This is the accepted version of the following paper:

Christopher D. Lloyd (2015) Assessing the spatial structure of population variables in England and Wales. *Transactions of the Institute of British Geographers*, 40 (1), 28–43.

doi: 10.1111/tran.12061

Assessing the Spatial Structure of Population Variables in England and Wales

Areas within England and Wales have population profiles which make them distinct from other locales; some areas have lower unemployment rates than others, while, in some places, there is a greater mix of ethnic groups than elsewhere. Thus, the degree of difference between areas differs geographically and between population sub-groups. Being able to measure change in these differences is crucial in assessing whether the population has become more or less similar over time. The spatial distribution of the population by, for example, ethnicity or employment status can be characterised and the resulting measures show how the population is geographically organised, and how this changes through time. For example, spatial concentrations of the population by age may be less obvious locally (e.g., within a town or city) or regionally (e.g., the north west of England) than by housing tenure. This paper makes two key contributions: (i) it introduces methods for the analysis of spatial distributions of population sub-groups and (ii) enhances our understanding of the characteristics of population sub-groups in England and Wales and how they have changed over time. Based on Census data for Output Areas (OAs), the analysis uses the index of dissimilarity (*D*), the Moran's *I* autocorrelation coefficient, and the variogram to measure (spatial) variation in variables representing population sub-groups by age, ethnic group, housing tenure, car or van ownership, qualifications, employment, limiting long term illness (LLTI) and National Statistics Socio-economic Classification (NS-SeC). The analysis shows that, between 2001 and 2011, unevenness in most population sub-groups reduced and the populations in individual Census zones across England and Wales became more similar. Neighbouring Census zones also became more similar (more 'clustered'). The findings suggest that there were decreased differences both within and between regions for many population variables between 2001 and 2011.

key words Spatial dependence, Scale, Census data, Segregation

---

Introduction

Studies of geographical patterns and temporal change in the social, economic and demographic profiles of the United Kingdom and its constituent parts are numerous, examples include the Census atlases produced by Champion et al. (1996) and Dorling and Thomas (2004). The maps on which these accounts are based contain information on how population characteristics such as, for example, employment, housing tenure and ethnicity are distributed spatially. Using maps, it is possible to gain a visual impression of how, for instance, differences in economic status have increased or decreased in one region as opposed to another. Summary statistics may also be compared for different regions (e.g., north west and north east England) at different time points (e.g., 2001 and 2011). But, there is much more that can be extracted from geographically-referenced population data. To give an example, if the standard deviation of the employment rates for zones (such as Census Output Areas (OAs)) has increased between 2001 and 2011 then the differences between zones has increased and inequalities are greater. However, changes may be confined to particular areas or spatial scales – the largest changes may be in more densely occupied urban areas and over only small regions (perhaps within urban areas only, with minimal change outside of urban cores). Thus, taking account of spatial locations, rather than using conventional *aspatial* summary statistics alone, allows assessment of changes in variation at a neighbourhood or

regional scale. This paper seeks to enhance our understanding of the spatial distribution of population sub-groups in England and Wales how they have changed between 2001 and 2011.

The present paper has strong links to debates about (residential) segregation. Segregation has been characterised by a set of indices including the index of dissimilarity, *D*, and the indices of isolation and exposure (see Massey and Denton 1988). The spatial scale of residential segregation is considered by Wong (2004) and Reardon et al (2008), with an analysis of change in the scale of segregation in US metropolitan areas between 1990 and 2000 by Reardon et al (2009). Previous research has used *D* to assess social polarisation in the population of Britain using Census data from each Census year from 1971 to 2001 (Dorling and Rees 2003). In that study, the focus was on a wide range of variables for local authorities (LAs); 2001 boundaries were used and 1991 counts were reaggregated to 2001 LA boundaries (with adjustment for undercount). Also, comparisons were made with a limited set of comparable datasets for 1971 and 1981. Voas and Williamson (2000) compared values of *D* for three different zonal systems (enumeration districts, wards and districts) for multiple demographic and socio-economic variables across England and Wales in 1991. In that case, scale was conceptualised as relating to variation across and within each of the sets of zones.

Questions such as the degree of social or economic difference between the north and south, or between the constituent parts of England and Wales, can be answered by comparing the population characteristics of these regions. But, there may be considerable variation *within* regions which is, on average, greater than variation *between* regions. Comparison of discrete fixed areas (e.g., regions or countries) is limited as there may be a contrast between places at the edges of two regions and this would be ignored if each region is treated as a separate entity. A better approach is arguably to use small areas (e.g., in the UK, Output Areas) and explore scales of variation, without direct reference to regions or countries. In such cases, no fixed hierarchy of areas is used. Lloyd (2010a) used the geographically-weighted Moran's *I* autocorrelation coefficient to assess how spatial dependence of several sub-group variables in Northern Ireland varied over spatial scales defined by different geographical bandwidths. In other words, clustering was characterised at different spatial scales and it was shown that some characteristics (most notably religion) cluster at all scales considered, while others cluster locally, but not over regional scales. Lloyd (2012) used variograms to characterise spatial variation, at multiple spatial scales, in the population of Northern Ireland by religion and the variogram was shown to provide a composite measure of clustering and polarisation of population sub-groups.

In this paper, the focus is on the analysis of demographic and socio-economic variables derived from 2001 and 2011 Census data for England and Wales. The contributions made by the paper are twofold and it – (i) details approaches for the analysis of spatial structure of population sub-groups and (ii) provides a set of results which enhance our understanding of the spatial characteristics of population sub-groups in England and Wales and how they have changed between the two most recent Censuses. The paper is innovative in providing the first systematic analysis of the spatial distribution of population sub-groups in England and Wales in 2001 and 2011, and how these have changed over time. The paper also uses methods which are rarely applied in the analysis of population sub-groups and it demonstrates the potentially considerable value of such approaches.

With respect to the first contribution, such approaches are invaluable in helping to understand how a society is becoming more or less divided, in what ways, and at what spatial scales and, as the paper later argues, should constitute a central part of the analysis of population change. In short, this element of the paper showcases methods for assessing how different areas are over a range of spatial scales (from neighbourhoods to large regions). Understanding spatial variations in

demographic, social, economic or cultural population characteristics, the focus of the second contribution, is a core element of human geography. Characterisation of the spatial structure of population variables is important for several reasons. Firstly, knowledge of how population sub-groups are distributed across spatial scales has direct links to several research areas including analyses of deprivation (see, for example, Norman 2010), residential segregation and health status, where spatial clustering of groups, for example, is a key concern. For any application concerned with concentrations of members of different groups, spatial structure and scales of variation are important. Secondly, any analysis of variables represented using area data (e.g., census zones) is partly a function of the size and shape of the zones, but also the spatial scale of variation in the variable of concern. Much effort has been expended on assessing how populations are, relatively speaking, under- or over-concentrated in particular regions (Voas and Williamson 2000). Characterising these spatial variations, and the scales over which variables are concentrated, is of direct policy relevance and the approaches used in this paper could be used to determine a meaningful or sensible scale over which to implement particular policies.

The second major contribution of the paper comprises results obtained from an analysis of a set of population characteristics for England and Wales, as represented by Census data for 2001 and 2011 using Output Areas, the smallest areas for which data have been released. Using the methods detailed as part of the first contribution, the characteristics of the population by these variables are summarised in several ways which enable assessment of (i) how the selected variables are structured at different spatial scales (e.g., does unemployment cluster over small areas, but not over regions?) and (ii) how the spatial distribution of the population by these variables has changed between 2001 and 2011. The variables relate to age, ethnicity, health status, employment status, qualifications and housing tenure.

While the timing of the Census as a decadal survey does not allow analysis of short-term changes, it is at least possible to assess if, for example, the global recession of 2008-2009 might have contributed to a more or less unequal population in England and Wales. A society may be unequal in many ways, and previous research has focused on ways in which the population of (parts of) the United Kingdom has become or less divided economically (e.g., Hills et al. 2010). In addition to focusing on spatial divisions, the present paper makes links to ongoing debates about the future of the Census in the United Kingdom by suggesting that fine-scale spatially aggregated data, such as OAs, are necessary to capture variation in key population characteristics. It is argued that small area data are essential if we are to properly assess geographic inequalities and that, therefore, a reduction in geographical detail is likely to correspond to a considerably diminished ability to assess how far the population of (parts of) the UK are becoming more or less similar.

Data

The analysis is based on counts released for Output Areas. OAs were generated using an automated zone design methodology whereby an intra-area correlation measure was used to maximise social homogeneity within areas with the constraint that the total population and household numbers were above a predefined threshold and close to the target size (Martin et al 2001). Use of OAs is appropriate for a study concerned with analysis of spatial scales of variation in population sub-groups. In 2001 there were 175,434 OAs in England and Wales (mean population = 297); the equivalent figure for 2011 was 181,408 (mean population = 309). Only some 2.6 per cent of 2001 OAs have been changed as a result of the 2011 Census[1]. Distances between OAs were measured using population weighted centroids. The variables used are derived

[1] http://www.ons.gov.uk/ons/rel/census/2011-census/population-and-household-estimates-for-wards-and-output-areas-in-england-and-wales/index.html

from counts by age, ethnic group, housing tenure, car or van ownership, qualifications, employment, limiting long term illness (LLTI) and National Statistics Socio-economic Classification (NS-SeC). The data are from the Key Statistics tables for 2001 and 2011, and they are specified in Table I.

