for each reduction in *k*. Figure 7 shows the difference in *R*-squared scores arising from increasing the frequency of *k* from *n* to *n+1*. The incremental increase in *R*-squared is consistently low beyond $k < 45$ with considerable increases for each incremental value for which $k \leq 25$.

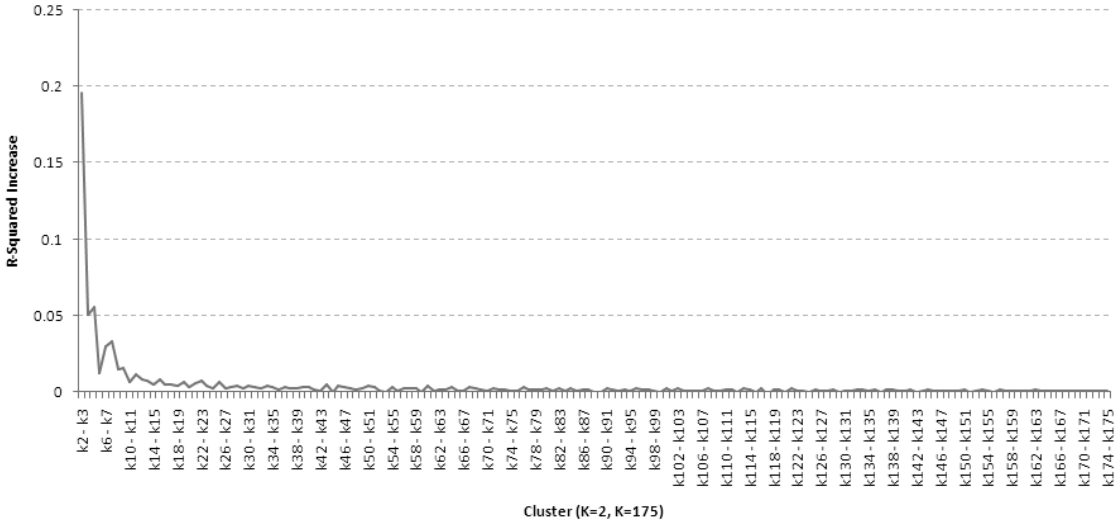


Figure 7: Incremental differences in R-squared values.

Geodemographic classifications typically consist of a hierarchical series of aggregations. This allows end users greater flexibility over the detail they can present and also the number of groups into which their own data are divided. Having a classification with a small number of large aggregations may be

useful when profiling data from a small population, e.g. an unpopular course against all courses at a university. In this context, it is useful to consider the numbers of clusters and levels suggested by a range of classification providers. These are summarised in Table 5. With the exception of OAC, little justification is given as to why particular levels of detail are chosen.

Table 5: Classification levels (source: adapted from Vickers 2005:35).

Classification System	*Clusters in Level 1 (<12 Clusters)*	*Clusters in Level 2 (>=12, <50 Clusters)*	*Clusters in Level 3 (>50 Clusters)*
Mosaic 2001	11	-	61
Cameo	10	-	58
ACORN	5	18	57
PRiZM	-	16	60
Super Profiles	10	40	160
OAC	7	21	52

The commercial classification used most widely in HE since 2001 has been Mosaic, with its two hierarchical levels of 11 and 61 clusters. It was considered preferable to keep our bespoke education classification in line with similar levels and cluster frequencies, in order not to confuse potential end users with radically different aggregations. The final classification should be fit for a range of purposes. Most importantly the classification should provide an effective way of discriminating between those areas of high and low participation in aggregate, and also disaggregated by course types to allow more specific targeting strategies. As in commercial geodemographic classification, it is useful for a bespoke educational classification to have multiple levels, since this creates flexibility when analysing target groups of different sizes. This classification will mainly be used to discriminate between 18-19 year olds as they form the majority of HE cohorts, and so a HE classification should aim to have a relatively even distribution of this age range between the final cluster assignments. The selection of variables was detailed earlier, although it is also appropriate to investigate the most appropriate value of *k*. Figure 6 shows that the average *R*-squared statistic does not show any discrete jumps in performance following successive changes in the number of clusters, and thus that there is no optimum value of *k* in terms of model parsimony. 10,000 separate cluster analyses were run for k=50 through k=65. These values of *k* are in a similar range to the finest level of aggregation of the Mosaic 2001 geodemographic product. The median, minimum and maximum *R*-squared results are presented in Figure 8.

