

The stability of geodemographic cluster assignments over an intercensal period

Alexander Singleton¹ · Michail Pavlis¹ · Paul A. Longley²

Received: 30 July 2014 / Accepted: 15 January 2016 / Published online: 22 March 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract A geodemographic classification provides a set of categorical summaries of the built and socio-economic characteristics of small geographic areas. Many classifications, including that developed in this paper, are created entirely from data extracted from a single decennial census of population. Such classifications are often criticised as becoming less useful over time because of the changing composition of small geographic areas. This paper presents a methodology for exploring the veracity of this assertion, by examining changes in UK census-based geodemographic indicators over time, as well as a substantive interpretation of the overall results. We present an innovative methodology that classifies both 2001 and 2011 census data inputs utilising a unified geography and set of attributes to create a classification that spans both census periods. Through this classification, we examine the temporal stability of the clusters and whether other secondary data sources and internal measures might usefully indicate local uncertainties in such a classification during an intercensal period.

Keywords Geodemographics · Output Area Classification · Cluster analysis · Uncertainty

JEL Classification J110 · C89 · R23

✉ Paul A. Longley
p.longley@ucl.ac.uk

¹ Department of Geography and Planning, University of Liverpool, Liverpool L69 3BX, UK

² Department of Geography, University College London, Gower Street, London WC1E 6BT, UK

1 Introduction

Geodemographic indicators are composite measures describing the socio-spatial structure of small geographic areas (Harris et al. 2005). They are typically used: to describe and hypothesise about processes of residential differentiation (e.g. Reibel 2011); to explore over- or under-represented behaviours exhibited between neighbourhood types (e.g. Singleton 2010); or to establish a basis to marketing activity or resource allocation in the private or public sectors, respectively (e.g. Singleton and Spielman 2014). Although geodemographic classification have received criticism (e.g. Goss 1995), they have developed and sustained a reputedly robust pedigree (Birkin et al. 2002) in both the public and private sectors (Longley 2005), with numerous successful areas of application including health (Petersen et al. 2011), retail (Thompson et al. 2012), education (Singleton et al. 2012), planning (Batey and Brown 2007) and policing (Ashby and Longley 2005).

Like their historical antecedents of social area analysis and factorial ecology (Timms 1971; Rees 1972), many modern geodemographic classifications are created entirely from cross-sectional census data. Although research has illustrated how such classifications could be updated and created in “real time” (Adnan et al. 2010), these methods have yet to enter mainstream use, potentially because of the computational complexity associated with their implementation (particularly when repeated iterations on large data sets are required in order to ensure a stable cluster solution), and the extent to which end-users are willing to engage in the process of classification creation.

The USA and the UK have particularly well-developed geodemographic markets, and here the majority of classifications are supplied by the commercial sector (Singleton and Spielman 2014), usually without comprehensive metadata and documentation of the techniques used. In addition, high licensing fees for the classifications (or composite data) can preclude their use by many potential end-users (Singleton and Longley 2009), and classifications may not pass the scientific requirement that they are reproducible by other researchers. In response to such issues, an alternate model of “open geodemographics” has been developed within some jurisdictions, where both the data and methods are all within the public domain. For example, in the UK, the 2001 output area classification (OAC) (Vickers and Rees 2007) became the most widely used open-source geodemographic classification and has now been updated using 2011 UK census data collected by the UK Office for National Statistics (ONS).¹ The variables making up this classification are selected to cover a series of domains and subdomains designed to provide a balanced general-purpose picture of the social, economic, built and population structure of the UK, with the final selection of variables also guided by the desire not to include variables that were strongly correlated with one another. Both the 2001 and 2011 OACs were created using the widely used *k*-means clustering procedure (Everitt et al. 2011), which is applied by iteratively assigning each UK census output area to one of a pre-specified number of clusters, so as to minimise the

¹ <http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/index.html>.

overall squared Euclidean distance of every area's attributes to their nearest cluster mean.

An important distinction between the 2001 and 2011 OACs and commercial geodemographic classifications is that the latter typically, although not universally, incorporate non-census data that may be collected more frequently, thus unshackling classification building in the UK from the ten yearly census cycle. The 2001 and 2011 OAC methodologies contained no mechanism through which accumulated uncertainty in the potential reliability of its cluster assignments could be assessed over time, and the remit for the 2001 and 2011 OAC projects did not allow for the classification to be reviewed and possibly updated.

The overarching aim of the analysis presented here is to explore how changes in the values of the census variables for given locations result in reassignment of geodemographic class over time, with particular focus upon the variables that are used in the open 2001 and 2011 OAC classifications. By examining the nature and patterning of change, we can provide an assessment of the degradation of accuracy in a census-based classification over an intercensal period and an assessment of the extent to which this is acceptable for typical applications. Thus, for this analysis we pool 2001 and 2011 census data using a common geography and set of attributes to create a classification that spans both census periods. The basic classification methodology broadly follows that of the 2011 OAC, which itself was an adaption of the classification of 2001 census data devised by Vickers and Rees (2007). Through our unified 2001–2011 temporal OAC, we examine the stability of the clusters obtained from the pooled data over time. We then discuss, in general terms, the nature of the changes that have occurred to the geodemographic structure of the UK over the period 2001–2011. Finally, as part of our discussion of the results, we also investigate how other secondary data sources might be used to accommodate measures of change during the intercensal period.

Any summary representation of the social similarities that characterise scattered neighbourhood areas is, however, necessarily incomplete, and the mix of variables that are included in general-purpose classifications are the outcome of choice, convention and chance. Voas and Williamson's (2001) prescient discussion of the importance of place in geodemographics focuses on the diversity of social attributes that occurs within neighbourhood clusters, but their conclusion that "while taxonomy has its uses, it is of little use in producing complete descriptions of particular areas" (p. 64) is unhelpful to the wide range of organisations in business, government and research who nevertheless continue to find them useful in characterising and comparing areas (Singleton and Spielman 2014). In what follows, we thus take as given that the socio-economic indicators that underpin geodemographic classifications remain sufficiently stable over an intercensal period to allow pooling of data from different time periods, and that the established procedures of cluster analysis provide a valid way of identifying the social similarities that characterise different neighbourhoods.

Table 1 Summary of the census input attributes

Domain	Subdomain	Variables
Demographic	Age structure	Age 0–4; age 5–14; age 25–44; age 45–64; age 65–89; age 90+
	Family structure	Single; married or in a registered same-sex civil partnership; divorced or separated; no children household; non-dependent children
	Ethnicity	White; mixed; Indian; Pakistani; Bangladeshi; Chinese; Black/African/Caribbean; UK and Ireland; Other EU
Housing	Composition	Population density; lives in communal establishment; occupancy rating; all students
	Type	Detached; semi-detached; terrace; flats
	Tenure	Owned or shared ownership; socially rented; private rented
Socio-economic	Health	Standardised illness ratio; unpaid care
	Employment	Schoolchildren and full-time student; unemployed; part-time; full-time
	Education	Levels 1 and 2; level 3; level 4+
	Mobility	Car ownership; public transport; private transport; active transport
	Occupation	Agriculture, forestry and fishing; mining, quarrying and construction; manufacturing; energy, water and air conditioning supply; wholesale and retail trade; transport and storage; accommodation and food service activities; financial, insurance and real estate activities; public administration and defence; education; human health and social work activities

2 Building a composite temporal 2001–2011 output area classification

The first stage of our analysis was to assemble a database of census variables that were collected in the censuses of both 2001 and 2011. Sixty attributes that were used to build OAC 2011 were selected initially, and then refined to a smaller subset of 55 variables that were also available in 2001, and with a consistent set of definitions (Table 1). To aid this process, the analysis was also restricted to England, given the different and changing remits of the census between UK countries and time periods. As such, our specification of variables was essentially guided by issues of data availability, but we believe that the small deviation from the 2001 and 2011 OAC specifications maintains the general-purpose nature of the hybrid analysis that follows.