TABLE I ABOUT HERE

Part of the analysis is based on raw counts, while the remainder of the analysis makes use of data on the proportion (expressed as a percentage) of people in an area who belong to a given group (x_i / t_i), where t_i is the total number of persons in area i . Percentages (and proportions) are referred to as compositional data and they sum to 100 (percentages) or 1 (proportions). Statistical analyses of raw percentages are problematic (Lloyd et al, 2012; see Filzomser et al 2009 for a discussion about univariate data analysis and compositions) and so the analysis presented in this paper makes use of an appropriate transform of the percentages, namely log-ratios.

Analysis

This section details, in turn, a summary of population sub-group counts and percentages in 2001 and 2011, measuring unevenness with the index of dissimilarity (using raw counts of people in sub-groups), transforming percentages with log-ratios, measuring clustering in population sub-groups using the Moran's *I* autocorrelation coefficient (using log-ratio transformed percentages), and analysing spatial variation at different scales with the variogram (again, using log-ratios).

Table II summarises counts and percentages for all of the variables used in this analysis. The largest percentage point changes relate (as they appear in the table) to ethnicity, housing tenure and qualifications. The increase in the size of the non-White population (and particularly the non White British) population of England and Wales between 2001 and 2011 has, since the release of the first 2011 Census data on ethnicity, been debated extensively in media and academic outlets[2]. In the present paper, the broad categories of 'White' and 'non-White' have been selected, and the implications of this choice are considered later in the paper.

TABLE II

The number of private rented households increased by some 1.6 million between 2001 and 2011 (note the definition of 'private rented' in the information associated with Table II). The numbers of owner-occupied households increased (according to the OA-level figures — small cell adjustment procedures for 2001 will impact on the counts) by 115,507, while the number of social rented households decreased by 39,197. The owner-occupier figures include households bought outright or through a mortgage. A tightening of lending requirements by banks and mortgage lenders, following the 2008 financial crisis, is one factor behind the growth of private renting and the proportional decline of owner-occupied households indicated in Table II (see ONS 2013a).

It is noted in Table II that the qualifications totals for 2001 and 2011 cannot be directly compared since the population bases used in the two years are different. Nonetheless, the difference in the percentages of people with (no) qualifications reflects, in part, higher qualifications rates amongst younger people. That is, a person born in 1980, for example, is more likely to attain qualifications than someone born in 1960, and so an increase in qualification rates would be expected between 2001 and 2011 (Thomson et al. 2010 discuss the measurement of adult attainment).

Measuring unevenness with the index of dissimilarity

[2] See http://www.ethnicity.ac.uk/

The index of dissimilarity, D , is one of the most widely used measures of segregation. *D* indicates the total differences between the spread of the two population groups over all of the areal units; it is given by:

$$D = 0.5 \times \sum_{i=1}^{n} \left(\left| \frac{x_1(\mathbf{s}_i)}{X_1} - \frac{x_2(\mathbf{s}_i)}{X_2} \right| \right) \quad (1)$$

where $x_1(\mathbf{s}_i)$ and $x_2(\mathbf{s}_i)$ are counts of population in two groups for areal unit i with centroid $\mathbf{s}_i$ and there are n units. X_1 and X_2 are the total population counts across the whole of the study area. *D* takes a value between 0 and 1 where a large value implies a high degree of unevenness.

The index of dissimilarity was computed for each stated sub-group versus the remainder in the category. For example, the proportion of those aged 0 to 15 was compared to the proportion of all others by age. Table III gives *D* for each sub-group. In 2001, the population was most unevenly distributed by White (i.e., White versus Non-White) and SocRent (social rented households versus all other households), and most evenly distributed by A30to64 (age 30 to 64 versus all others). This is not surprising given the inherently urban nature of population distributions by ethnicity and by social rented households. In 2011, *D* for both White and SocRent decreased, although that for White decreased more. The 'Right to buy' scheme, which offers some tenants of social housing the opportunity to purchase the property at a reduced cost[3], is likely to lead to reduction in unevenness; quite spatially homogenous areas of social housing become more mixed in terms of tenure as properties are purchased and thus household tenure converted. In addition, initiatives which support 'mixed tenure' as a part of new building schemes will reduce spatial variation in terms of differences between neighbourhoods. It is worth nothing at this stage that all of the results are dependent in part on the zonal systems – the small differences (in proportional terms) in OAs in 2001 and 2011 suggests, however, that these results are robust.

TABLE III ABOUT HERE

Transforming percentages: log-ratios

As detailed in the Data section, analysis of raw percentages is problematic. The remainder of the analysis is based on percentages and these are transformed to log-ratios which facilitates analysis using standard statistical methods. Log-ratios are used as input to analysis using (i) standard statistical measures, (ii) Moran's *I* and (iii) the variogram. There were zero values in some categories and, for computing log-ratios, the proportions are calculated from counts x_1, x_2, x_3 … with $x_1 + 1$, $x_2 + 1$, $x_3 + 1$ … (see Lloyd 2010a for a justification of this approach). That is, a value of one is added to all counts and the percentages, y_1, y_2, y_3 ... are calculated from the modified counts. The sensitivity of results to the addition of different values (e.g., 0.1 and 0.5) was assessed and the results were found to be robust.

Aggregations of population data are often arbitrary — for example, ages could be grouped by year (e.g., 0, 1, 2, 3… years) or by any number of years together. In the present analysis, five sets of counts are divided into two groups (ethnicity, cars and vans, qualifications, employment and LLTI), there were two three-part compositions (housing tenure and NS-SeC), and one four-part

[3] https://www.gov.uk/government/policies/helping-people-to-buy-a-home

composition (age). The percentages were computed from these sets of counts – so the compositions are two, three or four part and it is the ratio between the parts which is of interest.

The additive-log-ratio (alr) and the centred-log-ratio (clr) were introduced by Aitchison (2003) for the transformation of compositional data. Egozcue et al. (2003) developed the isometric-log-ratio (ilr) transform (see Egozcue and Pawlowsky-Glahn (2006), for a summary). The outputs from alr and clr transforms are subject to restrictions in their treatment by standard methods, whereas ilr transformed data can be analysed directly using standard univariate or multivariate statistical methods. Egozcue and Pawlowsky-Glahn (2005) developed a form of ilr coordinates called balances. Balances represent the relative variation in two groups of parts and they may have straightforward interpretations. Balances (as ilr coordinates generally) can be analysed using standard multivariate statistical approaches. Balances provide a means of analysing simultaneously variation within groups of parts and between groups of parts (Egozcue and Pawlowsky-Glahn (2005)). The present analysis is based on balances. The general equation for computing balances for groups (sets of parts) R_1 and R_2 is:

$$z_p = \sqrt{\frac{r_1 r_2}{r_1 + r_2}} \ln \left(\frac{\left(\prod_{j \in R_1} y_j \right)^{\frac{1}{r_1}}}{\left(\prod_{k \in R_2} y_k \right)^{\frac{1}{r_2}}} \right) \quad \text{for} \quad p = 1,..., P-1 \tag{2}$$

The products $\prod_j$ and $\prod_k$ refer to parts within groups R_1 and R_2; r_1 and r_2 refer to the number of parts in these two groups (the number of parts with, respectively, positive and negative signs in the partition, as illustrated below) for the p th order (Egozcue and Pawlowsky-Glahn, 2006), and P is the number of parts in the composition. An example partition for six variables $(y_1,..., y_6)$ can be given with:

y_1	y_2	y_3	y_4	y_5	y_6	Balance (z)
1	1	1	1	-1	-1	1
1	1	1	-1	0	0	2
1	1	-1	0	0	0	3
1	-1	0	0	0	0	4
0	0	0	0	1	-1	5

The five balances are computed, following Equation 2, with:

$$z_1 = \sqrt{\frac{8}{6}} \ln \frac{(y_1 y_2 y_3 y_4)^{\frac{1}{4}}}{(y_5 y_6)^{\frac{1}{2}}}, \; z_2 = \sqrt{\frac{3}{4}} \ln \frac{(y_1 y_2 y_3)^{\frac{1}{3}}}{y_4}, \; z_3 = \sqrt{\frac{2}{3}} \ln \frac{(y_1 y_2)^{\frac{1}{2}}}{y_3},$$

$$z_4 = \sqrt{\frac{1}{2}} \ln \frac{y_1}{y_2}, \; z_5 = \sqrt{\frac{1}{2}} \ln \frac{y_5}{y_6}$$

Partitions can be selected using expert knowledge or using a compositional biplot (see Lloyd et al 2012).