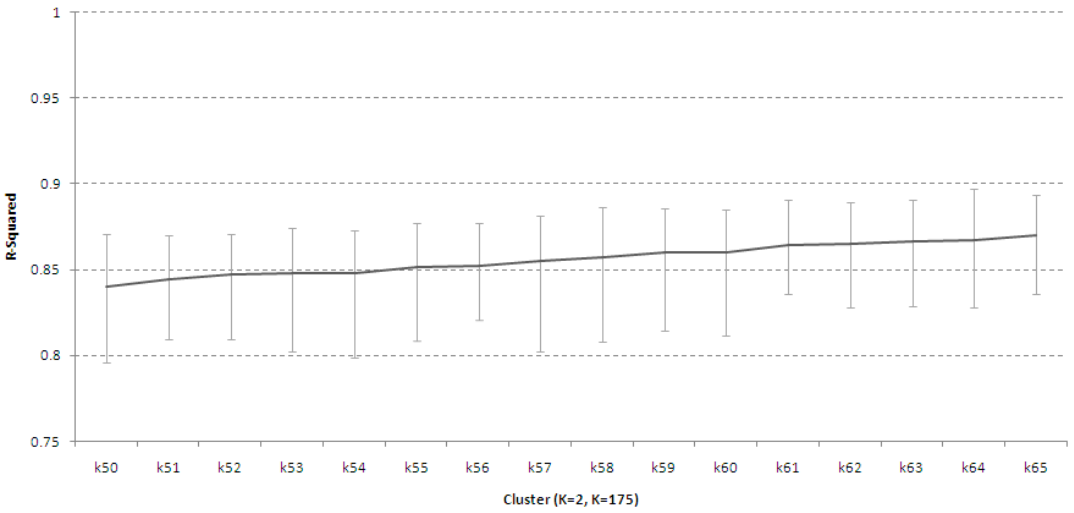


Figure 8: R-squared results from the 10,000 iterations of cluster analyses (the continuous dark line charts the median R-squared value and the whiskers link minimum and maximum values).

Each of these assignments of *k* appears to be successful in discriminating within the input data matrix and, as demonstrated in the earlier exploratory analysis, the minimum and maximum bars further illustrate the need to optimise each *k* allocation. The total 18-19 year old population from the 2001 Census were aggregated by *k*=50 to *k*=65 cluster models in order to ensure that no outliers of this key target population had been created in the clustering process. The model demonstrating the most even distribution of 18-19 year olds across the new clusters was *k*=53 (see Figure 9) and as such was chosen as the final model. In practice, the distribution of 18-19 year olds is nevertheless strongly skewed. The solid line shown in Figure 9 at 1.88% divides the principal applicant cohort equally between the 53 clusters (i.e. 100/53). Uneven distribution of household and population counts is characteristic of most geodemographic classifications: in the Mosaic classification, for example, the allocation of households to the 61 Mosaic Types ranges from 0.17% - 3.82%.

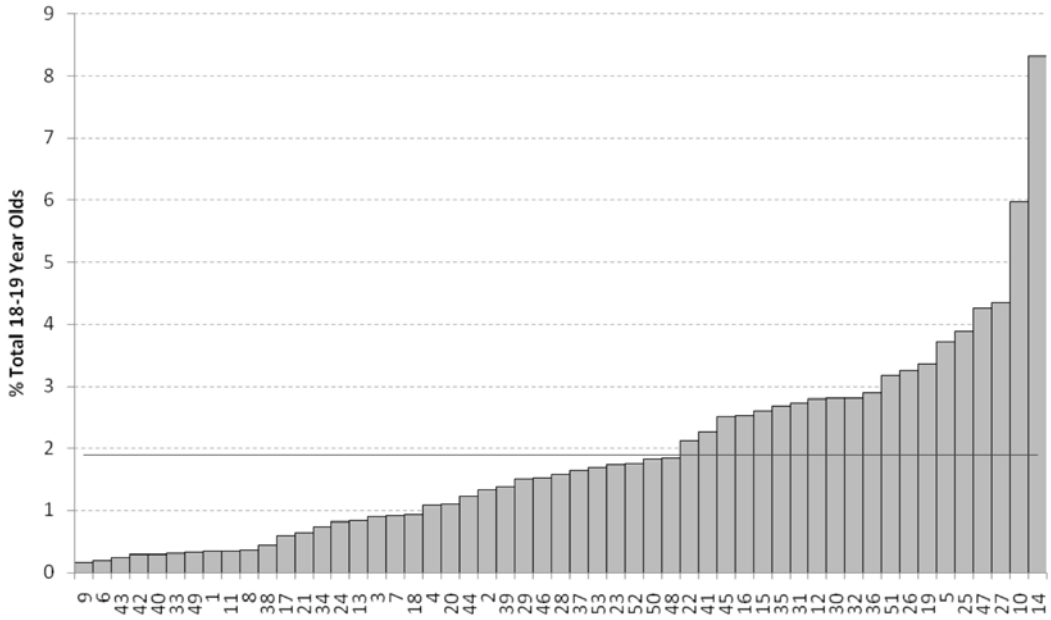


Figure 9: Distribution of 18-19 age cohort between geodemographic clusters (k = 53).