A largely common set of output area zones was used, each containing an average of 300 people and 130 households in 2011, and are the smallest zonal geography for which comprehensive census attributes are released in the UK. The vast majority of the 2011 zones are the same as those used in the 2001 census outputs, with approximately 2.6 % of zones formed from either merging or splitting to reflect underlying population changes during the intercensal period.² Local changes to census geography are, of course, an indicator of likely change in social, economic

² For specific details on the criteria for change, see the official Office for National Statistics (ONS) document: <http://www.ons.gov.uk/ons/guide-method/geography/products/census/report-changes-to-output-areas-and-super-output-areas-in-england-and-wales-2001-to-2011.pdf>.

and demographic circumstances, and so changes in local census administrative geography are itself an indicator of change. In the analysis that follows, we supplement the evidence from changing administrative geography with evidence of changing geodemographic characteristics of census areas throughout the study area. The target zonal geography used for this analysis was that of the 2011 output areas, identified using a lookup table available from the Office for National Statistics.³ The process of reconciling 2001 data with the complete set of 2011 boundaries entailed summation of constituent zones that had been merged in 2011, and apportioning 2001 zone totals proportionately to area for 2001 zones that had been divided for purposes of the 2011 census.

The 2001 and 2011 census input data were thus rendered compatible with the 2011 output area geography, resulting in two records for each area: one for 2001 and one for 2011. For each area, inputs were calculated as percentages, with the exception of population density and standardised limiting long-term illness. As with the creation of the 2011 OAC, the data were then transformed using an inverse hyperbolic sine transformation (Johnson 1949), in order to return inputs that were more normally distributed, with the aim of aiding cluster identification using *k*-means. Prior to clustering, the data required standardisation onto the same scale and so, in common with the 2011 OAC, variables were range-standardised onto a 0–1 scale. The final input data set comprised 55 variables and 342,744 output area records, equal to twice the number of 2011 output areas within England.

The input data were assembled in the statistical programming language R (R Core Team 2013), which was then used to run the *k*-means algorithm 10,000 times in order to identify an optimal and robust partitioning of the areas into an 8-cluster solution. Repeated runs are necessary as the algorithm outputs are sensitive to the initial seeding of the *k* clusters. Eight classes were identified as a parsimonious solution that also matched the 2011 OAC.

An alternative method might have been to cluster two separate classifications for 2001 and 2011, akin to developing a standard census geodemographic system for each. However, such classifications would be optimised against a different distribution of input values, and the aim here was to establish linked clusters drawn from the same input data and then to use these to examine how the relationship between areas and the cluster means changed over time. This was made possible because a common set of attributes were available for both 2001 and 2011. With this method, the clusters might be conceptualised as an optimised assignment of areas into groups derived from an average of the two time periods; i.e. the clusters represent the best fit for the whole time period, rather than on the census nights of either 2001 or 2011.

Once the common classification was created, this could be split post-clustering and mapped for the two time periods. For the purpose of this analysis, we only created a single tier ('Super Group') classification in order to map those main cluster changes between 2001 and 2011.

³ [https://geoportal.statistics.gov.uk/Docs/Lookups/Output_areas_\(2001\)_to_output_areas_\(2011\)_to_local_authority_districts_\(2011\)_E+W_lookup.zip](https://geoportal.statistics.gov.uk/Docs/Lookups/Output_areas_(2001)_to_output_areas_(2011)_to_local_authority_districts_(2011)_E+W_lookup.zip).

In common with the 2001 and 2011 OACs, we also created short descriptions of the clusters, and assigned representative labels to aid end-user interpretation of the main cluster characteristics. There are multiple ways in which this process can be accomplished, and our preferred technique was to calculate a “grand index”, representing the deviation of the classification input attributes within each cluster away from their national representation in the pooled data sets for 2001 and 2011. Such scores are typically standardised so that 100 would represent the national average over the entire data set, 200 a rate of double and 50 a half. Options for creating index scores included separating out the 2001 and 2011 clusters and input attributes, and calculating separately; or, combining both years together, and calculating index scores on the basis of both 2001 and 2011 inputs combined. The latter method was selected because it was the combined data that were used to form the clusters, thus also maintaining the unified approach for cluster description. The cluster index scores for the selected variables are presented in Table 2.

Cluster labels and descriptions are as follows:

Cluster 1—suburban diversity These areas are typically suburban in location, with very high ethnic diversity. Populations are typically young, and many families have dependent children. There are above average numbers of residents from newer EU countries, and crowded, privately rented terraced housing is common. Perhaps given lower rent values within these areas, they are also attractive to students. Although unemployment is higher than average, those who are in work tend to be employed in manual occupations such as warehousing, transport, accommodation and food services.

Cluster 2—ethnicity central These are areas of very high ethnic diversity, with especially high prevalence of Black and Bangladeshi residents. Many households have young children, and rates of divorce are higher than the national average. There are also high numbers of students living within these areas. The dominant housing stock is flats, with many overcrowded and rented from the public sector. Unemployment within these areas is typically high, and as might be expected given their central locations, public transport is heavily used.

Cluster 3—intermediate areas These areas have few distinctive features, apart from higher than average numbers of very elderly people living in communal establishments.

Cluster 4—students and aspiring professionals Undergraduate and postgraduate students, as well as those who are starting their careers, are over-represented within these areas. Residents are ethnically diverse, with higher than average numbers of people identifying their origins as Chinese, Indian or being born in countries that acceded to the EU prior to 2001. The dominant housing stock is flats, which are typically rented within the private sector, and there is some overcrowding.

Cluster 5—county living and retirement These rural areas are overwhelmingly White and house large numbers of people who work in agriculture, forestry and fishing. Of those not working, there are higher numbers of people who are past retirement age. Many people live in uncrowded detached houses, perhaps because children have aged and left the family home.

Cluster 6—blue-collar suburbanites These suburban areas are dominated by terraced or semi-detached housing, with a higher than average number being

Table 2 Eight cluster index scores for the input variables

Variables	Clusters							
	1	2	3	4	5	6	7	8
Age 0–4	132	132	97	73	76	109	81	111
Age 5–14	116	101	92	45	94	115	100	105
Age 25–44	113	132	102	137	79	100	85	96
Age 45–64	82	73	103	63	127	97	121	90
Age 65–89	66	60	104	62	129	94	117	113
Age 90 and over	60	58	146	107	129	64	84	131
Lives in a communal establishment	48	63	147	588	125	27	36	61
Single	120	157	95	186	69	95	66	108
Married or in a registered same-sex civil partnership	91	61	102	52	123	101	131	79
Divorced or separated	96	123	103	84	87	109	69	142
White	65	65	107	94	112	109	108	108
Mixed/multiple ethnic group	215	315	80	172	33	58	49	79
Asian/Asian British: Indian	443	119	58	130	9	30	68	25
Asian/Asian British: Pakistani	589	95	27	59	4	30	28	33
Asian/Asian British: Bangladeshi	349	666	24	76	5	22	13	38
Asian/Asian British: Chinese and other	279	287	74	296	20	28	45	41
Black/African/Caribbean/Black British	302	711	28	111	6	23	16	37
UK and Ireland	80	74	103	83	108	108	107	106
Other EU: member countries in March 2001	119	247	102	348	67	50	62	59
No children household	70	59	112	104	134	97	124	73
Non-dependent children household	109	73	89	36	102	125	120	93
Full-time student household	187	175	50	848	6	18	8	30
Whole house or bungalow: detached	32	8	104	13	237	43	221	24
Whole house or bungalow: semi-detached	94	17	96	21	92	153	137	86
Whole house or bungalow: terrace and end-terrace	168	59	106	83	52	160	22	141
Flats	102	372	94	333	18	10	8	154
Owned or shared ownership	88	40	115	66	117	103	142	64
Social rented	109	280	30	62	39	123	8	241
Private rented	140	147	113	292	87	63	34	87
Occupancy room rating –1 or less	178	351	62	267	28	57	18	114
Provides unpaid care	90	76	98	59	116	104	117	97
Highest level of qualification: level 1, level 2 or apprenticeship	90	76	103	57	102	114	109	105
Highest level of qualification: level 3 qualifications	101	94	109	200	94	82	96	75
Highest level of qualification: level 4 qualifications and above	101	129	121	181	112	54	110	49
Schoolchildren and full-time students: age 16 and over	143	148	80	310	64	72	74	69
2 or more cars or vans in household	72	22	117	46	166	86	157	51
Public transport	163	292	77	205	27	75	59	86
Private transport	74	36	109	50	118	113	127	95
On foot, bicycle or other	90	113	104	179	71	108	65	130