The example of the age variable is detailed below. In the two part case, the order of the variables (e.g., y_1 / y_2 to y_2 / y_1) is not important. However, if the order is reversed this changes the sign

of the log-ratio (and the sign of correlation coefficients for paired variables) and clearly it is important for purposes of interpretation to be aware of the variable order. In the three and four part case and for compositions with more than three parts, the order does matter and, in the case of the age composition, for example, age order was followed (i.e., the first group is the youngest and the last group is the oldest). In other cases with more than two parts logical ordering was also followed. The ilr transform was used in this analysis although, in the case of two part compositions, the results of all analyses presented in this paper would be identical if the widely-used alr transform was used (although alternative statistical methods would produce different results in some contexts). As noted previously, the ilr transformed variables can be analysed using any standard multivariate statistical methods (for example, principal components analysis). Log-ratio transforms, including the ilr transform (and back-transform) can be performed in the free software package CoDaPack (Thió-Henestrosa and Martín-Fernández 2005) and the R package 'compositions' (van den Boogaart and Tolosano-Delgado 2008). In this study, the log-ratios are derived using the following elements:

Two part compositions:

Ethnicity:	AllWhite / NonWhite
CarsVans:	NoCarsVans / CarsVans
Qual:	NoQual / Qual
Employ:	Employ / Unemploy
LLTI:	LLTI / NonLLTI

Three or four part compositions:

Tenure (**Denominator**)

No	y_1	y_2	y_3
1	OwnOcc	PrivRent	**SocRent**
2	OwnOcc	**PrivRent**	

NS-SeC

No	y_1	y_2	y_3
1	NSSEC12	NSSEC37	**NSSEC8**
2	NSSEC12	**NSSEC37**	

Age

No	y_1	y_2	y_3	y_4
1	A0to15	A16to29	A30to64	**A65plus**
2	A0to15	A16to29	**A30to64**	
3	A0to15	**A16to29**		

Taking Age as an example, this leads to:

$$\text{Age } 1 = \sqrt{\frac{3}{4}} \ln \frac{(\text{A0to15} \times \text{A16to29} \times \text{A30to64})^{\frac{1}{3}}}{\text{A65plus}}, \quad \text{Age } 2 = \sqrt{\frac{2}{3}} \ln \frac{(\text{A0to15} \times \text{A16to29})^{\frac{1}{2}}}{\text{A30to64}},$$

$$\text{Age } 3 = \sqrt{\frac{1}{2}} \ln \frac{\text{A0to15}}{\text{A16to29}}$$

Figures 1, 2 and 3 show respectively Ethnicity log-ratios, Tenure1 log-ratios and LLTI log-ratios. It is clear from visual examination of the maps that the structure of the three variables is quite different. Figure 1 (where large positive values correspond to OAs with large percentages of people identifying with a White ethnic group) shows that spatial variation in ethnicity is largely an urban phenomenon. This reflects settlement patterns of particular immigrant groups (discussion about second- and later-generation immigrants appears later). In Figure 2, large negative values (corresponding to large percentages of social rented households) are found in urban areas, but there are less visually distinct differences between urban and rural areas. Social housing is likely to be strongly structured in some places since local authority-run housing has tended to be constructed in discrete locales. The spatial patterning in LLTI log-ratios (Figure 3; large negative values correspond to smaller percentages of people with a LLTI), is again quite different. There appears to be a larger concentration of OAs with very large negative log-ratios in the south east than elsewhere, and generally smaller negative log-ratios in urban areas (with positive values in some areas) than in more rural areas. While differences may be expected between, say, housing tenure and LLTI or ethnicity, the nature and scale of these differences can only be explored through the use of some quantitative summary measure.

FIGURE 1 ABOUT HERE
FIGURE 2 ABOUT HERE
FIGURE 3 ABOUT HERE

Measuring clustering using Moran's I

In this section, clustering in log-ratios is measured using the Moran's I autocorrelation coefficient (Moran 1950, Cliff and Ord 1973). Moran's I with weights, w_{ij}, between locations $\mathbf{s}_i$ and $\mathbf{s}_j$ row-standardised (i.e., the weights for each i sum to one) can be given by:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(z(\mathbf{s}_i) - \bar{z})(z(\mathbf{s}_j) - \bar{z})}{\sum_{i=1}^{n}(z(\mathbf{s}_i) - \bar{z})^2} \tag{3}$$

where the values $z(\mathbf{s}_i)$ have the mean $\bar{z}$. Positive values of I indicate positive spatial autocorrelation (spatial dependence), while negative values indicate negative spatial autocorrelation.

In this analysis, the Moran's I coefficient was computed using a kernel of variable size (where its size is defined by its bandwidth). The kernel defines a weighting scheme which indicates weights assigned to observations as a function of distance from a particular location – close observations receive larger weights than more distant observations. An adaptive kernel is intuitively sensible when zones vary in size and population density. With an adaptive bandwidth based on observation density or population number, the kernel is small in areas with large numbers of zones and large in areas with small numbers of zones. As an example, the kernel may be adapted to include the nearest ten observations to each location. In the study presented below, the bi-square weighting function (see Fotheringham et al., 2002) is used:

$$\begin{aligned} w_{ij} &= [1 - (d_{ij}/\tau)^2]^2 \text{ if } d_{ij} \le \tau \\ &= 0 \text{ otherwise} \end{aligned} \tag{4}$$

Where d_{ij} is the distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$, and τ is the distance to the nth nearest neighbour. Moran's *I* is here computed using adaptive bandwidths of different sizes; Fortran 77 code was written to compute Moran's *I*. Table IV contains Moran's *I* values and standard deviation for each log-ratio for 2001 and 2011. The standard deviations reflect, of course, variability in the log-ratios and, in 2001, they were largest for LRTenure1 and LREthnicity and smallest for LRage2 and LRLLTI. The relative ordering of largest and smallest standard deviations is the same for 2011, but, for these variables, all are smaller for 2011 than for 2001, with the exception of LRLLTI. As was the case for *D*, the standard deviations suggest less variation (and thus less spatial division) in the population in 2011 than in 2001. Taking the example of ethnicity, it is worth noting that the standard deviation of raw percentages (not shown) for 2001 is smaller than that for 2011 (the reverse situation for log-ratios). This is because the percentages have a very strongly skewed distribution, with a very large proportion of OAs having small percentages of non-White persons. The log-ratio has a distribution much closer to normal; the standard deviation is not meaningful in the case of the raw percentages, while it is a better summary of dispersion in the case of the log-ratios. Thus the standard deviations (and other statistical measures) derived from the log-ratios are considered to accurately reflect change in population characteristics (in this case, ethnicity, as supported by the *D* values). The results discussed so far are aspatial and make no reference to the spatial configuration of values — for example, large values of a log-ratio (or percentage) could be clustered or dispersed across England and Wales. The remainder of the analysis focuses on the analysis of the spatial structure of the selected variables, and the Moran's *I* results are discussed next.

TABLE IV ABOUT HERE

Moran's *I* was computed for 20 and 100 nearest neighbours, using the bi-square weighting function defined in Equation 4, and the values are given in Table IV. The largest values for both 20 and 100 nearest neighbours were for LREthnicity in 2001 and 2011, and, for both neighbourhood sizes, the values have increased. Moran's *I* was, for most log-ratios, larger in 2011 for both neighbourhood sizes than it was in 2001. This should be considered in the context of a reduction in the standard deviation between 2001 and 2011 for most log-ratios (i.e., seven out of 12; with the largest proportional decrease for the standard deviations of LRNSSEC1). Taking LREthnicity as an example, this suggests a reduction in concentrations of White persons as against non-White persons – that is, it suggests greater residential mixing (see Catney 2013), as will be elucidated below. The ratio of *I* for 100 nearest neighbours to *I* for 20 nearest neighbours (see Table IV) reflects spatial structure in the log-ratios. Where the value is close to one, this suggests that the variable concerned is clustered over (relatively) large areas. Where the ratio is closer to zero, this indicates that the variable is locally clustered (i.e., for a neighbourhood of 20 OAs), but much less so for a larger area (100 nearest OAs). The largest ratio figure is for LREthnicity for both 2001 and 2011, while the smallest figure is for LRage1 (in 2001) and LRTenure1 (in 2011). So, LREthnicity is spatially structured and the large ratio indicates continuity in values over quite large areas (the scale of variation in LREthnicity is considered in more depth below). In contrast, LRage1 and LRTenure1 are less clustered over small areas (*I* for 20 nearest neighbours is quite small in both cases) than are other variables, but they are also less clustered over larger areas (100 nearest neighbours) than other variables.

Systematically relating *I* to the standard deviations of the log-ratios provides a summary of local clustering (*I* for 20 nearest neighbours) as against variation across England and Wales (the standard deviation). Ranking *I* and the standard deviation from largest to smallest for 2001 and 2011 indicates that variables with relatively large *I* and standard deviation (in the top six of 12 for both *I* and standard deviation) include LREthnicity, LRNSSEC1, LRCarsVans, and LRTenure1. These findings suggest that the variables are found in fairly distinct clusters – some neighbourhoods are

internally quite homogenous, but the differences between neighbourhoods may be quite large. Variables with small *I* and standard deviation (in the bottom six of 12 for both *I* and standard deviation) include LREmploy, LRage2, LRage3 and LRLLTI. For variables with small *I* and small standard deviations, there is little or no indication of spatial structure – there is no (or minimal evidence for) spatial dependence (neighbouring values tend to be dissimilar) and there do not tend to be large differences between regions. So, the results indicate that there are larger between-region differences (stronger spatial structure) for the first set of variables (in the top six) than for the second (in the bottom six).

Comparison of clustering or spatial variation in the selected variables should take place with the recognition that interpretations are variable dependent. For example, we would expect, as discussed later, the population to be clustered by ethnicity. But, it is problematic to directly interpret this as indicating a higher level of segregation between, in this study, White persons and non-White persons than there is between those who are, for instance, employed or unemployed. The measures used in this study provide useful summaries of how the population of England and Wales is spatially distributed but it is essential to bear in mind the quite distinct processes which have led to the patterns observed and there may be much greater meaningful mixing 'on the ground' between White persons and non-White persons than between, for example, those in NS-SeC1 or NS-SeC8. Despite this caveat, it is possible to note that the population is more spatially structured by some variables than others. The analysis now moves onto assessment of variation over multiple spatial scales using the variogram.