Our educational classification is therefore defined as comprising of 53 clusters (henceforth referred to as Types). However, as discussed earlier, it is often useful to have a second tier in a classification into which the Types fit hierarchically (henceforth referred to as Groups). The Ward hierarchical clustering algorithm (Ward, 1963) measures changes in variance or 'information loss' and was used to aggregate the 53 Types into Groups. Information loss by this merging procedure is defined by an error sum of squares criterion (ESS), which measures the total sum of the squared deviation for all variables from each of the 53 Types to the means of the clusters to which they might be assigned. At each step in this process the algorithm iterates through all possible unions for the 53 Types, and at each pairing an assessment is made to identify the increase in the ESS. The union that results in the smallest increase in ESS is actioned, and the process continues through further iterations until all 53 Types have been progressively assimilated into a single cluster. The hierarchical organisation of Types into Groups can have multiple arrangements depending on the frequency of Groups required. The performance of these Group classifications for predicting a target variable (e.g. participation) will depend upon the level of correlation between the variables used in cluster analysis and the target variable. Harris et al (2005) suggest that Group level classifications should ideally have populations no lower than 4% and no greater than 20%, and should also contain between 2 and 7 constituent Types. Using these guidelines the Ward method produced the classification shown in Figure 10.

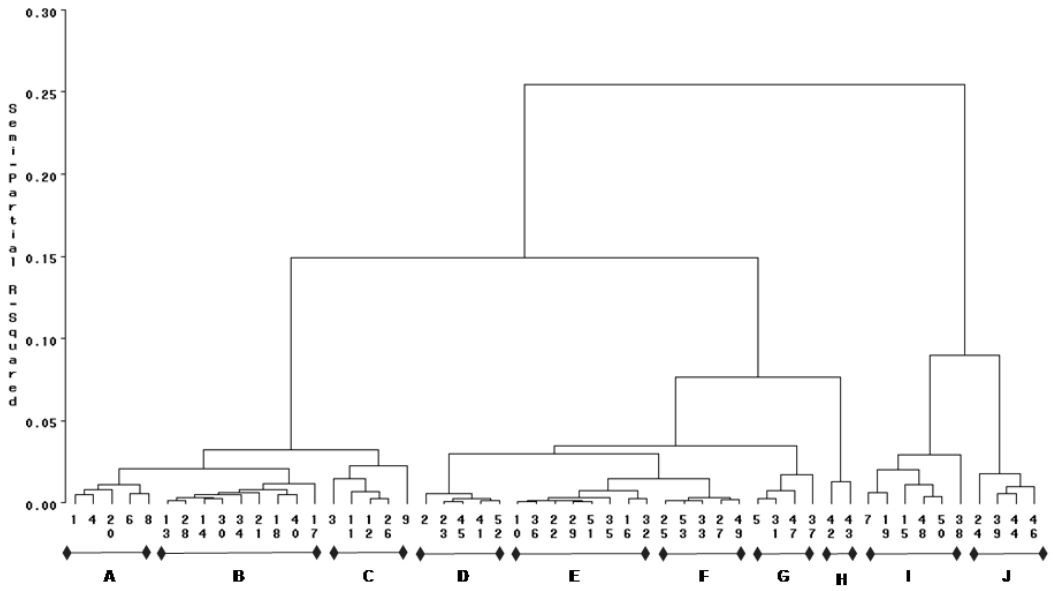


Figure 10: Dendrogram showing derivation of clusters.