Table 2 continued

Variables	Clusters							
	1	2	3	4	5	6	7	8
Unemployed	147	193	72	85	56	111	52	155
Part-time	106	89	105	74	110	98	105	92
Full-time	103	98	111	109	99	95	100	87
Agriculture, forestry and fishing	17	14	47	21	513	63	49	67
Mining, quarrying and construction	88	68	100	48	114	118	104	110
Manufacturing	76	42	89	50	99	136	113	124
Energy, water and air conditioning supply	88	50	106	63	95	115	115	104
Wholesale and retail trade; repair of motor vehicles and motor cycles	105	83	94	74	92	115	99	117
Transport and storage	130	101	91	67	76	116	92	120
Accommodation and food service activities	121	153	88	138	100	93	67	121
Financial, insurance and real estate activities	87	123	101	168	87	79	114	78
Public administration and defence; compulsory social security	87	84	120	92	97	94	115	80
Education	102	98	113	108	105	80	113	73
Human health and social work activities	103	103	101	87	93	103	98	106

socially rented. Employment is most typically in manufacturing, although many other blue-collar occupations are prevalent, such as construction.

Cluster 7—professional prosperity The populations of these areas are most typically White and towards the latter stages of successful careers in a range of white-collar professional occupations. Most are married, and if they have had children, these are of an age where they are no longer dependent. Housing within these areas is typically privately owned and detached; higher incomes enable many households to sustain multiple car ownership.

Cluster 8—hard-up households These deprived and predominantly White areas feature households from a full range of age groups. Those of working age experience higher than average rates of unemployment. Employed residents work in service or manual occupations. Housing within these areas is typically terraced or flats, with some overcrowding and very high rates of renting within the social housing sector.

These descriptions are exemplified in Fig. 1, which shows the changing complexion of three English cities: Bristol in the South West of England and Liverpool and Leeds in the north. Both Liverpool and Bristol are predominantly urban areas (cluster 5, “county living and retirement” has either no or limited representation), whereas Leeds represents a much larger local authority district, complementing its urban core with more extensive hinterland and rural areas. There is no radical change in the assignments over the 2001–2011 period, in large part because the systems of property ownership and planning control preclude this. That said, Liverpool in particular has seen significant redevelopment and housing

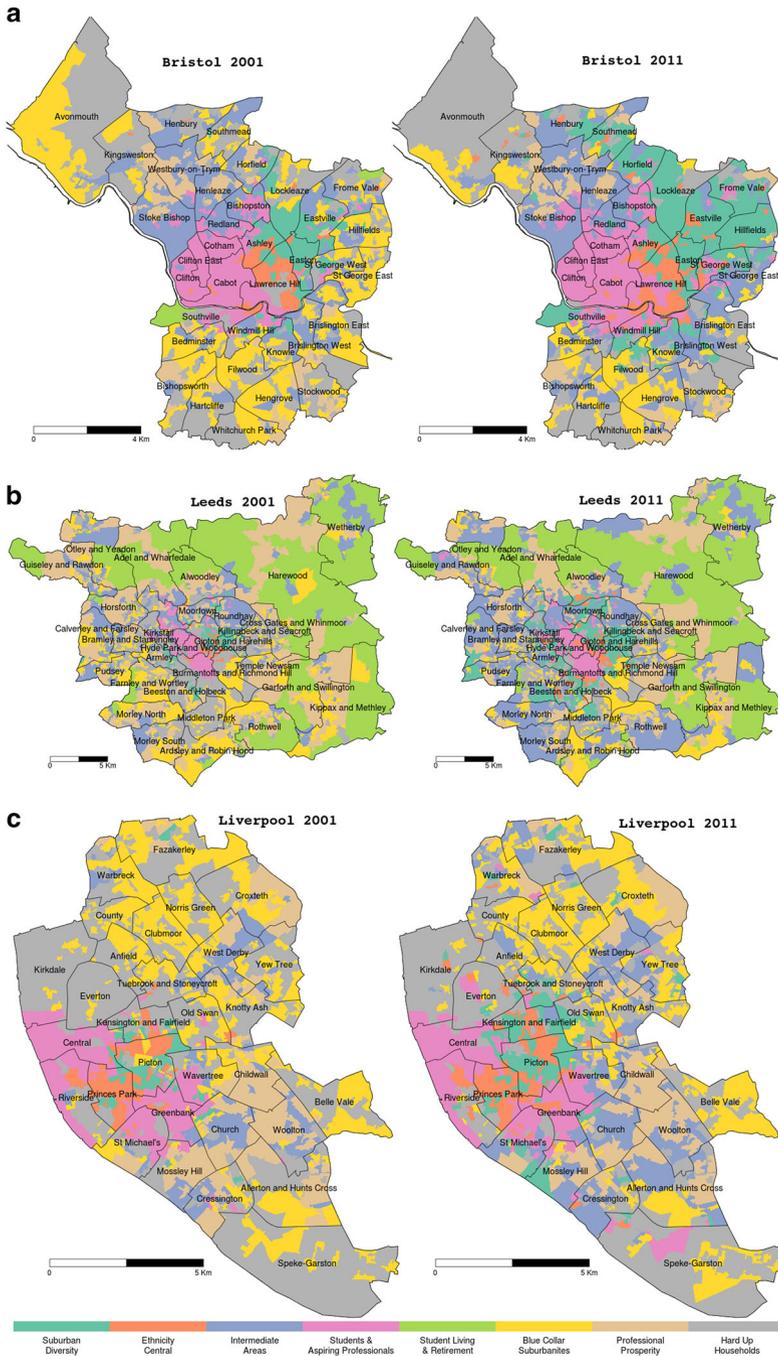


Fig. 1 Cluster results using unified OAC for three UK cities in 2001 and 2011: **a** Bristol; **b** Leeds; and **c** Liverpool

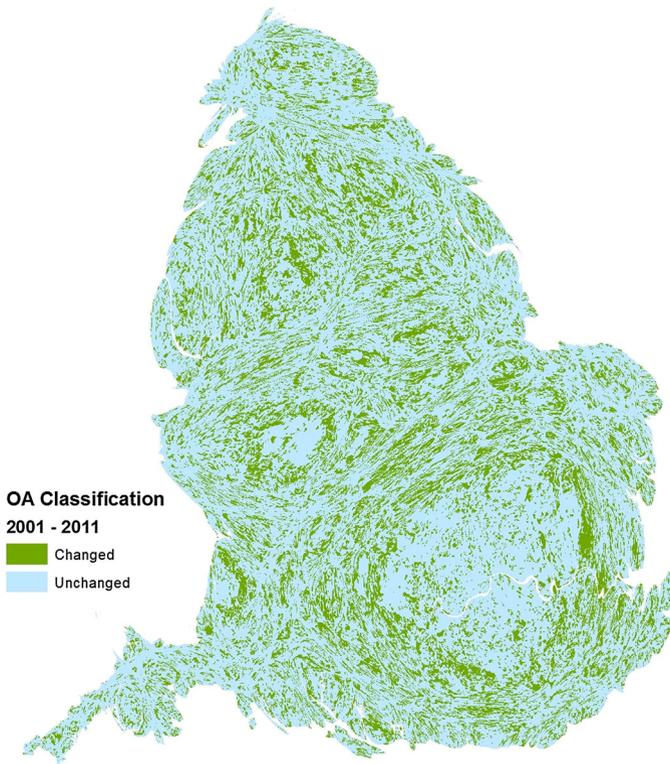


Fig. 2 Output areas that changed their assignment between 2001 and 2011. The changes are shown using a cartogram to illustrate the relative importance of change in urban and suburban areas

clearance during the last intercensal period (Sykes et al. 2013), and there is evidence of this within the core areas radiating from the Central Ward.⁴