Analysing spatial variation with the variogram

The variogram relates half the average of the squared differences between zones to the distances (in bins, or lags) separating their centroids (this is in contrast to the distance decay functions used for computing *I*). The variogram, $\gamma(\mathbf{h})$, provides a summary of spatial dependence at different spatial scales. The experimental variogram can be estimated for the $p(\mathbf{h})$ paired observations (here, log-ratios), $z(\mathbf{s}_i)$, $z(\mathbf{s}_i + \mathbf{h})$, $i = 1, 2, ..., p(\mathbf{h})$ with:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2\,p(\mathbf{h})} \sum_{i=1}^{p(\mathbf{h})} \{z(\mathbf{s}_i) - z(\mathbf{s}_i + \mathbf{h})\}^2 \qquad (5)$$

where $\mathbf{h}$ is the lag (distance and direction) by which two observations are separated. The correlogram (equivalent to spatially lagged *I*) could have been used instead for more direct comparison with *I*, but the variogram was preferred as a more widely-used structure function, and for consistency with previous research (i.e., Lloyd 2012). The use of population weighted variograms (where more weight is given to zones with larger populations; see Goovaerts et al, 2005) was assessed in this analysis, but standard unweighted variograms were considered appropriate in the study. It is important to note that the variogram is a function of the data support (the size and shape of the data zonal system (OAs in this case); see Lloyd, 2014), and determination of the point support variogram from the areal data variogram (e.g., variogram of OA data) is discussed by Goovaerts (2008). Pawlowsky and Burger (1992) and Pawlowsky-Glahn and Olea (2004) discuss the analysis of compositional data using variograms.

Variogram analysis (see Lloyd 2010b for an introduction) is often followed by fitting a mathematical model which is used to inform spatial prediction using kriging (see, for example, Webster and Oliver 2007). In this analysis, a model is fitted as end in itself as the model provides a summary of the spatial variation represented by the variogram. In this paper, the widely-used spherical model was applied. The spherical model is a bounded model – that is, the variance is

finite and the variogram reaches a plateau. A nugget effect, which represents measurement error and variation at a distance smaller than that represented by the sample spacing, was also fitted to the variograms in this study (see Lloyd 2012, 2014 for more details in a population analysis context). Figure 4 gives an example of a bounded variogram model with a nugget effect included. The range represents the spatial scale of variation, while the structured component represents spatially correlated variation. In practice, multiple structures may be fitted such that, for example, there may be a nugget effect, two ranges and two structured components. While models can be fitted by eye, a more common approach is to use a fitting procedure such as weighted least squares (WLS). Detailed accounts of the variogram and other structure functions are provided by Wackernagel (2003) and Webster and Oliver (2007). Variograms were estimated and modelled (with WLS) using the Gstat software (Pebesma and Wesseling 1998).

FIGURE 4 ABOUT HERE

Where variograms are estimated from the same variables for two Census years, the difference in the forms of the variogram provides a useful summary of change in the spatial structure of the population sub-group. Aside from no (or minimal) change, there are eight basic possibilities:

1. If the nugget decreases and the total sill increases then this suggest that the population sub-groups have become more polarised; that is, local areas have become more homogenous, but differences between these local areas are larger
2. If the nugget decreases and the total sill decreases then the population sub-groups are more mixed at all spatial scales
3. If the nugget increases and the total sill decreases then there is more variation locally, but less variation regionally (zones tend to be more different from one another over small areas, but there is less difference, on average, between regions).
4. If the nugget increases and the total sill increases then there is more variation at all spatial scales (zones are more different from one another)

The interpretations of the remaining four scenarios follow easily from the above:

5. Increased nugget, sill remains similar
6. Decreased nugget, sill remains similar
7. Nugget remains similar, sill increased
8. Nugget remains similar, sill decreased

Variograms were estimated for three contrasting cases (as suggested by the preceding analyses): LREthnicity, LRTenure1, and LRLLTI; in all cases a lag size of 2km was used. The variograms for LREthnicity (Figure 5A), for both 2001 and 2011, show very strong spatial structure, as indicated by small nugget effects and large total sills. The models, comprising for both years a nugget effect and two spherical model components, are good fits. The two variograms indicate that clustering over small areas increased between 2001 and 2001 (this is equivalent to an increase in Moran's *I*), but that the variance (as indicated by the total sill) decreased. So, there was strong spatial structure in both years, but regions are becoming more similar (this is consistent with the standard deviation). These changes correspond to category 2 – more mixing at all spatial scales. The range components for the first structured component of the model are for distances of approximately 6km for both Census years while the ranges for the second components are 51km (2001) and 56km (2011). These suggest variation across urban areas (6km range) and between urban areas (or regionally; 51km and 56km). That they are so similar for the two Census years indicates that the *scale* of variation has remained similar, but a reduced total sill indicates that the *magnitude* of variation has decreased. By conveying such information, the variogram adds to the previous analyses based

on *D*, *I* and the standard deviation of the log-ratios. The variograms for LRTenure1 (Owner occupied and Private rented HH / Social rented HH) (Figure 5B) suggest less spatial structure for LRTenure1 than for LREthnicity in that the difference between the nugget and the total sill is small for both 2001 and 2011. The nuggets (reflecting local clustering) for 2001 and 2011 are quite similar, while the total sill (i.e., the variance) for 2011 is smaller than the total sill for 2001.

FIGURE 5 ABOUT HERE

For LRTenure1, clustering over small areas increased slightly between 2001 and 2001 (smaller nugget in 2011 and also larger *I*), but the variance decreased (matching most closely scenario 8 above). In short, in 2011 regional variation in the population by tenure (specifically, LRTenure1) was smaller than in 2001; so, localised clustering remained similar but differences between regions decreased. The variograms for LRLLTI (Figure 5C) show little evidence of spatial structure. There is some suggestion of clustering at a local scale (small semivariances at smaller lags, suggestive of variation in urban areas). Models were not fitted in these cases as they would likely be best represented by pure nugget models (a model parallel with the *x* axis; see Webster and Oliver 2007), indicating zero spatial structure. The variance increased between 2001 and 2011 suggesting that differences between OAs have, on average, increased. This is supported by an increase in *D* and in the standard deviation of LRLLTI between 2001 and 2011. The variance increased at all spatial scales, and this corresponds most closely to scenario 4 (although recall that no model was fitted to the variogram). Note that there is a large scale trend in LLTI – the south and east of England and Wales as against the north and west – but spatial structure is weak at the scales considered here.

As noted previously, the nugget effect indicates measurement error and variation at a distance smaller than that which is represented by the sample. For the variograms of LRTenure1 and LRLLTI, the nugget effects are a large proportion of the total sill. This implies that there may be a large amount of variation at distances smaller than 2km (the lag size used in estimating the variograms). Estimating variograms using smaller lag sizes (in this case, 500m lags were used) allows assessment of spatial variation in some urban areas (where OAs are small) and shows that both LRTenure1 and LRLLTI are spatially structured over relatively small distances (up to around 5km for LRLLTI, but over larger distances for LRTenure1, as suggested by Figure 5B) (these variograms with smaller lag sizes are not shown for reasons of space). These variables are spatially structured in some regions over larger scales and use of local variograms (see Lloyd 2012) would allow this to be assessed further.

Discussion and conclusions

Two key contributions of the paper were outlined in the introduction, and these related to (i) detailing approaches for the analysis of spatial structure of population sub-groups and (ii) providing an analysis of the spatial distribution of population sub-groups in England and Wales in 2001 and 2011. In terms of the first contribution, the paper considered several measures which can be used to analyse population distributions; these included the index of dissimilarity (*D*), the Moran's *I* autocorrelation coefficient and the variogram. These provide measures of unevenness (*D*), clustering (*I*) and spatial structure (spatial dependence) at different scales (the variogram) and the second contribution detailed the rich information offered by this combination of measures in the case of England and Wales in the 2000's.

The analysis reveals how the population of England and Wales varies by several demographic and socio-economic characteristics. It is shown that the population tends to be more evenly distributed and less clustered by age than by the other variables. Between 2001 and 2011 unevenness in most population sub-groups in England and Wales reduced. Over the same period, there was an increase in localised clustering in the population by most of the demographic and socio-economic variables

assessed. Other research has pointed to an increase in, or persistence of, inequalities in the UK over the last decade. Tunstall (2011) shows that some dimensions of social exclusion (namely income, employment, and neighbourhood quality) for those in social housing had seen a small reduction in England between 2000 and 2011, but that there were suggestions of an increase in concentrations of disability. Hills et al (2010) argue that there remain major inequalities in earnings and incomes, and that, while there has been some narrowing over the last decade, large social gaps between the earnings of men and women with respect to the educational attainments of different ethnic groups. Certainly, England and Wales (and the UK as a whole) continues to be socially and economically divided but, at least in terms of the quite coarse categorisations of sub-groups employed in this paper, there is evidence for a decline, on average, in *spatial* differences with less distinct concentrations of population sub-groups and smaller differences between regions. In short, the population is, in at least some key respects, becoming more similar and the population features which makes some regions different from others have changed over the last ten years through processes of migration, population growth/decline, social and economic transitions by individuals, and external influences such as recession. With respect to unemployment, it is worth noting that during, or after, a recession it could appear that inequality has decreased as unemployment rises in areas where formally it had been very low. It is not possible to deconstruct the nature of these changes using data for static time points (i.e., 2001 and 2011 Census data), and obviously motivations cannot be discerned using Census data, but the paper reveals some considerable changes in the population distribution of England and Wales over the last decade.

Taken together, the findings suggest that local areas have become more similar but, for many variables, this is against a background of reduced regional variation. The example of ethnicity is used to illustrate this point. In terms of ethnicity, variation reduced at all spatial scales between 2001 and 2011. The reduction in Moran's *I* and the standard deviation suggests, in this case, a smoothing effect between the two Census years — places are, in general, becoming more similar in terms of their proportion of White and non-White residents. Given the increase in the non-White share of the population, the reduction in *I* does not suggest increasing clustering of White persons or non-White persons – rather it indicates that local areas are becoming more similar, but with a larger proportion of non-White persons. For example, in 2001, an area may have had non-White percentages in the range 5% to 25%, but in 2011 the range could be 15% to 25% — the range of values has decreased and the area is more homogeneous with a smaller value of *I* with respect to values in zones in this area. These findings are consistent with an interpretation of dispersal by members of ethnic minority groups from immigrant settlement areas, and common migration patterns for those in similar socio-economic classes irrespective of ethnic group (Simpson and Finney 2009; Catney and Simpson 2010). New migration streams from the 2004 European Union accession countries will also have contributed to changes in the spatial distribution of the (mostly White) population.