5. An illustrative case study: Gospel Oak in North London.

We illustrate how our bespoke classification discriminates between OAs, with respect to the Gospel Oak area of North London that comprises a full range of neighbourhoods ranging from the very affluent to the very deprived. A criticism of the standard OAC geodemographic classification has been its performance within London, where large areas are assigned to the umbrella "Multicultural" Supergroup. These assignments apparently fail to discriminate the more subtle characteristics of the people living within these areas, and for this reason, this issue provides a useful test bed for our application specific classification. More important, however, is the impact that the cocktail of standard census variables and even more refined measures of affluence is likely to have upon attitudes towards human capital formation, choice of vocational versus academic subjects, and so forth. The census, and indeed most all commercial classifications, contain only legacy information about participation of past generations of students, and is not differentiated according to institution or programme of study.

Figure 11 shows the distribution of the educational OAC Group assignments by OA. Using these data one can explore some of the patterns that have emerged in the classification for North London. When the Gospel Oak Ward is examined it can be seen that the Output Areas to the North and South West are mainly categorised as belonging to Group G, with the remaining OA in Group I. These divisions seem

to reflect the geographical distribution of affluence. Measures of affluence are not included in the 2001 Census and as such are not included in the OAC classification, although they are in the commercial Experian Mosaic™ product (see Figure 12). Comparison of these figures suggests that one effect of introducing additional variables which correlate with wealth and educational opportunity (Singleton, 2007), such as educational performance and participation, is that we begin to highlight the spatial variation in these dimensions which may be hidden in classifications utilising only census data.

Figure 11: Educational OAC Groups in North London.

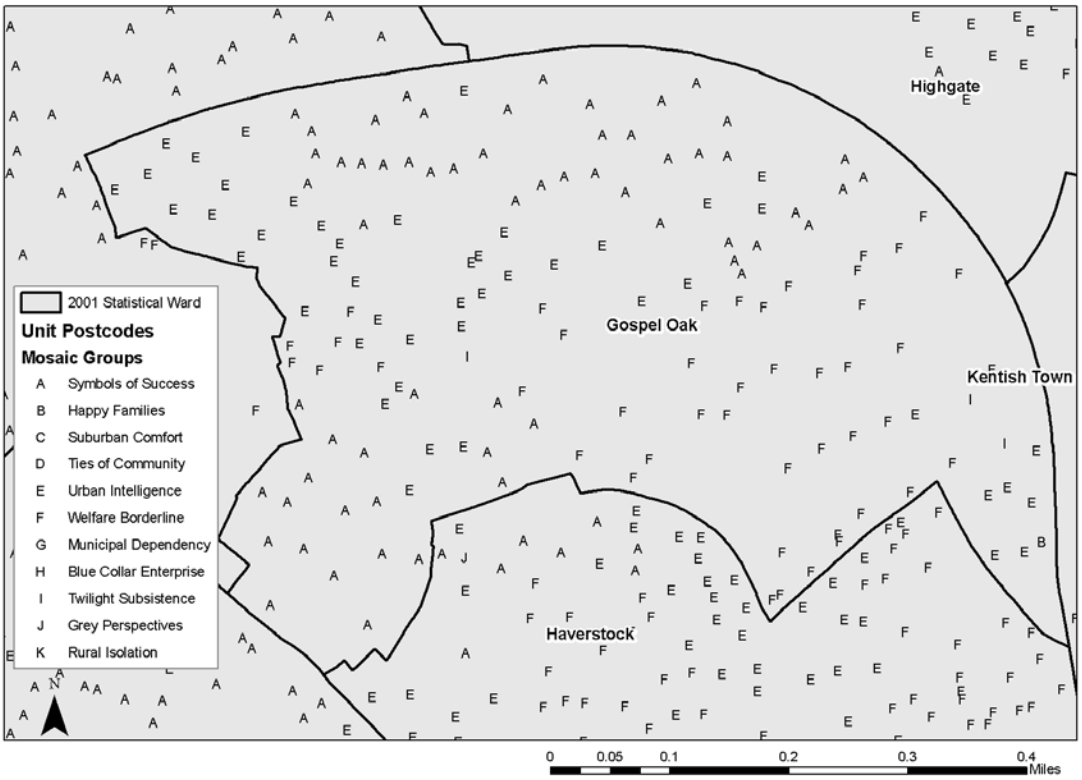


Figure 12: Experian Mosaic™ Groups in North London.

