ERSA 2003 Congress, 27-30 August 2003, Jyvaskyla, Finland

Spatial dependence and heterogeneity in patterns of urban deprivation

Paul Longley and Carolina Tobón

Department of Geography and Centre for Advanced Spatial Analysis (CASA)
University College London
1-19 Torrington Place, London, WC1E 7HB, UK
{plongley, ctobon}@geog.ucl.ac.uk

Abstract

Developments in the provision and quality of digital data are creating possibilities for finer resolution spatial and temporal measurement of the properties of socio-economic systems. In this paper, we suggest that the ‘lifestyles’ datasets collected by private sector organisations provide one such prospect for better inferring the structure, composition and heterogeneity of urban areas. Using a case study of Bristol, UK, we compare the patterns of spatial dependence and spatial heterogeneity observed for a small area (‘lifestyles’) income measure with those of the census indicators that are commonly used as surrogates for it. This leads to specification of spatial dependence using a spatially autoregressive model, and accommodation of local heterogeneity using geographically weighted regression (GWR). This analysis begins to extend our understanding of the determinants of hardship and poverty in urban areas: urban policy has hitherto used aggregate, outdated or proxy measures of income in a less critical manner; and techniques for measuring spatial dependence and heterogeneity have usually been applied at the regional, rather than intra urban, scales. The consequence is a limited understanding of the geography and dynamics of income variations within urban areas. The advantages and limitations of the data used here are explored in the light of the results of our statistical analysis, and we discuss our results as part of a research agenda for exploring dependence and heterogeneity in spatial distributions.

1. Introduction

The scale and pace of change in urban systems is without historic precedent. Today's increasingly affluent populations have ever more diverse lifestyles, and it is increasingly untenable to think of intra urban social patterning by analogy to an inert mosaic of internally homogeneous statistical reporting zones (Johnston, 1999). Moreover, there is mounting evidence to suggest that, in developed countries, differences in social conditions and levels of participation in society are becoming starker at fine spatial scales (Hall and Pfeiffer, 2000: 82). Small area spatial differentiation in physical and social conditions thus remains an important and developing focus for policy concern: most cities are restricted in spatial extent by planning policy, and so experience complex processes of change in neighbourhood composition over even quite short time periods (Harris, 1999). Appropriate allocation of public resources within and between urban areas requires that social conditions be represented in ways that are open to scrutiny (Gordon and Pantazis, 1997), while there is increasing realisation that 'if government has no settled and adequate measure of poverty, then it cannot reliably assess how its policies are contributing to reducing poverty' (White, 2002). Such measures need to be generalisable, transparent, pertinent, up-to-date and safe to use (Campbell, 1999). Taken together, one of the greatest challenges to policy-relevant urban geography is to keep pace with the fission of many urban lifestyles and effectively measure and monitor pertinent social conditions at fine spatial scales.

In recent years our ability to measure and monitor the morphology and extent of urban areas has improved considerably, in view of the development of urban remote sensing (Donnay *et al.*, 2001), 3-D cadastral systems (Lemmen and van Oosterom, 2002) and new digital mapping products (Morad, 2002; Longley and Mesev, 2002). A major UK research programme (NERC URGENT: Swetnam *et al.*, 2002) has demonstrated various ways in which urban regeneration initiatives can be furthered through improved measures of conditions across extensive urban tracts. There is a range of procedures that can be used to cross validate detailed maps of a range of land cover and administrative criteria. Yet there has been no commensurate improvement in our ability to measure or infer urban social conditions. The barriers to creation of pertinent, robust and defensible small area measures are well known: confidentiality strictures dictate the maximum resolution at which census variables may be differentiated (Rees and Martin, 2002); population censuses are carried out only every ten years (UK researchers will still be using 1991 Census data well into 2003); resource constraints on public sector surveys limit the sampling interval and hence increase zone size for sample estimates (e.g. Dale and Teague, 2002); non-response is an increasing problem in most national settings (see the discussion of the 'missing millions' in the 1991 Census: Champion *et al.*, 1996); and the remit of many public sector surveys is restricted to what are deemed acceptable intrusions into private lives (and in the UK this precludes any income question in population censuses). The consequence for quantitative urban geography is representation of cities as crude mosaics of rough-hewn tiles, coloured according to very imperfect surrogate attribute measures, coincident with reporting zones that may be arbitrary, and remaining inert for the duration of successive decennial census periods.