3 Mapping the temporal output area classification changes between 2001 and 2011

Nationally, between 2001 and 2011, 46,078 of the 171,372 2011 output areas were reassigned between clusters, and this geography of change is highlighted in Fig. 2, using a cartogram in order to render visible the changes that occurred in urban areas. Of the 46,078 output areas (39 % of the 2011 total) that were reassigned between clusters, 39,444 (85.6 %) lay within urban areas as defined by ONS.⁵ Only a small fraction of these are areas where change was indicated by splitting or merging of

⁴ To further explore national patterns of temporal change, these are visible through our interactive website: (<http://www.maps.cdrc.ac.uk/#/geodemographics/toac/>).

⁵ <http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/2011-rural-urban/index.html>.

Table 3 Cluster reassignments between 2001 and 2011

	2011							
	1—Suburban diversity	2—Ethnicity central	3—Intermediate areas	4—Students and aspiring professionals	5—County living and retirement	6—Blue-collar suburbanites	7—Professional prosperity	8—Hard-up households
1—Suburban diversity	90.4	6.0	1.0	2.3	0.0	0.1	0.1	0.1
2—Ethnicity central	4.1	89.1	0.0	6.5	0.0	0.0	0.0	0.3
3—Intermediate areas	11.0	0.2	76.6	3.5	2.2	1.7	2.3	2.5
4—Students and aspiring professionals	6.7	10.4	3.1	78.2	0.0	0.0	0.1	1.6
5—County living and retirement	0.1	0.0	12.4	0.0	78.4	2.4	6.2	0.5
6—Blue-collar suburbanites	11.0	0.0	15.2	0.4	2.4	61.7	3.0	6.3
7—Professional prosperity	2.2	0.0	15.6	0.3	7.3	2.7	71.7	0.2
8—Hard-up households	15.7	3.9	6.0	2.6	0.2	5.4	0.1	66.1

Row percentages sum to 100, but column totals are not so constrained, as the total number of areas assigned to a category may change over the intercensal period

OAs. Of particular note in Fig. 2 are a series of circular patterns that represent the suburbs of a number of urban areas, most notably in the South East of England in London, but also other large metropolitan areas such as Birmingham in the West Midlands and Manchester in the North West.

The national aggregate patterns of change can be examined further by cross-tabulation of the assignment of areas in 2001 with those in 2011. The analysis shown in Table 3 compares assignments in 2001 (rows) with those in 2011 (columns). The cell values are percentages, summing to 100 over each row. Thus, the principal diagonal scores identify the percentage of areas in each cluster that remain the same in both 2001 and 2011. Nationally, the largest single geodemographic transition was between areas assigned to “8—hard-up households” in 2001, of which 15.7 % transitioned to “1—suburban diversity” in 2011. Although it is beyond the scope of this paper to comment about the detailed pattern of transitions, this particular change evidently reflects a national trend for suburbs to become less dominated by those describing themselves as White, instead becoming more ethnically diverse (see Catney 2015). A further interesting national pattern is the increased prevalence of cluster “4—students and aspiring professionals” in 2011, reflecting the intensification of student residential developments in many central urban areas.

Further insight is gained when flows are disaggregated by region and are perhaps best exemplified by changes in London (25,053 output areas) and the South East (27,638 output areas) that are presented in Tables 4 and 5 and the North West (23,343 output areas) and North East (8802 output areas) in Tables 6 and 7; Fig. 3.

These data illustrate that some of the aggregate patterns of change observed in the national cross-tabulation shown in Table 3 mask regional variations that are quite striking. For example, the cluster “2—ethnicity central” is much more stable between 2001 and 2011 in the South East and London, than in the North West and North East. Such changes bear testimony to the changing ethnic composition of the UK as a whole and the trend for the regions to become more like London in terms of ethnic composition. Conversely, areas classified as “8—hard-up households” are more stable in the North West and North East than the South East or London where economic conditions improved faster during the first part of the intercensal period. The changing assignments from this cluster within the South East, and acutely so in London, are predominantly towards “1—suburban diversity” which are more ethnically diverse and reflect those national changes highlighted earlier. In London just 16.7 % of output areas remaining within “8—hard-up households” by 2011, only 27.1 % of “6—blue-collar suburbanites” output areas remain in this cluster by 2011. The changing composition of these areas results in their assignment most prevalently swapping to “1—suburban diversity”. Such processes of suburban change appear to be most visible in London, which might be a result of historically larger ethnic minority populations resident in central areas, or pressures on housing affordability and space.

Table 4 Cluster assignment changes between 2001 and 2011 (London)

	2011							
	1—Suburban diversity	2—Ethnicity central	3—Intermediate areas	4—Students and aspiring professionals	5—County living and retirement	6—Blue-collar suburbanites	7—Professional prosperity	8—Hard-up households
1—Suburban diversity	87.3	9.7	0.8	2.1	0.0	0.0	0.1	0.0
2—Ethnicity central	3.0	90.6	0.0	6.3	0.0	0.0	0.0	0.0
3—Intermediate areas	40.2	1.0	50.9	5.4	0.1	0.7	1.6	0.2
4—Students and aspiring professionals	5.9	15.9	1.0	77.2	0.0	0.0	0.0	0.0
5—County living and retirement	0.0	0.0	33.3	0.0	57.1	0.0	9.5	0.0
6—Blue-collar suburbanites	55.3	0.5	13.3	0.1	0.0	27.1	2.8	0.9
7—Professional prosperity	14.2	0.1	18.2	0.1	0.1	1.9	65.4	0.0
8—Hard-up households	59.4	18.6	1.8	2.7	0.0	0.8	0.0	16.7

Table 5 Cluster assignment changes between 2001 and 2011 (South East)

	2011							
	1— Suburban diversity	2— Ethnicity central	3— Intermediate areas	4—Students and aspiring professionals	5—County living and retirement	6—Blue-collar suburbanites	7— Professional prosperity	8— — Hard-up households
2001								
1—Suburban diversity	93.1	2.5	1.1	3.0	0.0	0.3	0.0	0.0
2—Ethnicity central	7.7	83.1	0.0	5.6	0.0	0.0	0.0	3.5
3—Intermediate areas	9.4	0.1	80.0	3.9	1.5	1.2	1.9	2.1
4—Students and aspiring professionals	6.4	4.4	5.3	80.8	0.1	0.0	0.2	2.8
5—County living and retirement	0.3	0.0	22.4	0.0	66.0	1.9	8.8	0.4
6—Blue-collar suburbanites	14.3	0.1	22.0	0.6	2.0	52.0	1.7	7.3
7—Professional prosperity	2.0	0.0	22.5	0.4	6.5	1.9	66.4	0.3
8—Hard-up households	17.1	4.0	8.7	4.2	0.1	2.3	0.1	63.5

Table 6 Cluster assignment changes between 2001 and 2011 (North West)

	2011							
	1—Suburban diversity	2—Ethnicity central	3—Intermediate areas	4—Students and aspiring professionals	5—County living and retirement	6—Blue-collar suburbanites	7—Professional prosperity	8—Hard-up households
1—Suburban diversity	91.1	1.7	2.7	3.5	0.0	0.3	0.2	0.4
2—Ethnicity central	18.7	72.0	0.0	7.0	0.0	0.0	0.0	2.3
3—Intermediate areas	3.5	0.0	82.8	1.9	2.0	3.0	3.8	2.9
4—Students and aspiring professionals	7.0	6.6	2.8	79.3	0.1	0.0	0.0	4.1
5—County living and retirement	0.0	0.0	11.6	0.0	78.2	2.0	8.1	0.2
6—Blue-collar suburbanites	7.0	0.0	14.0	0.4	1.2	66.2	4.1	7.0
7—Professional prosperity	0.7	0.0	10.8	0.2	4.2	3.3	80.8	0.1
8—Hard-up households	10.9	3.1	4.9	1.9	0.0	6.9	0.1	72.3