Taking the findings about the ethnicity variable further, the variogram contains information on spatial variation at multiple scales and it supports the conclusions reached using *I* and the variance. To help put the findings in context, it is useful to refer to a contrasting case. Lloyd (2012), in a study of spatial variation by religion in Northern Ireland, found that the log-ratio of Catholics to non-Catholics became more clustered locally between 1971 and 1991, but that the variance (indicated by the variogram total sill) also increased. This was interpreted as evidence that local areas were becoming more similar, but that differences between these areas were increasing – Catholic areas were becoming more Catholic, while non-Catholic areas were becoming more non Catholic. In other words, the population become more polarised by religion. This was defined is indicating an increase in both microsegregation (clustering locally) and macrosegregation

(increased difference between localities)[4]. Between 1991 and 2001 the nugget decreased but the total sill was very similar, indicating little change between these years. In the present study, the smaller variance in 2011 than in 2001 indicates that White and non-White areas are becoming more similar, as also indicated by *D*. It is worth reiterating that if the summary statistics (Table IV) (and also the variograms) were computed for raw percentages instead of log-ratios then the results may not be consistent (e.g., increases rather than decreases in the standard deviation could be indicated). This is largely because the percentages are, at least in some cases, strongly skewed and this alone makes the analysis of raw percentages problematic.

The results for the White/non-White variable are consistent with the ranking of *D* values presented by Voas and Williamson (2000), and they reflect concentrations of members of several non-White groups in large urban areas including London. However, the results outlined here present a different perspective on scales of spatial variation which is linked to a neighbourhood derived using a distance decay function (as used here in computing Moran's *I*), rather than a nested hierarchy of zones. In addition, the variogram is based on Euclidean distances between points (here, zone centroids). *D* and *I* may be completely unrelated. While *D* relates to variance and *I* to clustering at different scales (thus, it can be said to be a spatial measure of variance (or more accurately covariance)), the spatial independence of *D* means that a variable could have a large variance (large *D*) while being, for example, either clustered (large positive *I*) or dispersed (*I* close to zero). Note that spatial versions of *D* have been developed and these could also be applied (see, for example, Shuttleworth et al., 2011).

The kind of analysis outlined here is useful in providing information about clustering (how far people (as recorded in zones) with particular characteristics tend to live close to others with similar characteristics), and about the spatial scales over which this clustering persists (i.e., spatial structure). This may provide valuable information in policy contexts in that targeting resources to alleviate particular problems, such as particular forms of deprivation, may need to operate over different spatial scales. In addition, these approaches can provide summaries of how a population has changed. One benefit of geographically-weighted approaches like Moran's *I*, as applied here, is that they are quite robust to changes in the zonal system. If the number of zones (e.g., wards) does not change markedly between one Census and another, the results of a geographically-weighted Moran's *I* analysis are likely to be comparable and, thus, the results presented here could be compared with those from an analysis of, for example, 1991 data as well as those for 2001 and 2011.

The analysis uses quite broad categories such as White/non-White. Of course, the categories could have been subdivided and there are likely to be considerable differences between subsets of the groups used here (for example, see Catney 2013 with respect to ethnic groups). Further subdivisions of, for example, the age, tenure, qualifications and NS-SeC categories used here are likely to be beneficial and allow a fuller assessment of the distribution of the population of England and Wales. In addition, the range of variables used in the analysis could be expanded; the profile of the population by religion was a particular focus for the media following the release of 2011 Census data in the UK, and the spatial structure of religion could also be assessed. Other variables such as marital status and general health might also usefully be included in a future analysis which builds on the principles established here. In addition, the inclusion of, for example, variables such as the numbers of households for three or more cars, or houses with many rooms, may help to better assess geographical inequalities between the most wealthy and the rest of the population.

[4] Using these definitions, microsegregation and macrosegregation refer to different dimensions of segregation (see Massey and Denton, 1988), as opposed to the use of these terms to describe one dimension at different spatial scales (see Reardon et al., 2008, 2009)

The analyses of spatial structure (particularly as evidenced by the variograms) suggest that spatial variation in many population sub-groups is at a very fine spatial scale. Thus, zones which are larger than OAs may be insufficient to capture important variation in these groups. Given debates about the future of the UK Census (see ONS 2013b), such information is potentially useful and there is scope to use these approaches to determine how much information might be lost if future survey mechanisms are less rich in terms of attribute and spatial detail. The results, though provisional, suggest than zones larger than OAs (and which are designed to be internally homogeneous; see Data section) would be insufficient to represent important spatial variations in population sub-groups and that, therefore, any replacement for the Census would need to provide counts over similarly small areas. The variograms for Tenure1 and LLTI log-ratios indicate suggest that spatial variation in these variables is found over very short distances and thus, zones with larger widths (on average) than these distances would be too large to resolve this spatial variation. The analyses presented here are based on univariate data and it is possible that sources other than the Census may provide sufficient information on some population characteristics. For example, Norman and Bambra (2007) suggest that sickness benefit could serve as a regularly updatable indicator of health over small areas. The methods detailed in this paper could provide the basis for the assessment of how much information is lost by moving from one set of zones to another coarser set, and this is a focus for ongoing research.

Future work will include increasing the range of variables analysed, assessment of alternative approaches for characterising spatial structure in individual variables, and extension of the analysis to multivariate frameworks. Direct analyses of raw percentages is problematic even in the analysis of single variables, as well as has been often demonstrated for analysis of multiple variables, thus the use of log-ratios was preferred in this study. There is scope to further assess the relative pros and cons of analyses based on raw percentages and log-ratios. Adding a constant to counts as in this study, to allow log-ratios to be computed, reduces variation and alternative strategies should be considered. Another area for consideration relates to local variations in population sub-group distributions. In some cases, small levels of measured clustering reflect high levels of clustering in some areas, but high variability elsewhere — social housing is a key example. Thus, the use of local measures of clustering (see Anselin 1995; Johnston et al., 2011; Lloyd 2011) will help to deconstruct spatial variations by population sub-groups. This paper suggests that, contrary to the case for Britain over the period 1971–2001 (Dorling and Rees, 2003), and despite the economic difficulties of the late 2000s, spatial divisions between most population sub-groups reduced in England and Wales between 2001 and 2011.

Acknowledgements

The Office for National Statistics are thanked for provision of the data. Office for National Statistics, 2001 and 2011 Census: Digitised Boundary Data (England and Wales) [computer file]. ESRC/JISC Census Programme, Census Geography Data Unit (UKBORDERS), EDINA (University of Edinburgh)/Census Dissemination Unit. Census output is Crown copyright and is reproduced with the permission of the Controller of HMSO and the Queen's Printer for Scotland. The editor and three anonymous reviewers are thanked for their very helpful comments on an earlier version of the paper.

References

Aitchison J 1986 *The Statistical Analysis of Compositional Data* Chapman and Hall, London

Anselin L 1995 Local indicators of spatial association — LISA *Geographical Analysis* 27 93–115

Catney G 2013 Has neighbourhood ethnic segregation decreased? *The Dynamics of Diversity: Evidence from the 2011 Census* Centre on Dynamics of Ethnicity, Manchester (http://www.ethnicity.ac.uk/census/885_CCSR_Neighbourhood_Bulletin_v7.pdf) Accessed 11 July 2013

Catney G and Simpson L 2010 Settlement area migration in England and Wales: assessing evidence for a social gradient. *Transactions of the Institute of British Geographers*, 35, 571–84

Champion T, Wong C, Rooke A, Dorling D, Coombes M and Brunsdon C 1996 *The population of Britain in the 1990s: A social and economic atlas* Clarendon Press, Oxford.

Cliff A D and Ord J K 1973 *Spatial autocorrelation* Pion, London

Dorling D and Rees P 2003 A nation still dividing: the British census and social polarisation 1971–2001 *Environment and Planning A* 35 1287–1313

Dorling D and Thomas B 2004 *People and places: A 2001 Census atlas of the UK* The Policy Press, Bristol.

