Assessing the Spatial Structure of Population Variables in England

Christopher D. Lloyd[1]

[1]Department of Geography and Planning, School of Environmental Sciences,
University of Liverpool, Roxby Building, Chatham Street, Liverpool, L69 7ZT UK
Tel. (+44) 0151 794 2857
C.D.Lloyd@Liverpool.ac.uk

KEYWORDS: Spatial dependence, Scale, Census data, Moran's *I*

1. Introduction

Characterisation of the spatial structure of population variables is important for several reasons. Firstly, knowledge of how population subgroups are distributed across spatial scales has direct links to several research areas including analyses of deprivation, residential segregation and health status. 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. This paper explores the spatial scale of variation in population subgroups across England in 2001. The variables relate to age, ethnicity, health status, employment status, qualifications and housing tenure. The contributions made by the paper are twofold – (i) approaches for the analysis of spatial structure of population subgroups are detailed and (ii) the results enhance the understanding of the characteristics of population subgroups in England.

Previous research on the spatial structure of population subgroups has compared values of the index of dissimilarity, *D*, for three different zonal systems (enumeration districts, wards and districts) for multiple demographic and socioeconomic variables across England and Wales (Voas and Williamson, 2000). In that case, scale was conceptualised as relating to variation across and within each of the sets of zones. In the present paper, an alternative conceptualisation is used and the research builds on the work of Lloyd (2010), who used the geographically-weighted Moran's *I* autocorrelation coefficient to assess how spatial dependence of several subgroup 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. In the present paper, a similar analytical framework is extended to the analysis of demographic and socioeconomic variables derived from 2001 Census data for England.

2. Data and methods

2.1 Data

The variables are for Census Area Statistics (CAS) Wards and they relate to age, ethnic group, limiting long term illness (LLTI), unemployment, qualifications and housing tenure. The variables are derived from the Key Statistics tables, and they are specified in Table 1.

Table 1. Key Statistics Census Tables and Derived Variables

Table	Table description	Variables	Description
KS002	Age structure	14:17/1	Over 65s/all
KS006	Ethnic group	2:4/1	Whites/all
KS008	Health and provision of unpaid care	2:3/1	LLTI/all
KS09A	Economic activity — all persons	5/2:5	Unemployed/all economically active employed or unemployed
KS013	Qualifications and students	2/1	No qualifications/All people aged 16-74
KS018	Tenure	5:6/1	Social rented*/all households

*Council (local authority), Housing Association or Registered Social Landlord

Statistical analyses of raw percentages are problematic (Lloyd et al, 2012) and so the analysis presented in this paper makes use of an appropriate transform of the percentages, although raw percentages are also used for comparative purposes. In this study, the variables are two-part and the percentages of one group (e.g., those with a LLTI) are given by x_1 and the percentage of all others (e.g., all non LLTI) is given by x_2. There were zero values in some categories and, for computing log-ratios, the proportions are calculated from counts n_1, n_2 with n_1+1 and n_2+1 (see Lloyd 2010 for a justification of this approach). That is, a value of one is added to both sets of counts and the two percentages x_1, x_2 are calculated from the modified counts. Raw percentages computed from the raw counts (i.e., one is not added to the counts) are utilised for comparative purposes. The log-ratios, y_1, are computed with:

$$y_1 = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2} \tag{1}$$

Variables transformed in this way can be analysed using standard statistical methods. The log-ratios are computed with x_1 corresponding to the first variable in the last column of Table 1 and x_2 equalling the remainder of 100 (that is, $100 - x_1$).

2.2 Methods

The Moran's I coefficient with weights, w_{ij}, between locations i and j row-standardised (i.e., they sum to one) can be given by:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

where the values y_i have the mean $\bar{y}$ (I is computed for log-ratios and raw percentages). In this analysis, the Moran's I coefficient was computed using the Gaussian weighting scheme (Fotheringham *et al*., 2002):

$$w_{ij} = \exp[-0.5(d_{ij}/\tau)^2] \tag{3}$$

where d_{ij} is the distance between locations i and j and τ is the bandwidth which determines the size of the kernel. In the present study, I is computed for several bandwidths to allow assessment of spatial dependence over different spatial scales.

$$I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{4}$$

The use of y_i indicates that log-ratios are used, although raw percentages are also used for comparison.

3. Analysis

Figure 1 shows (a) LLTI (%) and (b) social rented households (%) for comparative purposes. It is clear from visual examination of the maps that the structure of the two variables is quite different. For example, the small LLTI percentages in the south east are apparent, while there is no such clear pattern in terms of social rented households percentages (with the caveat that visual interpretation is based on two different sets of divisions of percentage values). These differences are not surprising given that the later variable is a function of the structure of the housing sector, and this explains the high visibility of the London area in Figure 1b. While differences may be expected between, say, housing tenure and LLTI or unemployment, the nature and scale of these differences can only be explored through the use of some quantitative summary measure. The main focus here is on the use of Moran's I using the weighting scheme defined in Equation 3.