Table 7 Cluster assignment changes between 2001 and 2011 (North East)

	2011							
	1—Suburban diversity	2—Ethnicity central	3—Intermediate areas	4—Students and aspiring Professionals	5—County living and Retirement	6—Blue-collar suburbanites	7—Professional prosperity	8—Hard-up households
1—Suburban diversity	89.0	3.7	0.0	2.4	0.0	1.2	0.0	3.7
2—Ethnicity central	17.6	70.6	0.0	5.9	0.0	0.0	0.0	5.9
3—Intermediate areas	1.4	0.0	84.4	1.6	2.5	2.5	3.3	4.4
4—Students and aspiring professionals	4.9	5.3	4.5	80.0	0.0	0.0	1.6	3.7
5—County living and retirement	0.0	0.0	8.7	0.0	81.9	3.3	5.7	0.4
6—Blue-collar suburbanites	2.0	0.0	8.8	0.0	1.9	79.8	2.7	4.8
7—Professional prosperity	0.3	0.0	11.0	0.0	3.4	3.3	82.0	0.1
8—Hard-up households	4.1	2.2	6.5	2.8	0.3	11.1	0.1	73.0

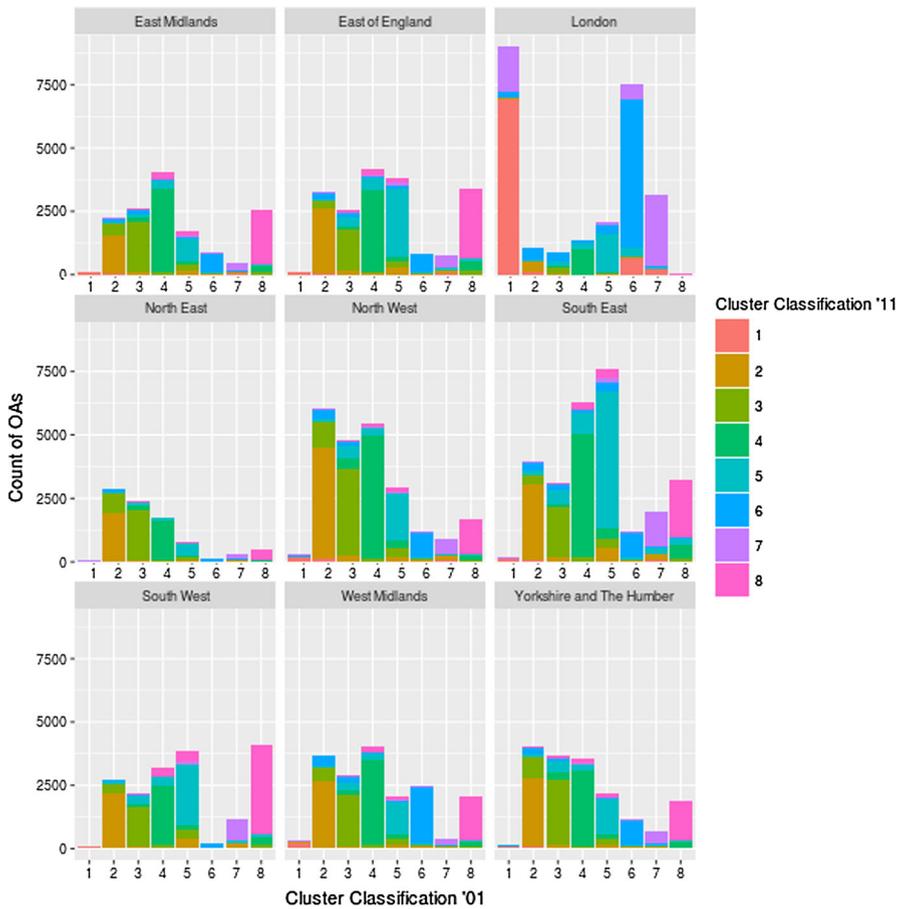


Fig. 3 Percentage of areas changing class within each region (1 country living, 2 ethnicity central, 3 students, 4 professional prosperity, 5 blue collar, 6 suburban diversity, 7 intermediate areas, and 8 hard-up professionals)

4 Ancillary secondary data as indicators of likely geodemographic change

In this section, we discuss a range of indicators of probable change in local geodemographic structure: indicators arising out of changing boundaries used by ONS; indicators based upon single variables that are made available at small area level throughout intercensal periods; indicators that are composites of more than one variable available in intercensal periods; and indicators of centrality derived from the clustering procedure. The first three of these are indicative of changes in the nature and composition of neighbourhood areas, while the fourth identifies the neighbourhoods that are least well accommodated by the classification and which might thus be more likely to transition between categories of it.

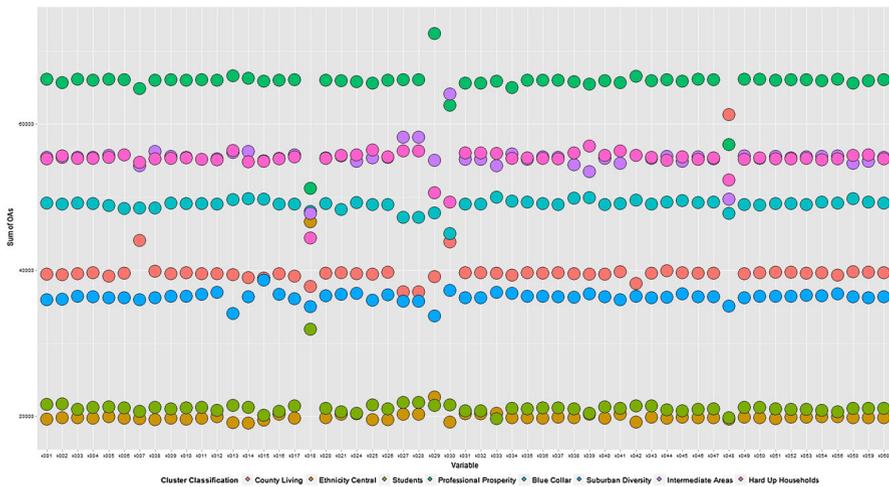


Fig. 4 Frequency distribution of output areas per cluster for each iteration of cluster analysis when one of the variables was removed (for key to variables shown on x axis, see “Appendix”)

As discussed above, a challenge for census-based geodemographics is that their input data are only renewed periodically, which results in decay in accuracy over time. With respect to the 2001 and 2011 OACs, as pointed out above, changes in the areal extents of census output areas that are designed to accommodate data disclosure and granularity requirements provide a first obvious indicator of change and affected 2.6 % of all OAs over the 2001–2011 period. A more developed methodology for identifying change might either supplement census attributes with relevant and more frequently updated open data, such as those derived from ancillary administrative sources. A second approach is to utilise such measures external to the classification, in order to create a composite indicator of temporal uncertainty (Gale and Longley 2013). However, the input variables to a geodemographic classification have different impacts on the assignment of areas into geodemographic clusters: as a consequence, use of intercensal data sources may be more or less useful in evaluating change in the different domains underpinning a classification input. To provide a guide to such effectiveness in the context of the classification created in the previous section, we implement a sensitivity analysis to inform which of the input variables might benefit from ancillary sources, given their impact on the aggregate classification structure.