Egozcue J J, Pawlowsky-Glahn V, Mateu-Figueras G and Barcelo-Vidal C 2003 Isometric logratio transformations for compositional data analysis *Mathematical Geology* 35 279–300

Egozcue J J and Pawlowsky-Glahn V 2005 Groups of parts and their balances in compositional data analysis *Mathematical Geology* 37 773–93

Egozcue J J and Pawlowsky-Glahn V 2006 Simplicial geometry for compositional data in **Buccianti A, Mateu-Figueras G and Pawlowsky-Glahn V** eds *Compositional Data Analysis in the Geosciences: From Theory to Practice* Geological Society Special Publications No 264 Geological Society, London 145–60

Filzmoser P, Hron K and Reimann C 2009 Univariate statistical analysis of environmental (compositional) data: problems and possibilities *Science of the Total Environment* 407 6100–8

Fotheringham A S, Brunsdon C and Charlton M 2002 *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* John Wiley and Sons, Chichester

Goovaerts P 2008 Kriging and semivariogram deconvolution in the presence of irregular geographical units *Mathematical Geosciences* 40 101–28

Goovaerts P, Jacquez G M and Greiling D 2005. Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms *Geographical Analysis* 37 152–82

Hills J, Brewer M, Jenkins S, Lister R, Lupton R, Machin S, Mills C, Modood T, Rees T and Riddell S 2010 *An Anatomy of Economic Inequality in the UK: Report of the National Equality Panel* CASE report 60 Centre for Analysis of Social Exclusion, The London School of Economics and Political Science, London

Johnston R, Poulsen M and Forrest J 2011 Evaluating changing residential segregation in Auckland, New Zealand, using spatial statistics *Tijdschrift voor Economische en Sociale Geografie* 102 1–23

Lloyd C D 2010a Exploring population spatial concentrations in Northern Ireland by community background and other characteristics: an application of geographically weighted spatial statistics *International Journal of Geographical Information Science* 24 1193–1221

Lloyd C D 2010b *Spatial Data Analysis: An Introduction for GIS Users* Oxford University Press, Oxford

Lloyd C D 2011 *Local Models for Spatial Analysis* Second Edition CRC Press, Boca Raton

Lloyd C D 2012 Analysing the spatial scale of population concentrations in Northern Ireland using global and local variograms *International Journal of Geographical Information Science* 26 57–73

Lloyd C D 2014 *Exploring Spatial Scale in Geography* John Wiley and Sons, Chichester, in press

Lloyd C D, Pawlowsky-Glahn V and Egozcue J J 2012 Compositional data analysis for population studies *Annals of the Association of American Geographers* 102 1251–66.

Martin D, Nolan A and Tranmer M 2001 The application of zone-design methodology in the 2001 UK Census *Environment and Planning A* 33 1949–62

Massey D S and Denton N A 1988 The dimensions of residential segregation *Social Forces* 67 281–315

Moran P A P 1950 Notes on continuous stochastic phenomena *Biometrika* 37 17–23

Norman P 2010 Demographic and deprivation change in the UK, 1991–2001 in **Stillwell J, Norman P, Thomas C and Surridge P** eds *Spatial and Social Disparities* Understanding Population Trends and Processes Volume 2 Springer, Dordrecht 17–35

Norman P and Bambra C 2007 Incapacity or unemployment? The utility of an administrative data source as an updatable indicator of population health *Population, Space and Place* 13 333–352

ONS (Office for National Statistics) 2013a 2011 *Census Analysis, A Century of Home Ownership and Renting in England and Wales* (http://www.ons.gov.uk/ons/rel/census/2011-census-analysis/a-century-of-home-ownership-and-renting-in-england-and-wales/index.html) Accessed 20 May 2013

ONS (Office for National Statistics) 2013b 2011 *The Census and Future Provision of Population Statistics in England and Wales: Public Consultation* (http://www.ons.gov.uk/ons/about-ons/get-involved/consultations/consultations/beyond-2011-consultation/beyond-2011-consultation-doc-c1.pdf) Accessed 16 January 2014

Pawlowsky V and Burger H 1992 Spatial structure analysis of regionalized compositions *Mathematical Geology* 24 675–91

Pawlowsky-Glahn V and Olea RA 2004 *Geostatistical Analysis of Compositional Data* Oxford University Press, New York

Pebesma E J and Wesseling C G 1998 Gstat, a program for geostatistical modelling, prediction and simulation *Computers and Geosciences* 24 17–31

Reardon S F, Matthews S A, O'Sullivan D, Lee B A, Firebaugh G, Farrell C R and Bischoff K 2008 The geographic scale of metropolitan racial segregation *Demography* 45 489–514

Reardon S F, Farrell C R, Matthews S A, O'Sullivan D, Bischoff K and Firebaugh G 2009 Race and space in the 1990s: changes in the geographic scale of racial residential segregation, 1990–2000 *Social Science Research* 38 55–70

Shuttleworth I G, Lloyd C D and Martin D J (2011) Exploring the implications of changing census output geographies for the measurement of residential segregation: the example of Northern Ireland 1991–2001 *Journal of the Royal Statistical Society: Series A* 174 1–16

Simpson L and Finney N 2009 Spatial patterns of internal migration: evidence for ethnic groups in Britain *Population, Space and Place* 15 37–56

Thomson D and Knight T with Buscha F, Urwin P and Sturgis P 2010 *Research into Measuring Adult Attainment Using the Labour Force Survey: Final Report* RM Data Solutions

Thió-Henestrosa S and Martín-Fernández J A 2005 Dealing with compositional data: the freeware CoDaPack *Mathematical Geology* 37 773–93

Tunstall R 2011 Social housing and social exclusion 2000–2011. *CASE Paper*, no. 153 Centre for Analysis of Social Exclusion, London School of Economics and Political Science

van den Boogaart K G and Tolosano-Delgado R 2008 "compositions": A unified R package to analyze compositional data *Computers and Geosciences* 34 320–38

Voas D and Williamson P 2000 The scale of dissimilarity: concepts, measurement and an application to socio-economic variation across England and Wales *Transactions of the Institute of British Geographers*, *New Series* 25 465–81

Wackernagel H 2003 *Multivariate Geostatistics: An Introduction with Applications* Third edition Springer, Berlin

Webster R and Oliver M A 2007 *Geostatistics for Environmental Scientists* Second Edition John Wiley and Sons, Chichester

Wong D W S 2004 Comparing traditional and spatial segregation measures: a spatial scale perspective *Urban Geography* 25 66–82.

Table I Key Statistics Census tables and derived variables

Table 2001	Table 2011	Table description	Description
KS002	KS102	Age structure	Age 0 to 15; 16 to 29; 30 to 64; 65 plus
KS006	KS201	Ethnic group	All Whites; Non-Whites
KS018	KS402	Housing tenure	Owner occupied; Social rented*; Private rented
KS017	KS404	Cars and vans	Cars or vans; No cars or vans
KS013	KS501	Qualifications and students	No qualifications; Qualifications**
KS09A	KS601	Economic activity — all persons (aged 16-74)	Unemployed economically active; Employed economically active
KS008	KS301	Health and provision of unpaid care	LLTI; Non LLTI
KS14A	KS611	NS-SeC (persons aged 16-74)	NS-SeC1,2; 3 to 7; 8

*Council (local authority), Housing Association or Registered Social Landlord
**2001 population was all persons aged 16-74; 2011 population was all persons aged 16 plus

Table II Counts and percentages for 2001 and 2011

Variable	Definition	2001	2001 %	2011	2011 %	% change
A0to15	Persons aged 0 to 15	10488725	20.15	10579132	18.87	-1.29
A16to29	Persons aged 16 to 29	9112104	17.51	10495245	18.72	1.21
A30to64	Persons aged 30 to 64	24127561	46.36	25778462	45.97	-0.39
A65plus	Persons aged 65 plus	8313626	15.97	9223073	16.45	0.47
White	White persons	47521002	91.31	48209395	85.97	-5.34
Non-White	Non-White persons	4520653	8.69	7866517	14.03	5.34
OwnOcc	Owner occupied HH	14916407	68.86	15031914	64.33	-4.53
SocRent	Social rented HH	4157658	19.19	4118461	17.63	-1.57
PrivRent	Private rented HH	2586617	11.94	4215669	18.04	6.10
CarsVans	HH with cars or vans	15859121	73.21	17376274	74.37	1.15
NoCarsVans	HH with no cars or vans	5802283	26.79	5989770	25.63	-1.15
Qual*	Persons with qualifications	*26670364*	70.92	*35189453*	77.34	6.43
NoQual*	Persons with no qualifications	*10937039*	29.08	*10307327*	22.66	-6.43
EAEmploy	EA employed persons	22795512	94.76	25449863	93.40	-1.36
EAUnemp	EA unemployed persons	1260880	5.24	1799536	6.60	1.36
NSSEC12	NS-SeC 1,2	10173130	36.04	12792224	34.19	-1.85
NSSEC37	NS-SeC 3-7	16650680	58.99	22324839	59.66	0.68
NSSEC8	NS-SeC 8	1404083	4.97	2301614	6.15	1.18
No LLTI	Persons with no LLTI	42557060	81.77	46027471	82.08	0.31
LLTI	Persons with a LLTI	9484856	18.23	10048441	17.92	-0.31

HH are households; EA is economically active. NB: Figures are sums of OA counts (for 2001 this means there is inconsistency with, for example, national-level counts due to small cell adjustment); PrivRent for 2001 includes 'Private landlord or letting agency' and 'Other'; PrivRent for 2011 includes 'Private rented: Private landlord or letting agency', 'Private rented: Other' and 'Living rent free'. *As noted in Table I, Qual and NoQual figures (in italics) for 2001 and 2011 use 16-74 and 16 plus population bases respectively, and so should not be directly compared.