In order to go one step further and examine the HE characteristics of these areas, the educational Group level classification was appended to UCAS acceptance data for 2002 - 2004 by georeferencing the home unit postcodes of accepted applicants to the educational classification at the Output Area scale. Index scores for educational groups were calculated using Equation (1) for the following variables, where propensity refers to the extent that a target variable is overrepresented within a Group when compared to the total population:

- Propensity for course level participation.
- Propensity to attend a Russell Group[3] institution

The index scores for the first of these variables are shown in Figures 13 and 14. The Joint Academic Course Coding System (JACS) is a hierarchical classification of course types that has been used by UCAS and HESA since 2002 to classify courses of study into a fine level of 1281 'Lines' which aggregate up into 19 'Groups' (UCAS, 2007). Examining the Groups present in Gospel Oak, Figure 13 illustrates the variable propensity to participate across the range of JACS course Groups within Educational OAC Group G and Figure 14 shows the same data but for Group I. The data used to calculate these index scores are taken from the total population of participants to single honours JACS courses during 2002 - 2004 as classified by UCAS acceptances. The course level participation rates

differ markedly between these two groups, with neighbourhoods belonging to Group G showing a much higher propensity to supply medical students, for example. Group G has an index score of 150 with respect to acceptances of places at Russell Group Institutions, whereas Group I has a score of just 67.

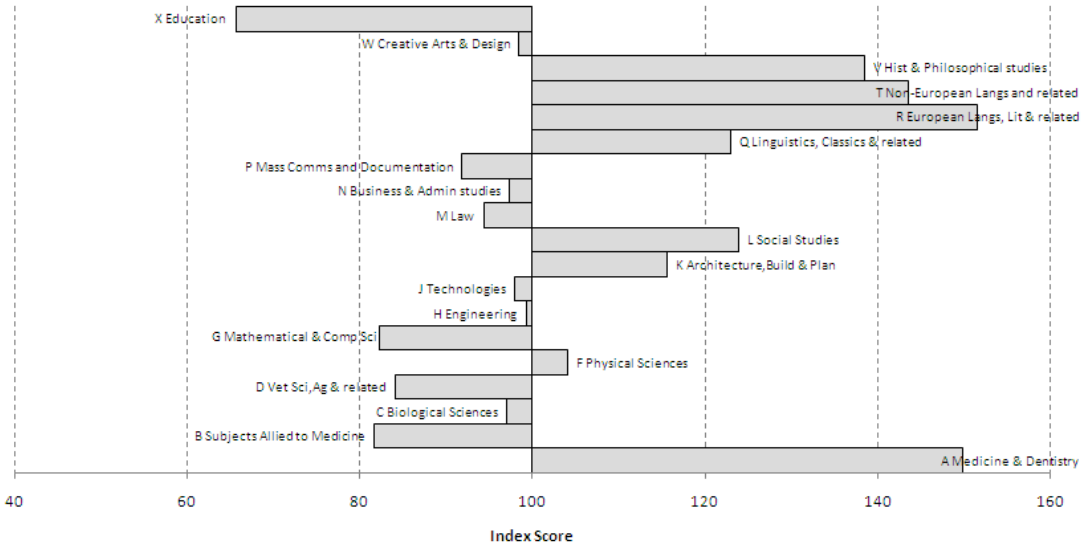


Figure 13: Propensity to accept HE places by course type in Educational OAC Group G.

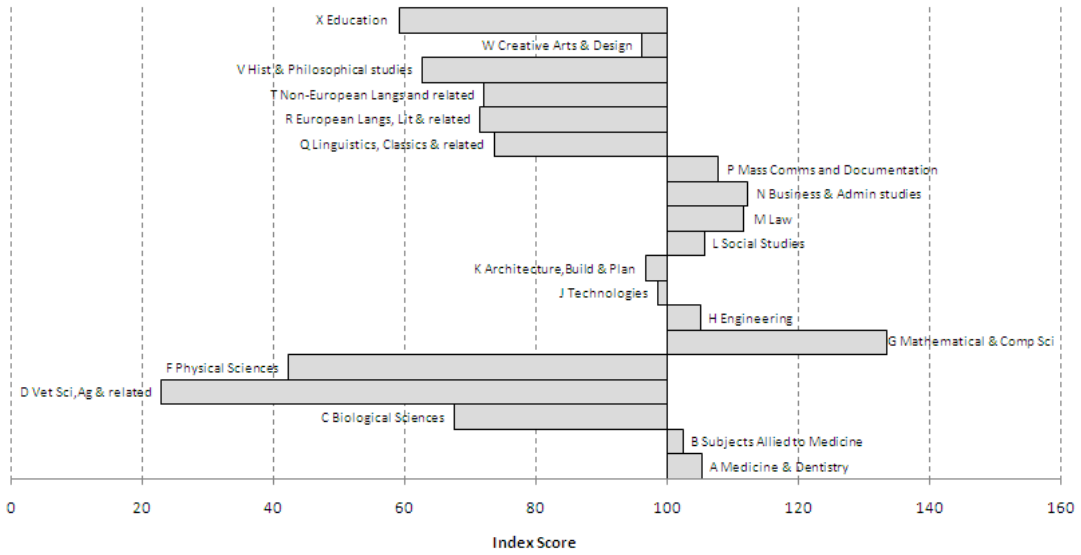


Figure 14: Propensity to accept HE places by course type in Educational OAC Group I.