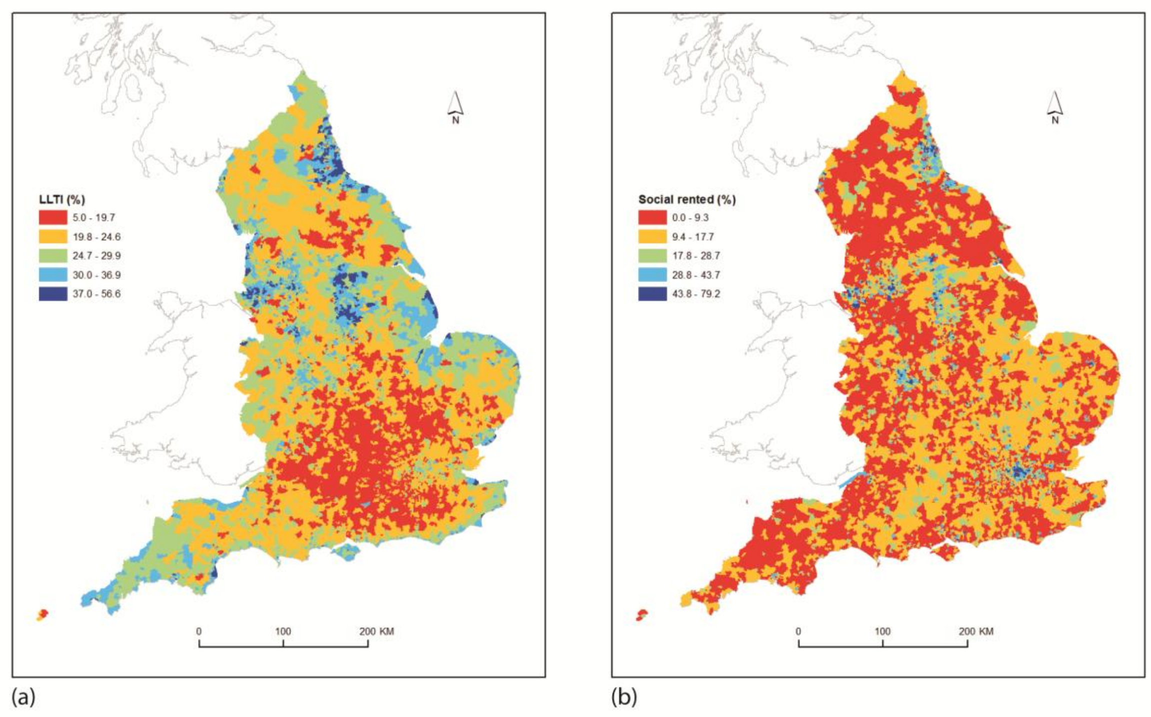

Figure 1. (a) LLTI (%), (b) Social rented households (%), by CAS Wards.

Source: 2001 Census, Output Area Boundaries. Crown copyright 2003

Figure 2 presents Moran's *I*, computed using raw percentages, for a range of spatial bandwidths. In Figure 3, Moran's *I* is given for log-ratios. The changes in *I* with an increase in spatial bandwidth are, predictably, different for the raw percentages and the log-ratios. In both cases, the variables relating to LLTI and ethnicity (albeit a very coarse categorisation of White/non-White) correspond to larger values of *I* at all spatial scales than any other variable (with the exception of unemployment in the log-ratio case, and no qualifications in the raw percentage case). The results show that some variables are more clustered at some spatial scales than others, but that this relationship is not consistent – for example, for log-ratios, the unemployment variable is more clustered than the no qualifications log-ratio at bandwidths up to 20km, but the order is reversed for larger bandwidths.

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, rather than a nested hierarchy of zones. Indeed, the analysis presented here suggests marked spatial clustering of LLTI, but a much smaller degree of clustering in, for example, housing tenure (social rented households specifically), findings which are contrary (in ranking terms) to the *D* values of Voas and Williamson. Comparative work on the two approaches would be useful to allow assessment of the reasons behind the differences. Of course, unevenness, the dimension of segregation represented by *D*, is not the same as clustering, and alternative measures will produce different rankings of variables.