Table III Index of dissimilarity, *D* (for stated sub-group versus remainder in category (e.g., A0to15 versus all others by age and White vs Non-White))

Variable (vs. rest)	2001	2011
A0to15	0.159	0.161
A16to29	0.197	0.208
A30to64	0.110	0.102
A65plus	0.258	0.274
White	0.623	0.592
OwnOcc	0.491	0.446
SocRent	0.613	0.592
PrivRent	0.384	0.371
NoCarsVans	0.391	0.402
NoQual	0.260	0.255
EAUnEmploy	0.329	0.300
NSSEC12	0.273	0.265
NSSEC37	0.230	0.207
NSSEC8	0.429	0.374
LLTI	0.197	0.199

Table IV Moran's *I* and standard deviation of log-ratios

	LA 20 NN **2001**	LA 100 NN **2001**	LA 20 NN **2011**	LA 100 NN **2011**	Ratio: 100/20 NN **2001**	Ratio: 100/20 NN **2011**	SD **2001**	SD **2011**
LRAge1	0.345	0.216	0.449	0.322	0.626	0.717	0.692	0.731
LRAge2	0.395	0.279	0.440	0.322	0.705	0.732	0.301	0.294
LRAge3	0.431	0.336	0.524	0.418	0.779	0.797	0.411	0.418
LREthnicity	0.754	0.718	0.841	0.797	0.953	0.948	1.168	1.133
LRTenure1	0.388	0.256	0.400	0.249	0.659	0.621	1.409	1.309
LRTenure2	0.475	0.361	0.578	0.457	0.761	0.791	0.832	0.739
LRCarsVans	0.585	0.450	0.645	0.518	0.770	0.802	0.776	0.776
LRQual	0.598	0.470	0.564	0.439	0.786	0.779	0.493	0.507
LREmploy	0.445	0.361	0.475	0.368	0.811	0.776	0.602	0.522
LRNSSEC1	0.490	0.397	0.592	0.477	0.810	0.805	0.939	0.771
LRNSSEC2	0.665	0.534	0.698	0.565	0.804	0.809	0.499	0.479
LRLLTI	0.397	0.298	0.411	0.308	0.750	0.749	0.371	0.388

LA is locally adaptive; NN is nearest neighbours; Variable names indicate LR (log-ratio) and labels used in the text

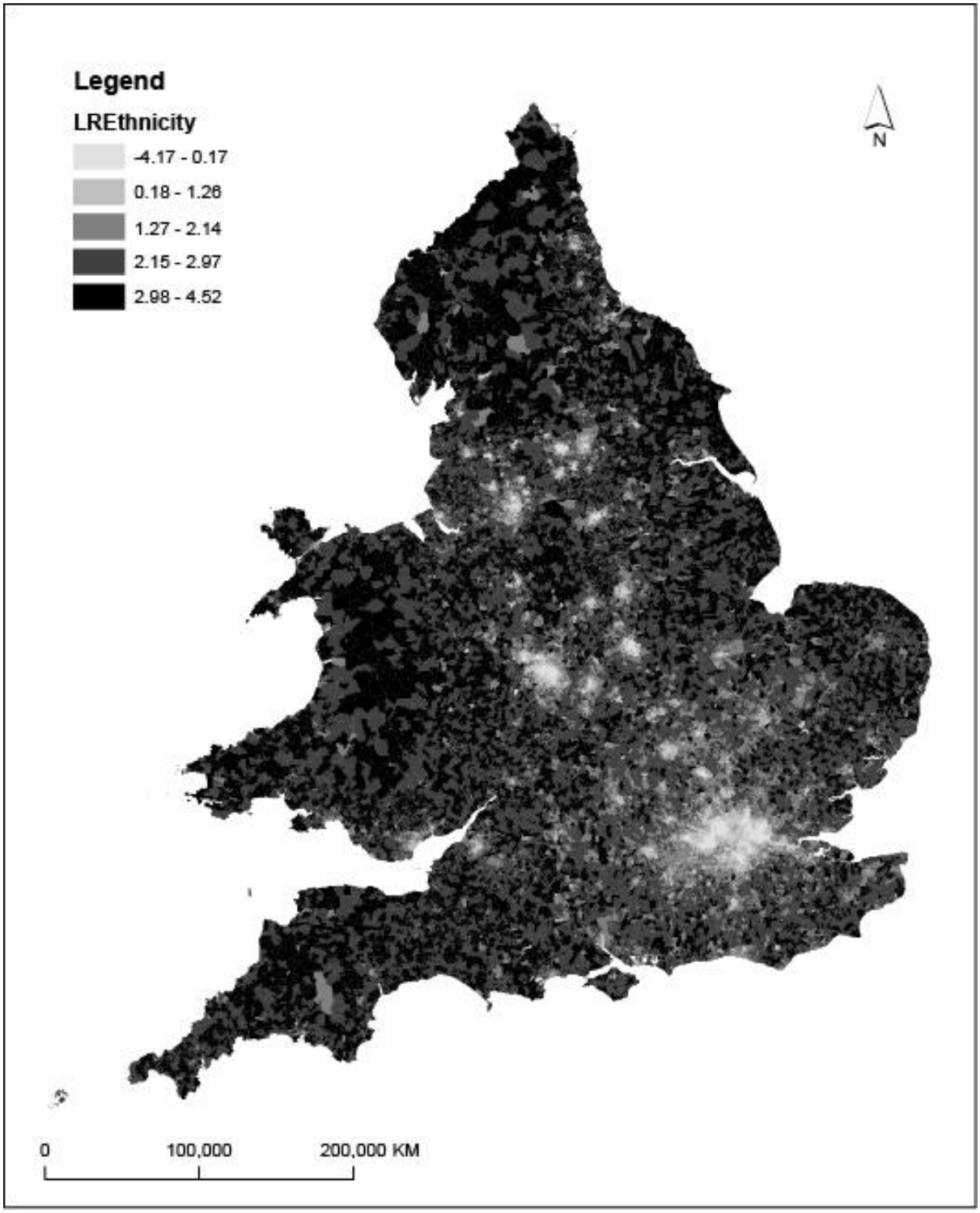

Figure 1 Ethnicity log-ratios, 2011. Contains Ordnance Survey data © Crown copyright and database right 2011

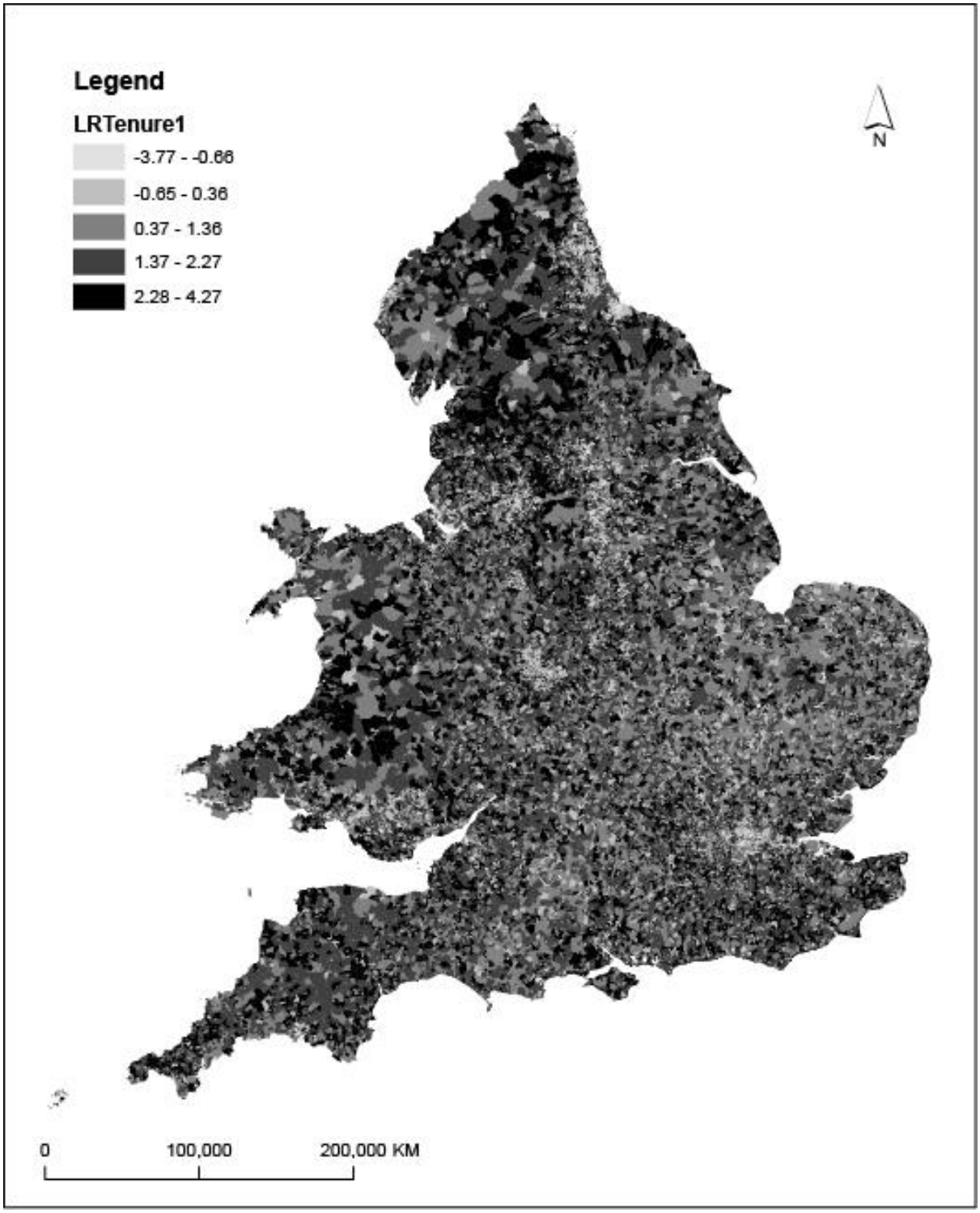

Figure 2 Tenure1 log-ratios, 2011. Contains Ordnance Survey data © Crown copyright and database right 2011