The applicability of using index scores created from a national dataset to predict local variation of behaviours between areas can be assessed by comparing the predicted rates one would expect within an area against those that actually occur. The Wards shown in Figure 11 and 12 are all from within the London Borough of Camden. Within Camden during 2004 there were a total of 969 people attending

HE and studying single honours degree courses from within the main JACS groups A-X. These are distributed across eleven educational OAC Types (See Table 6).

Table 6: Distribution of Gospel Oak students between Educational OAC Types.

Educational OAC Types	*E26*	*G39*	*G41*	*I44*	*I45*	*I46*	*I47*	*I48*	*I49*	*J53*	*G38*
Sum	*3*	*15*	*150*	*112*	*220*	*242*	*144*	*16*	*53*	*2*	*12*
Mean	*0.16*	*0.79*	*7.89*	*5.89*	*11.58*	*12.74*	*7.58*	*0.84*	*2.79*	*0.11*	*0.63*

Expected values of what one would expect if all students studying JACS courses were distributed evenly across all educational OAC Types can be devised by dividing the total population within these groups by the total number of JACS Group (19) - for example, Type I46 is expected to have 12.74 students in Camden (=242/19). Differences between observed and expected numbers of students can be calculated by taking the student average (expected value) within each of the Educational OAC Groups and multiplying it by the index scores, i.e. the difference from the average (see Table 7).

Table 7: Predicted minus observed scores.

	E26	***G39***	***G41***	***I44***	***I45***	***I46***	***I47***	***I48***	***I49***	***J53***	***G38***
A Medicine & Dentistry	0.2	1.0	-1.2	0.3	3.4	-5.1	1.2	0.8	0.5	0.2	1.1
B Subjects Allied to Medicine	0.2	-0.3	0.3	-2.8	-5.0	2.0	-4.5	0.9	-1.9	0.1	0.5
C Biological Sciences	0.1	0.8	-4.4	-3.9	-6.4	-5.9	-7.6	-0.3	0.8	0.1	0.6
D Veterinary Sci., Agric. & Related	0.1	0.6	3.6	0.2	0.3	4.9	0.8	0.3	0.4	0.0	0.6
F Physical Sciences	0.1	0.8	0.3	0.6	1.6	0.0	-2.0	0.4	1.0	0.1	-0.3
G Mathematical & Comp. Science	-0.9	-0.3	1.6	2.3	-0.8	-1.9	-8.0	0.2	1.1	0.2	0.6
H Engineering	0.1	0.8	2.7	2.2	2.2	6.6	5.2	0.9	1.1	0.1	-0.3
J Technologies	0.2	0.7	8.7	6.9	10.4	13.4	4.1	0.5	1.1	0.1	0.6
K Architecture, Build. & Planning	0.1	-0.2	3.4	5.7	9.0	9.3	1.3	-0.3	1.3	0.1	-0.3
L Social Studies	-1.8	-2.1	-14.5	-2.6	-9.7	-6.1	1.0	-2.2	-1.3	-0.9	-4.2
M Law	0.1	-0.3	-0.9	1.7	1.6	-5.2	0.0	-0.1	0.8	-0.9	0.7
N Business & Admin. Studies	0.1	-1.3	-1.1	-5.8	-29.0	-17.8	-13.6	1.0	-8.5	0.2	-1.3
P Mass Comms. and Documentation	0.2	0.8	3.8	-0.3	4.8	0.5	5.2	0.0	2.7	0.1	-0.4
Q Linguistics, Classics & Related	0.2	0.0	-0.7	-5.3	-1.8	-1.4	-1.3	-2.4	-0.2	0.1	0.7
R European Lang., Lit & Related	0.2	0.2	10.9	3.8	3.4	6.1	2.8	0.3	0.1	0.1	0.9
T Non-European Lang. and Related	0.2	0.9	9.9	6.7	2.6	10.6	1.5	0.5	1.4	0.1	0.9
V Hist. & Philosophical Studies	0.2	0.0	-5.3	-3.5	-1.8	2.7	-5.4	0.4	0.1	0.1	0.8
W Creative Arts & Design	0.2	-2.2	-15.0	-14.8	-30.8	-30.1	-15.2	-3.2	-8.9	0.1	-0.4
X Education	0.1	0.6	3.0	1.4	4.2	4.7	3.8	-0.3	-0.6	0.1	0.4