In successive analyses, each variable was removed in turn and the clustering process repeated. This procedure thus led to 55 iterations of the cluster analysis, and the effect of removing each variable in succession was assessed by examining the frequency distribution of the output areas assignments to each of the eight clusters (Fig. 4: the key to variables listed on the x axis is provided in “Appendix”). This highlighted the individual variables that have the greatest influence upon cluster formation across the two time periods, as shown in Table 8. These included population density, Black/African/Caribbean ethnicity, terraced houses, flatted

Table 8 Most significant variables affecting the results of cluster analysis

Cluster	Description	Household variables associated with increased assignment to category	Household variables associated with decreased assignment to category
1	Suburban diversity	Population density, flatted housing, agriculture (etc.) occupation	Semi-detached housing, detached housing, Black/African/Caribbean, two or more cars
2	Ethnicity central	Black/African/Caribbean, couple with no children, terraced housing, owner tenure	Indian, Pakistani, non-dependent children, flatted housing, public transport to work
3	Intermediate areas	Children aged 0–4, or aged 5–14, Black/African/Caribbean	Pakistani, couple with no children, private, qualification level 4, agriculture/fishing
4	Students and aspiring professionals	Household head aged 25–44, mixed ethnicity, Indian, flatted housing, public transport to work	Population density, Black/African/Caribbean, flatted housing, occupancy rating, agriculture (etc.) occupation
5	County living and retirement	Indian, Pakistani, owner tenure, qualification level 3, qualification level 4, public administration occupation	Black/African/Caribbean, detached housing, semi-detached housing, flatted housing, agriculture (etc.) occupation
6	Blue-collar suburbanites	White, Pakistani, couple with no children, flatted housing, owner tenure, occupancy rating	Mixed ethnicity, Black/African/Caribbean, detached housing, semi-detached housing, terraced housing, agriculture (etc.) occupation
7	Professional prosperity	Lives in communal establishment, mixed ethnicity, Indian, detached housing, semi-detached housing, flatted housing	Population density, Black/African/Caribbean, owner tenure, qualification levels 3 and 4, agriculture (etc.) occupation
8	Hard-up households	Mixed ethnicity, non-dependent children, detached housing, semi-detached housing, qualification level 4, two or more cars	Population density, Indian, Black/African/Caribbean, terraced housing, flatted housing, agriculture (etc.) occupation

housing and the percentage of people employed in agriculture, forestry or fishing (i.e. variables k018, k029, k030 and k048 as identified in “Appendix”).

The next stage was to explore sources of ancillary data that were publicly available at the same spatial scale to our classification and with the promise for creating either temporally updatable variables or measure of change that is central to the classification as suggested by Table 8. Candidate sources that were identified included the Office for National Statistics annual estimates of population structure,⁶ and schools data⁷ derived from the National Pupil Database, which contains details of the ethnicity of state school pupils and, by extension, estimates of changing ethnicity in the general population, particularly if benchmark associations can be established for census years. An alternative method might include the application of names-based classifications of probable ethnicity to enhanced public versions of the

⁶ Available for a larger aggregate here: <http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcn%3A77-320.861> and for more disaggregate geography by request.

⁷ Available after project approval: <https://www.gov.uk/national-pupil-database-apply-for-a-data-extract>.

Register of Electors in order to derive annually updated spatial estimates of ethnicity (Longley et al. 2011). Surrogates for socio-economic status were identified through data that were collected during the process of registering for unemployment benefits (e.g. Jobseeker's allowance).⁸ Changes in the housing stock within England are captured in the Council Tax Valuation List,⁹ which is constantly updated as a basis to local property taxation. Banded property values provide a valuable output area scale indicator of potential changes in the volume and nature of the housing stock. Land Registry¹⁰ house sale transactions are also available at an address level and provide an indicator of property churn and residential mobility within an area, albeit with the caveat that changes within the rented sectors are not monitored. Other domains of the classification are less well represented using ancillary data. For example, there is a notable absence of data about travel behaviour or occupation. However, in the future, there might be potential to use surveys or even social media data to create small area estimates, although the former approach raises issues of generalisation between spatial scales (Spielman and Singleton 2015) and the latter on representativeness (Arribas-Bel 2014).

Taken together, the ancillary data that we have identified were deemed likely to capture key change dynamics. In the remainder of this section, we explore their usefulness in developing an indicator of the likely decay in the relevance of decennial census-based classifications. We focus upon three candidate indicators: mid-year population estimates; Council Tax valuation bands; and Land Registry house sale transactions. These are relevant to the results of the sensitivity analysis presented in Table 8, and all are available at the census output area scale, annual mid-year population estimates are available for the period 2002–2010, while Council Tax bandings and house sale transactions are available for the period 2001–2011. In order to obtain an indication of change, three metrics were developed: from the population estimates, the total population was used; for the Council Tax bands, the sum of the first four bands were combined (bands A, B, C, D); and for the house sales data, the total number of transactions was used. These individual metrics were developed after testing a number of different attribute combinations, and selecting the combination that demonstrated greatest utility in detecting change. The measures were calculated at output area level as the maximum absolute deviation for the period for which the data were available:

$$\text{Maximum Absolute Deviance} = \sum |x_i - \max(x)|/n \quad (1)$$

where x_i is the value at year i , $\max(x)$ the maximum value for the period for which the data were available, and n is the cumulative number of years in the calculation. Higher values indicate greater change during the period of study, in turn suggesting greater unreliability in the origin (2001) classification for a given output area.

As a first step, the three indicators were compared for output areas that did or did not change cluster between 2001 and 2011. The boxplots in Fig. 5 show that the

⁸ Available through: <https://www.nomisweb.co.uk/>.

⁹ Available through: <http://www.neighbourhood.statistics.gov.uk/>.

¹⁰ <http://landregistry.data.gov.uk/>.

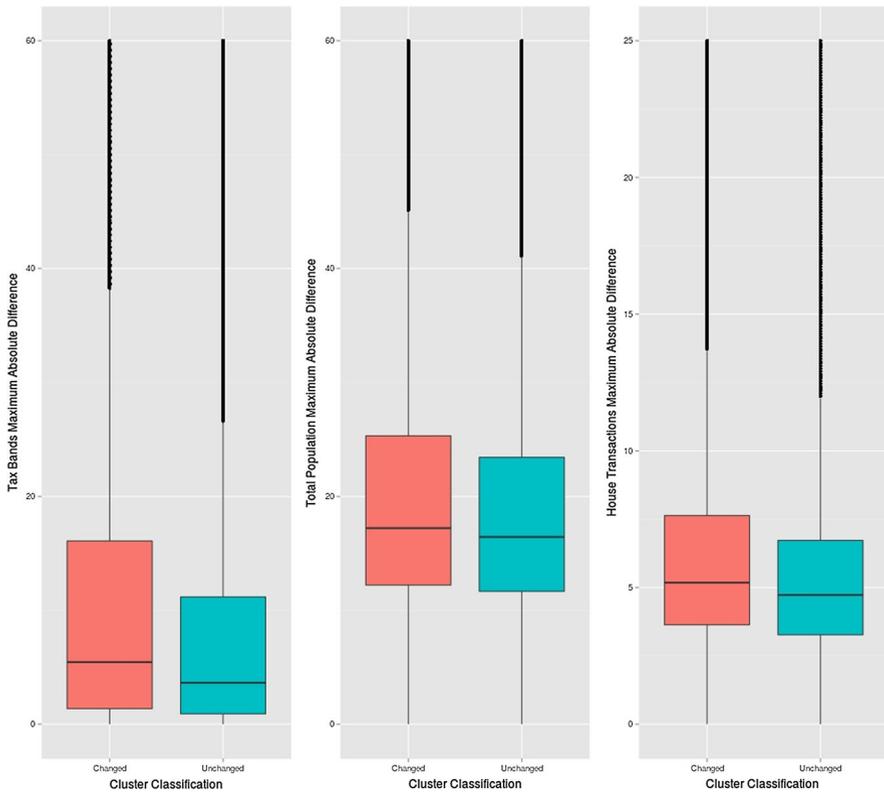


Fig. 5 Data distribution of the uncertainty indicators conditional on cluster reassignment

output areas that were reassigned have higher median and third quartile values for each of the four indicators. The difference between these two groups was shown to be statistically significant using the Mann–Whitney–Wilcoxon nonparametric test ($p < 0.001$).