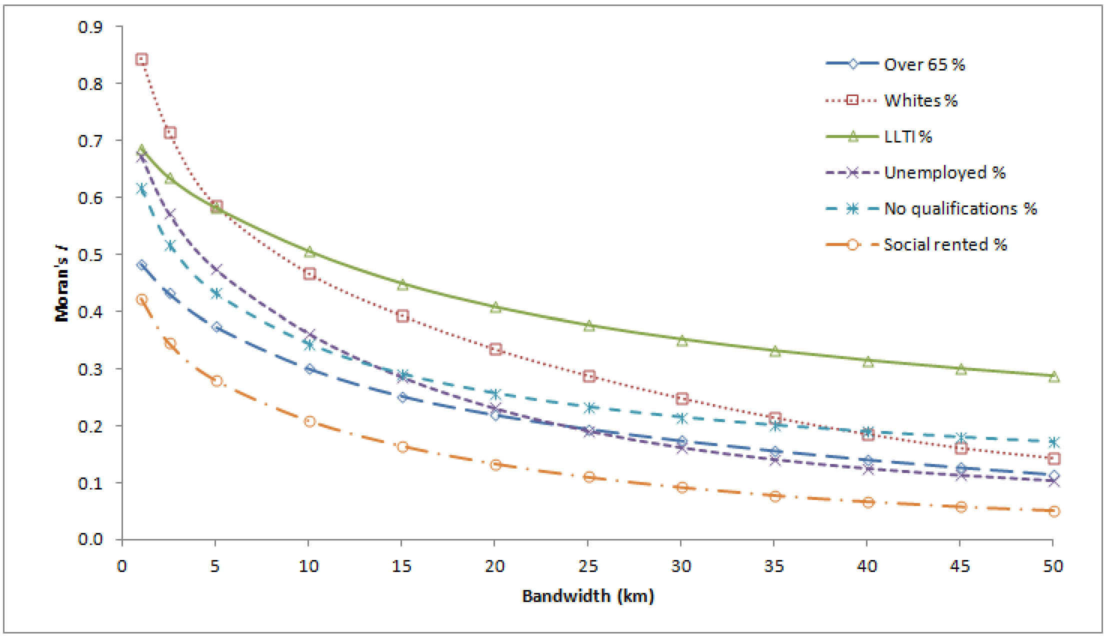

Figure 2. Moran's *I* for percentages

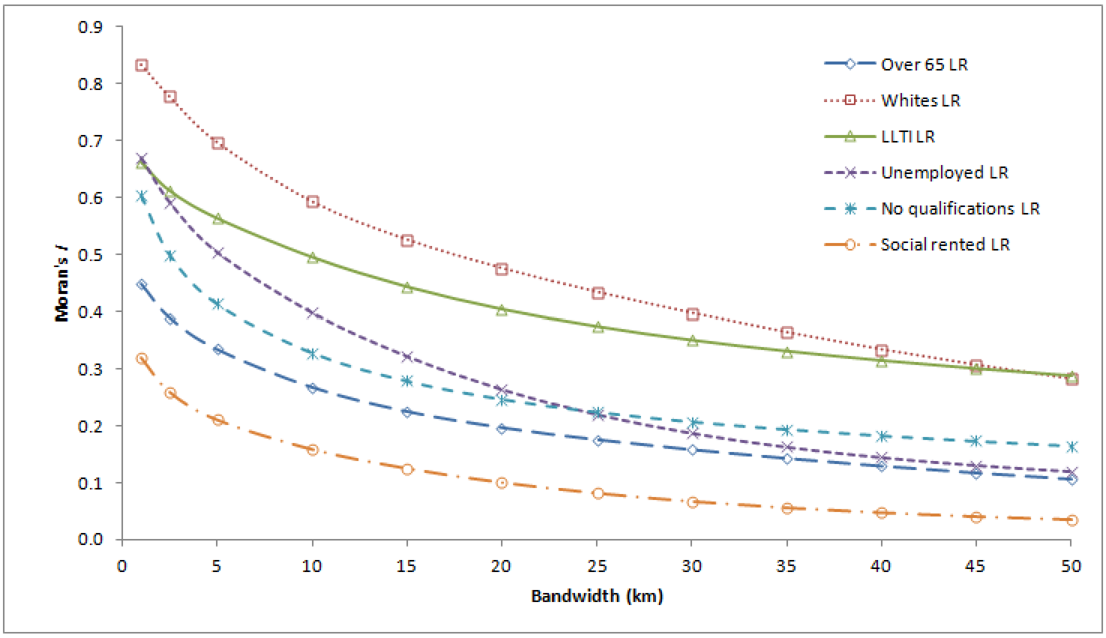

Figure 3. Moran's *I* for log-ratios (LR)

4. Discussion and conclusions

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 2011 (not available at the time of writing).

The differences in the results for raw percentages and log-ratios are partly a function of the degree of skewness of the raw percentages, which may be partly adjusted by the log-ratio transformation. However, given that 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, use of log-ratios is preferable.

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.

5. Acknowledgements

The Office for National Statistics are thanked for provision of the data. Office for National Statistics, 2001 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.

References

Fotheringham AS, Brunsdon C and Charlton M (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley and Sons, Chichester.

Lloyd CD (2010) 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(8)** 1193-1221

Lloyd CD, Pawlowsky-Glahn V and Egozcue JJ (2012) Compositional data analysis for population studies *Annals of the Association of American Geographers* **102(6)** 1251–1266.

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(4)** pp 465-481

Biography

Chris Lloyd is a Senior Lecturer in Human Geography. His research focus is on spatial data analysis, and in particular on local spatial statistics and the exploration of spatial scale. He has a particular interest in the conceptualisation and measurement of residential segregation, and is also engaged in analyses of distortion in historic maps.