6. Concluding comments

The differentiation according to subject and HE institution type in our Camden case study suggests that there is clear value in using bespoke geodemographic indicators to predict course choice. *A priori* one would not expect these differences to be identified in such sharp relief by discriminators based upon census variables alone, or upon data derived from the consumption of goods and services. However, we

suggest that this is only the starting point for the development of geodemographic discriminators that are tuned to the requirements of public service providers. This case study suggests that there are also externalities which are either not adequately modelled by this classification, or that arise because of local variations which are missed by using index scores based on a national datasets. The most obvious of these local variations identified in this limited test is the systematic under prediction of "N: Business and Administration studies" and almost all of "W: Creative Arts and Design" subjects across all Educational OAC Types present in Camden. Local externalities which may have induced these errors could include, for example, a local school/ college with specialisms in these subjects, or the existence of prestigious local institutions with strong outreach links or sponsorship arrangements. In either a revised classification or through the creation of locally weighted index scores, this technique should prove very useful to a range of end users to model potential markets to target. An example could be an HE institution wishing to target recruitment for a particular course in a selection of schools with a known demographic.

Nevertheless, this paper has demonstrated a method by which bespoke classifications for a particular sector or application can be created using pertinent public sector data sources. The motivation for this analysis lies in the observation that typologies created by commercial classification providers supply no evidence to justify why the inclusion of data relating to private consumption of goods is appropriate for predicting public consumption. Furthermore, the exact nature of the weighting schemes and data used to derive such commercial classification systems is closed to the public, which should be of concern to public services that may apportion real life chances, rather than simply provide consumer products and services. While other interesting research has sought to make commercial geodemographic classifications relevant to public service provision (Batey and Brown, 2007; Ashby and Longley, 2005), we believe that the addition of higher education sector data is seen as a positive step beyond use of generic and re-labelled classification for purposes which they were not originally designed. As such, this research presents a challenge to the implied assumption that the nature of individual use of public services such as education should directly correspond with the ways in which consumers use private goods and services. This work also responds to concerns that the data inputs used to create generic commercial geodemographic classifications come from disparate private sector and closed sources, that their provenance is often unknown, and that the assumptions used to create such classifications cannot

be scrutinised or tested by end users. The negative potential social implications of using such classifications in areas of public service provision should not be under-estimated, since they potentially significantly impact upon the life chances of stakeholders in public services.

The methodology has shown how a classification built using the 2001 Census can be refined for a specific purpose through the augmentation of sector specific data. Through an illustrative example of using the classification to predict course participation rates within a diverse Ward in London, it has demonstrated problems in using national index scores alongside geodemographic groups to predict phenomena on a local scale. Future work is required to examine the causes of such local variation and assess how they might be incorporated into the data model. Furthermore, should any model be disseminated amongst the wide range of potential end users (e.g. schools, universities, colleges, local education authorities), a method of creating both nationally and regionally variable descriptive material to accompany the clusters should be devised to allow for more accurate profiling relevant to local geographical area. The broader challenge to regional science is to assimilate these rich descriptive indicators of revealed preferences for courses and higher education institutions with systematic analysis of student flows to the different HE institutions within the national (and indeed increasingly international) system (Wilson, 2000).

References

Ashby DI, Longley, PA (2005) Geocomputation, geodemographics and resource allocation for local policing. *Transactions in GIS* 9(1): 53–72

Batey PWJ, Brown PJB (2007) The spatial targeting of urban policy initiatives: a geodemographic assessment tool. *Environment and Planning A*: 39, in press

Bourdieu, P., Passeron, J.C. (1977) Reproduction in Education, Society and Culture. London: Sage.

Brunsdon C, Charlton M (2006) Local trend statistics for directional data - A moving window approach, *Computers, Environment and Urban Systems* 30: 130-42

Corver M (2005) *Young participation in higher education* (online). Available from: http://www.hefce.ac.uk/pubs/hefce/2005/05_03/05_03.pdf [Accessed 28th April 2007]

Debenham J (2001) Understanding geodemographic classification: creating the building blocks for an extension (online). *The School of Geography, University of Leeds Working Paper*. Available from: http://www.geogleeds.ac.uk/wpapers/02-1.pdf [Accessed 1st July, 2006]

Debenham J, Clarke G, Stillwel, J (2002) Deriving new variables to extend geodemographic classification (online). *CyberGeo*. Available from: http://www.cybergeopresse.fr/ectqg12/stillwell/stillwell.htm [Accessed 25th May, 2007]

DfES.(2003) The Future of Higher Education. CM5735. London: HMSO.