To better understand this outcome, an indicator of cluster stability was developed to compare the difference in the Euclidean distance of an output area to its assigned cluster centroid in 2001 and 2011. The advantage of applying a single cluster analysis using data from the 2001 and 2011 census periods is evident for the development of this indicator, since the centroids of the clusters remain the same for 2001 and 2011, and hence, comparison is more meaningful than would be the case if the cluster analysis had been applied to each year independently. The assumption of using this indicator is that the greater the distance between an output area and cluster centroid, the more probable it would be for a cluster reassignment to occur between years. However, as can be seen in Fig. 6a, there are only very small differences between the uncertainty scores for areas that were reclassified and areas that were not.

A series of additional internal indicators are also presented in Fig. 6b–e. One hypothesis is that the greater the squared Euclidean distance between the cluster

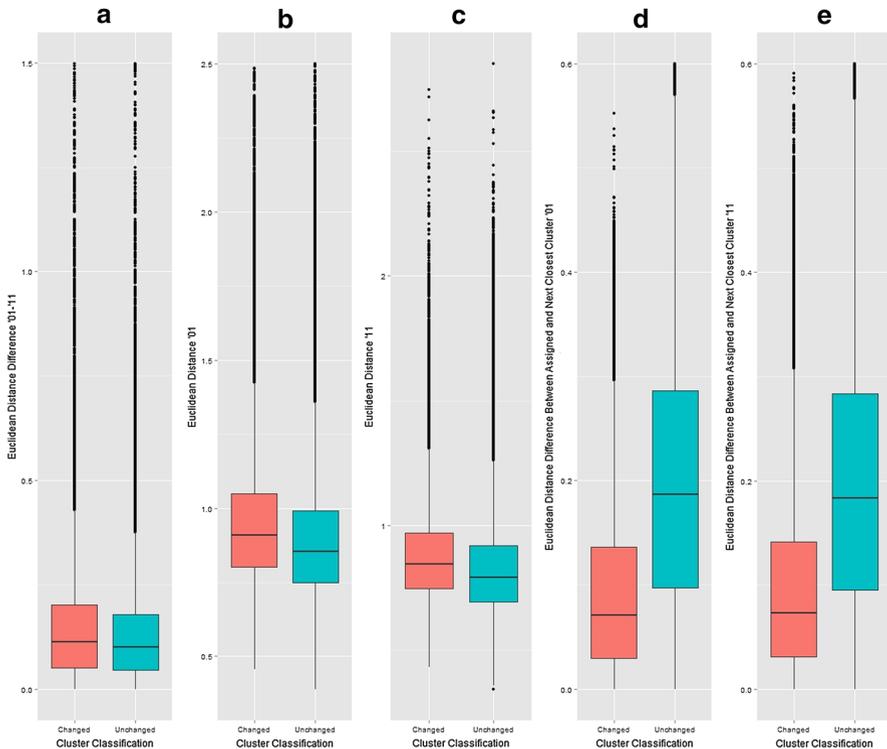


Fig. 6 Data distribution of the internal uncertainty indicators conditional on cluster reassignment

attribute values representing an output area and its assigned cluster centroid, the closer the zone would be to the margin of the cluster, and therefore, the more likely it is that the zone would have been reassigned between 2001 and 2011 (boxplots b and c—larger scores equate to greater distance). A further indicator measures the absolute difference in distance between each output area attribute values, and the second centroid in closest proximity (boxplots d and e in Fig. 6). Taken together, these measures demonstrate that cluster reassignment is most likely to occur because output areas are closer to the margin of their clusters, rather than because they have moved considerable distances because they have experienced profound changes in cluster attributes.

5 Conclusions

The underlying motivation for this study has been to evaluate the prospects for “open” geodemographics, over a period during which various countries have made more and more open data available than ever before. Yet the most widely used classifications remain commercial and closed, because of the widely held perception that classifications based upon open census data become obsolete over the 10-year

periods following publication of census results. This perception is substantiated by the analysis presented here, although it is also clear that (a) some geodemographic segments undergo more rapid change than others and (b) some areas of the country are more affected than others by the different dynamics of change. The methodology of pooling census data between censuses provides a valuable benchmark for direct comparison, and stepwise removal of successive variables provides a means of assessing the sensitivity of cluster outcomes to key variables. The results suggest, however, that these effects can be assessed using ancillary open data. The variables that we have identified here suggest directions in which small area open geodemographics might move in the coming years. Still more might be achieved using ancillary open data sources at coarser granularities (e.g. lower super output areas in the UK), or whether the promise of small area estimation techniques can successfully be applied to geodemographic classifications (Birkin and Clarke 2012).

Geodemographics works fundamentally by assigning clusters of high- and low-order data (nominal, ordinal, interval and ratio scales) to discrete categories. There are many subjectivities inherent in this process—not least, faith that multivariate space is indeed populated by clusters of similar areas and that it is not merely being almost arbitrarily dissected in analysis, and that there is an “optimum” number of clusters that is most appropriate to a full range of end uses. The introduction of the temporal dimension in this paper further increases the ambiguities inherent in this process. The approach that we have adopted is one extreme, in that we assume a temporal invariance to the ways in which society may be profiled. At the opposite extreme lie the commercial classifications (such as Acorn: CACI, London) that create entirely new classes using different variables in successive major releases. An intermediate position is that of the output area classifications (e.g. Vickers and Rees 2007) which use broadly the same *mélange* of census variables and data transformations to cluster data from successive censuses, but use entirely new cluster classes to describe the results. Societies are dynamic, variables take on different connotations in different locations and time periods (car ownership is a prominent example), and the structure of any classification is very much an outcome of choice, convention and indeed chance. We do not see the pooling of data across time periods as an unusual or radical departure from these practices.

Eight clusters were identified in our pooled analysis, namely “suburban diversity”, “ethnicity central”, “intermediate areas”, “students and aspiring professionals”, “county living and retirement”, “blue-collar suburbanites”, “professional prosperity” and “hard-up households”. An important assumption underpinning our analysis is that the socio-economic structure of the study area remains fundamentally unchanged over the 2001–2011 period. This is a crucial conjecture, and it is for future research in the fast developing field of geotemporal demographics to confirm whether or not this is reasonable: our own interim view, however, is that the inherent inability to fully specify temporal change should be treated no differently to the fundamental inability to fully specify the place effects that underpin the patterns of social similarity that are observed in conventional geodemographics. In substantive terms, the results of the cluster analysis and sensitivity analysis illustrate the changing importance over time of ethnicity as a

driver of cluster structure, with two of the eight clusters impacted by structural changes in the ethnic composition of the nation.

In addition, as one might expect, different processes can be identified at different spatial scales. Thus, at the national scale, cluster analysis shows evidence of increasing ethnic minority flows to the suburbs as well as patterns of gentrification in central areas because of new residential developments and redevelopments that are attractive to younger professionals and students. At the regional scale, however, it is obvious that the greatest flows of ethnic minorities to suburbs actually occurs mostly in London, while other regions have more stable classifications of suburban areas.

Census geodemographics are by definition constrained to looking at patterns of population and built structure every 10 years. This is a critical constraint given how rapidly populations can change; however, at present, available UK open data sources do not offer a wide enough set of attributes that might enable classifications to be compiled entirely from these resources. The surrogate indicators of change presented in this paper represents a step towards how such data might be used prospectively, but also tentatively points towards potential future research directions. Investments within UK spatial data infrastructure are leading to the expansion of open data resources, and furthermore, through a variety of heuristic and matching processes, are enabling the creation of linked open data resources (see <http://www.adm.ac.uk/>). Such developments hold great potential for future open geodemographic classification.