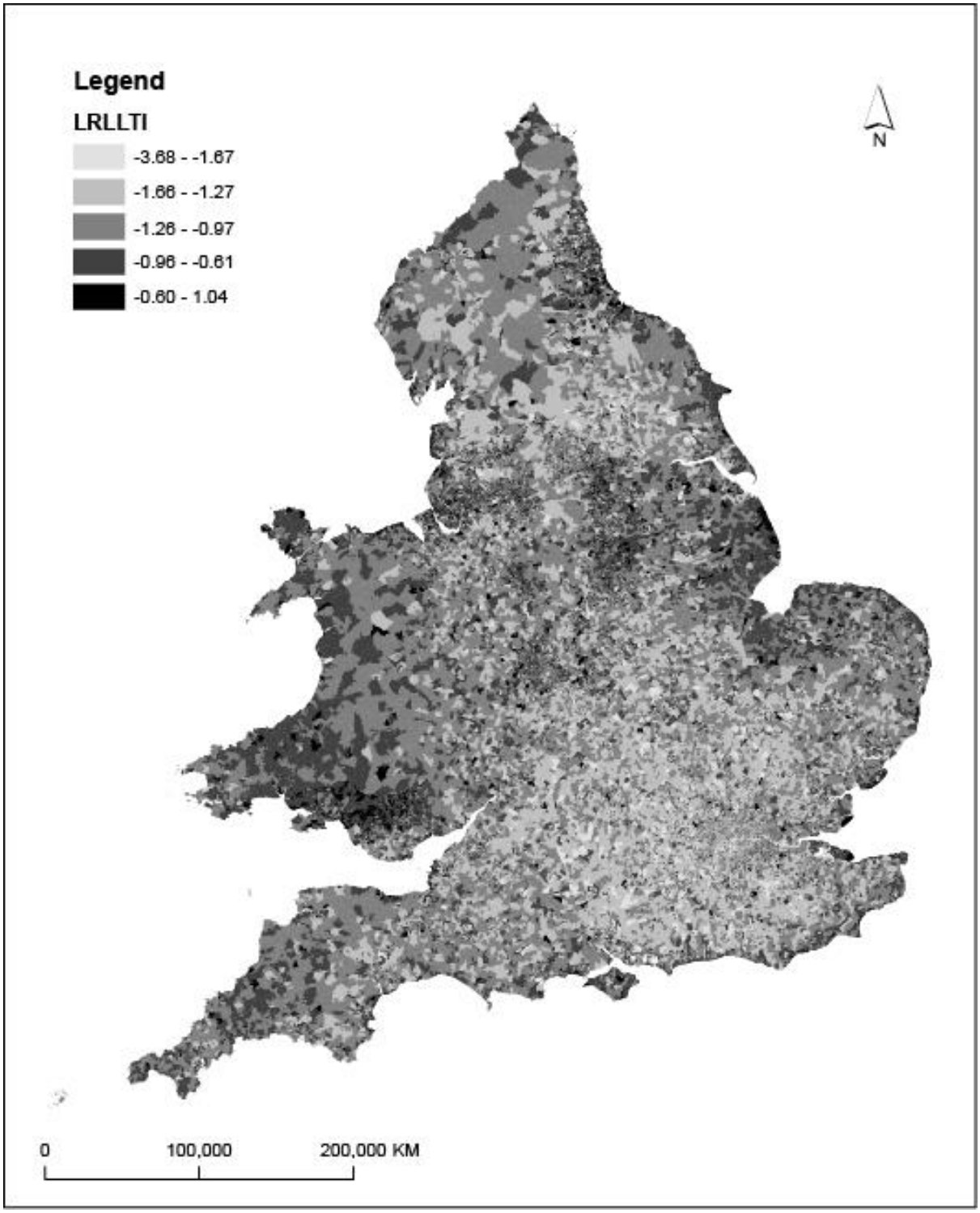

Figure 3 LLTI log-ratios, 2011. Contains Ordnance Survey data © Crown copyright and database right 2011

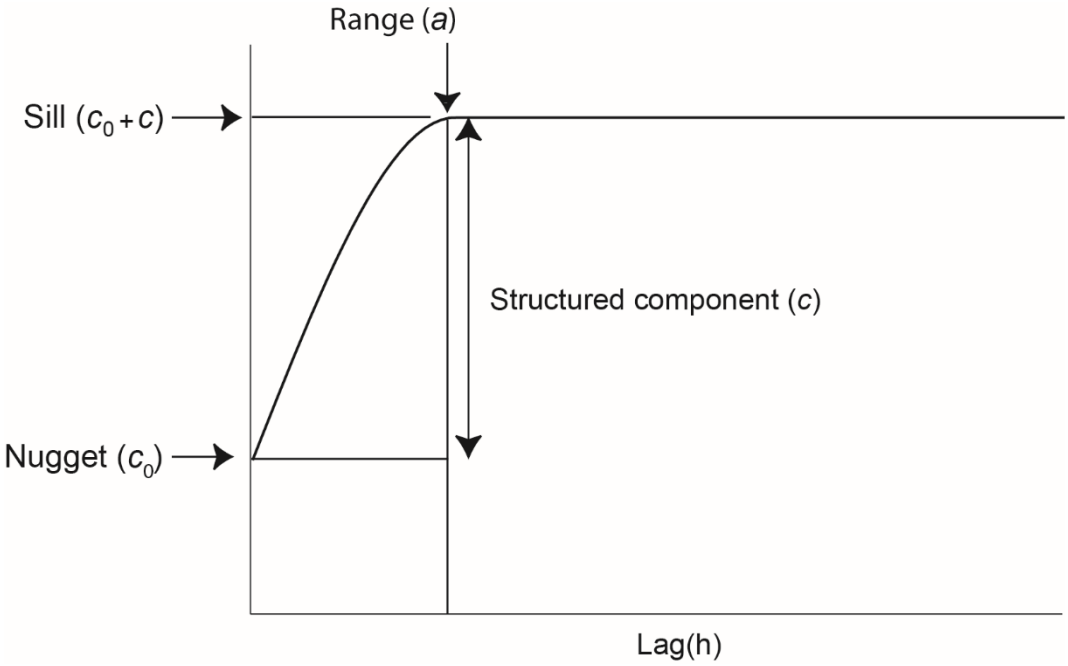

Figure 4 Bounded variogram model

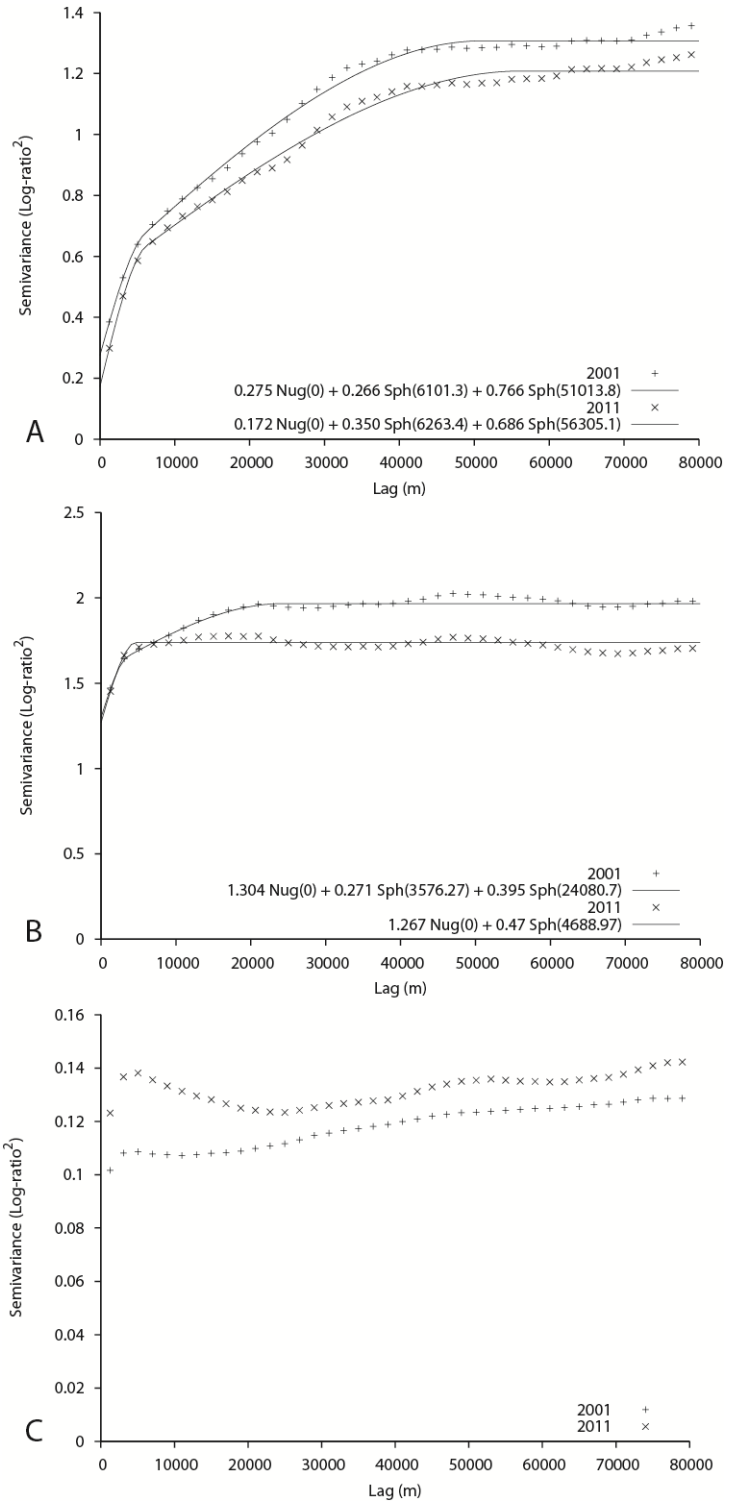

Figure 5 Variograms: (A) Ethnicity log-ratios, (B) Tenure1 log-ratios, (C) LLTI log-ratios, 2001 and 2011