Everitt B, Dunn G (1983) *Advanced methods of data exploration and modelling.* London: Ashgate

Everitt B (1974) *Cluster analysis* London: Heinemann Educational Books

Gilchrist R, Philips D, Ross A (2003) Participation and potential participation in UK higher education In: Archer L, Hutchings M, Ross A (eds) *Higher Education and Social Class.* London: RoutledgeFalmer

Gordon AD (1981) *Classification: Methods for the exploratory analysis of multivariate data*. Norwell: Kluwer Academic Publishers

Harris R, Sleight P, Webber R (2005) *Geodemographics, GIS and Neighbourhood Targeting.* London: Wiley

HEFCE (2005) *POLAR: A short guide* (online). Available from: http://www.hefce.ac.uk/widen/polar/guide/ [Accessed 1st July, 2006]

HESA (2006) *Higher Education Statistics Agency* (online). Available from: http://www.hesa.ac.uk/ Cheltenham: HESA [Accessed 10th June, 2006]

Huang JZ, Ng M J, Rong H, Li Z (2005) Automated variable weighting in *k*-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(5): 657-668.

Leathwood, C., Hutchings, M. (2003) Entry Routes to Higher Education. In L. Archer, M. Hutchings, A. Ross (eds) Higher Education and Social Class. London: Routledge Falmer.

MacQueen JB (1967) Some Methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability Berkeley*, pp 281-297

Reay D, Miriam DE, Ball S (2005) *Degrees of choice: social class, race and gender in Higher Education.* Staffordshire: Trentham Books

Reid I (1998) *Class in Britain*. Cambridge: Polity Press

Robbins LC (1963) *Higher Education: report of the committee appointed by the Prime Minister under the chairmanship of Lord Robbins 1961-3*. London: HMSO

Sa R, Florax RJGM, Rietveld P (2003) Determinants of the regional demand for higher education: A gravity model approach. *Tinbergen Institute Discussion Paper* 2003-013/3

Singleton, AD, Farr, M (2004) Widening access and participation in higher education. *Proceedings of the GIS Research UK 12th Annual Conference*, pp 265-267

Singleton A (2007) *A spatio-temporal analysis of access to higher education*. Unpublished Thesis. London: University of London

UCAS (2007) *Data management: UCAS course coding* (online). Available from: http://www.ucas.ac.uk/higher/courses/coding.html [Accessed 28th April 2007]

Vickers, D., Rees, P., Birkin, M. (2005) Creating the National Classification of Census Output Areas: Data Methods and Results. [online]. The School of Geography, University of Leeds Working Paper. Available from http://www.geog.leeds.ac.uk/wpapers/05-2.pdf [Accessed 21st August, 2006]. Leeds: University of Leeds.

Vickers D, Rees P (2007) Creating the UK national statistics 2001 output area classification. *Journal of the Royal Statistical Society, Series A* 170(2): 379-403

Voas D, Williamson P (2001) The diversity of diversity: a critique of geodemographic classification. *Area* 33(1): 63-76

Walker, M. (2003) Framing Social Justice in Education: What Does the Capabilities Approach Offer? British Journal of Educational Studies. 51(2), 168-187.

Ward, J.H. (1963) Hierarchical grouping to optimise an objective function. Journal of the American Statistical Association. 58, 236-234.

Wilson AG (2000) The widening access debate: student flows to universities and associated performance indicators. *Environment and Planning A* 32: 2019-31

Endnotes

[1] http://www.p2peopleandplaces.co.uk/

[2] Full details on the implementation of the *k* means algorithm in SAS can be found at: http://v8doc.sas.com/sashtml/stat/chap27/

[3] The Russell Group is association of leading UK research-intensive Universities whose membership include: University of Birmingham, University of Bristol, University of Cambridge, Cardiff University, University of Edinburgh, University of Glasgow, Imperial College London, King's College London, University of Leeds, University of Liverpool, London School of Economics & Political Science, University of Manchester, Newcastle University, University of Nottingham, Queen's University Belfast, University of Oxford, University of Sheffield, University of Southampton, University College London, University of Warwick