Acknowledgements This research was funded under grants ES/K004719/1 ('Using secondary data to measure, monitor and visualise spatio-temporal uncertainties in geodemographics'), ES/L011840/1 ('Retail Business Datasafe') and ES/L013800/1 ('The analysis of names from the 2011 Census of Population').

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix

See Table 9.

Table 9 Variables used in the cluster analysis

Variable	Description
k001	% Persons aged 0–4
k002	% Persons aged 5–14
k003	% Persons aged 25–44
k004	% Persons aged 45–64
k005	% Persons aged 65–89
k006	% Persons aged 90 and over
k007	Population density
k008	% Persons living in a communal establishment
k009	% Persons aged over 16 who are single
k010	% Persons aged over 16 who are married or in a registered same-sex civil partnership
k011	% Persons aged over 16 who are divorced or separated
k012	% Persons who are white
k013	% Persons who have mixed ethnicity or are from multiple ethnic groups
k014	% Persons who are Asian/Asian British: Indian
k015	% Persons who are Asian/Asian British: Pakistani
k016	% Persons who are Asian/Asian British: Bangladeshi
k017	% Persons who are Asian/Asian British: Chinese and Other
k018	% Persons who are Black/African/Caribbean/Black British
k020	% Persons whose country of birth is the UK or Ireland
k021	% Persons whose country of birth is in the EU
k024	% Households with no children
k025	% Households with non-dependent children
k026	% Households with full-time students
k027	% Households who live in a detached house or bungalow
k028	% Households who live in a semi-detached house or bungalow
k029	% Households who live in a terrace or end-terrace house
k030	% Households who live in a flat
k031	% Households who own or have shared ownership of property
k032	% Households who are social renting
k033	% Households who are private renting
k034	% Occupancy rating
k035	Individuals day-to-day activities limited a lot or a little (standardised illness ratio)
k036	% Persons providing unpaid care
k037	% Persons aged over 16 whose highest level of qualification is level 1, level 2
k038	% Persons aged over 16 whose highest level of qualification is level 3 qualifications
k039	% Persons aged over 16 whose highest level of qualification is level 4 qualifications
k040	% Persons aged over 16 who are schoolchildren or full-time students
k041	% Households with 2 or more cars or vans
k042	% Persons aged between 16 and 74 who use public transport to get to work
k043	% Persons aged between 16 and 74 who use private transport to get to work
k044	% Persons aged between 16 and 74 who walk, cycle or use an alternative method to get to work

Table 9 continued

Variable	Description
k045	% Persons aged between 16 and 74 who are unemployed
k046	% Employed persons aged between 16 and 74 who work part-time
k047	% Employed persons aged between 16 and 74 who work full-time
k048	% Employed persons aged between 16 and 74 who work in the agriculture, forestry or fishing industries
k049	% Employed persons aged between 16 and 74 who work in the mining, quarrying or construction industries
k050	% Employed persons aged between 16 and 74 who work in the manufacturing industry
k051	% Employed persons aged between 16 and 74 who work in the energy, water or air conditioning industry
k052	% Employed persons aged between 16 and 74 who work in the wholesale and retail trade; repair of motor vehicles and motor cycles
k053	% Employed persons aged between 16 and 74 who work in the transport or storage industries
k054	% Employed persons aged between 16 and 74 who work in the accommodation or food service activities industries
k056	% Employed persons aged between 16 and 74 who work in the financial, insurance or real estate industries
k058	% Employed persons aged between 16 and 74 who work in the public administration or defence; compulsory social security industries
k059	% Employed persons aged between 16 and 74 who work in the education sector
k060	% Employed persons aged between 16 and 74 who work in the human health and social work activities industries

References

- Adnan M, Longley P, Singleton A, Brunson C (2010) Towards real-time geodemographics: clustering algorithm performance for large multidimensional spatial databases. *Trans GIS* 14(3):283–297
- Arribas-Bel D (2014) Accidental, open and everywhere: emerging data sources for the understanding of cities. *Appl Geogr* 49:45–53
- Ashby DI, Longley PA (2005) Geocomputation, geodemographics and resource allocation for local policing. *Trans GIS* 9:53–72
- Batey P, Brown P (2007) The spatial targeting of urban policy initiatives: a geodemographic assessment tool. *Environ Plan A* 39(11):2774–2793
- Birkin M, Clarke G (2012) The enhancement of spatial microsimulation models using geodemographics. *Ann Reg Sci* 49(2):515–532
- Birkin M, Clarke G, Clarke M (2002) Retail geography and intelligent network planning. Wiley, Chichester
- Catney G (2015) The changing geographies of ethnic diversity and mixing in England and Wales, 1991–2011 population, space and place. doi:[10.1002/psp.1954](https://doi.org/10.1002/psp.1954)
- Everitt BE, Landau S, Lees M (2011) Cluster analysis. Wiley, Chichester
- Gale CG, Longley PA (2013) Temporal uncertainty in a small area open geodemographic classification. *Trans GIS* 17(4):563–588
- Goss J (1995) Marketing the new marketing: the strategic discourse of geodemographic information systems. In: Pickles J (ed) Ground truth. Guildford Press, New York, pp 130–170
- Harris R, Sleight P, Webber R (2005) Geodemographics, GIS and neighbourhood targeting. Wiley, Chichester

- Johnson NL (1949) Systems of frequency curves generated by methods of translation. *Biometrika* 36(1–2):149–176
- Longley PA (2005) Geographical information systems: a renaissance of geodemographics for public service delivery. *Prog Hum Geogr* 29(1):57–63
- Longley PA, Cheshire JA, Mateos P (2011) Creating a regional geography of Britain through the spatial analysis of surnames. *Geoforum* 42(4):506–516
- Petersen J, Gibin M, Longley PA, Mateos P, Atkinson P, Ashby DI (2011) Geodemographics as a tool for targeting neighbourhoods in public health campaigns. *J Geogr Syst* 13(2):173–192
- R Core Team (2013) R: a language and environment for statistical computing R Foundation for statistical computing. Austria, Vienna
- Rees P (1972) Problems of classifying subareas within cities. In: Berry BJJ, Smith KB (eds) *City classification handbook: methods and applications*. Wiley-Interscience, New York, pp 265–330
- Reibel M (2011) Classification approaches in neighborhood research: introduction and review. *Urban Geogr* 32(3):305–316
- Singleton AD (2010) The geodemographics of educational progression and their implications for widening participation in higher education. *Environ Plan A* 42(11):1560–2580
- Singleton AD, Longley PA (2009) Geodemographics, visualisation, and social networks in applied geography. *Appl Geogr* 29(3):289–298
- Singleton AD, Spielman SE (2014) The past, present and future of geodemographic research in the United States and United Kingdom. *Prof Geogr* 66(4):558–567
- Singleton AD, Wilson AG, O'Brien O (2012) Geodemographics and spatial interaction: an integrated model for higher education. *J Geogr Syst* 14(2):223–241
- Spielman SE, Singleton AD (2015) Studying neighborhoods using uncertain data from the American community survey: a contextual approach. *Ann Assoc Am Geogr*. doi:[10.1080/00045608.2015.1052335](https://doi.org/10.1080/00045608.2015.1052335)
- Sykes O, Brown J, Cocks M, Shaw D, Couch C (2013) A city profile of liverpool. *Cities* 35:299–318
- Thompson C, Clarke G, Clarke M, Stillwell J (2012) Modelling the future opportunities for deep discount food retailing in the UK. *Int Rev Retail Distrib Consum Res* 22(2):143–170
- Timms D (1971) *The urban mosaic: towards a theory of residential differentiation*. Cambridge University Press, Cambridge
- Vickers D, Rees P (2007) Creating the UK National Statistics 2001 output area classification. *J R Stat Soc Ser A Stat Soc* 170(2):379–403
- Voas D, Williamson P (2001) The diversity of diversity: a critique of geodemographic classification. *Area* 33(1):63–76