In important respects it is possible to caricature two prevalent responses to urban geography. Postmodernism entails a plurality of analytical perspectives, and many geographers have become frustrated and overwhelmed by the inadequacies of data to create plausible representations of living conditions. They have by and large abandoned the quest for generalised analysis of urban areas. At the opposite extreme, the applied, task-centred geography practised by marketers has embraced the revolution in the capture and handling of geographic information in order to create datasets attuned to the needs of specific projects. But the foundations to the resulting representations are provided by commercial datasets, which make no claims to generality or social, economic and demographic inclusiveness. The sources of bias in such data and the ways in which they operate are usually unknown. With respect to this latter approach, the collection of commercial datasets clearly does not correspond with best scientific practice (Goodchild and Longley, 1999). Yet, in important respects they do allow the creation of rich and generalisable depictions of contemporary living conditions, the data are frequently collected, and they are quickly processed. There is thus a developing realisation that empirical regularities between datasets collected across different domains can allow inference about the new polarities of social structure (Webber and Longley, 2002).

In this paper, we address some of these issues in the context of the debate about the intra-urban geography of hardship and social exclusion. Low income fundamentally restricts the abilities of people to participate actively in society (Harris and Longley, 2002), yet reliable, up-to-date income measures at fine spatial scales are rarely available from conventional sources. As a consequence, Lee (1999) reflects that many indicators of deprivation are reliant upon data sources that are out of date and/or entail use of crude surrogate measures. Some measures bear little clear correspondence with hardship at all – for example, even following re-engineering of socio-economic classifications (Rose and Pevalin, 2002), diversity in terms and conditions of employment make occupational classifications an increasingly opaque indicator of household circumstances. Other widely-used indicators are spatially variable in their operation. For example: low car ownership is a poor bellwether of socio-economic conditions in transit rich metropolitan centres; lacking freezers has different connotations in densely-populated urban areas where residents are almost invariably close to a shop (which will often stay open late); and socio-economic descriptors of occupation can have different connotations in different settings (e.g. the conditions of employment and remuneration of personal assistants, accountants and solicitors in the English provinces compared to their City of London counterparts). One response that has been better received (Gordon *et al.*, 2000) is to use a 'democratic majority' to identify the goods and services that, by common consent, define social exclusion; to then infer the relationship between these goods and services and census variables in a nationally representative sample survey; and to then infer the detailed geography of hardship through the geography of the census variables. This is essentially a return to the 'budget standards' approaches pioneered in Victorian Britain (White, 2002) and represents progress of sorts (Senior *et al.*, 2000), although it also entails problematic ecological inferences and insensitivity to spatial setting. Specifically, the indicators deemed to define deprivation are not allowed to vary spatially, even though the goal of subsequent linkage to the UK Census through surrogate indicators is identification of spatial variation. Thus the specification of deprivation is aspatial (in what settings, for example, is car ownership or freezer ownership a necessity?), and it is only when parameter estimates are used predictively that spatial variation is assumed to occur. Can a model of deprivation that is mis-specified in this way provide a reliable indication of geographic variation in the phenomenon of interest? And can 'top down' inferences from a national survey to local scales be used reliably to infer characteristics of small areas?

The broader issue concerns the scale and extent of 'pockets' of hardship and the scale ranges at which difference is deemed manifests. Where geography underpins policy, there is often a prevailing sense of muddled thinking – for example, a recently introduced UK policy on property sales tax breaks has led to 40% of the entire area of Wales receiving relief, while the same policy can blight small inner city designations. The problems are further compounded if each of the range of surrogate measures used to specify a concept operates at different scales. Taken together, it remains unclear whether meaningful indicators of social conditions can ever be adequately specified, or whether generalised representations can be sufficiently sensitive to place. For example, at what scale range does the impact of predictor variables hold constant, and what are the possible distorting effects on the scale of hardship?

Our own view at this point is pragmatic. Notwithstanding evidence of geographies to the informal economy (e.g. Williams and Windebank, 2001) and mismatches between the spatial distribution of earned income and unearned income and wealth (e.g. Ball, 1994), it is clear that deprivation and hardship are inextricably linked to incomes from earnings and transfer payments. In many countries (e.g. the UK) no small area income measures are collected at all, and this forces reliance upon commercial sources. Yet, as we discuss below, the use of such data in academic research is not without considerable problems. In the same spirit as Gordon and Pantazis (1995) we thus think it necessary to retain some linkage to population census data – but in a way which is much more sensitive to spatial context. A critical issue is thus to understand the scales at which both income, and the variables that are used to predict it, vary (see also Rees, 1998; Harris and Longley, 2002). In our empirical analysis we examine a number of the facets to this issue through our detailed empirical case study. We begin by addressing the issue of multicollinearity in the specification of a model of small area incomes. Next we examine the over-all level of spatial autocorrelation in small area incomes and the predictor variables used to represent it. Further analysis is used to show that the local patterns of spatial autocorrelation vary between the independent variables. This is consistent with small area heterogeneity in each of the independent variables and we use geographically weighted regression (GWR) to appraise local variation in the parameter estimates of the indicator variables. We then draw conclusions with respect to the specification, estimation and testing of deprivation measures, and whether such methods might be developed using commercial data sources. The broader issue, which may be relevant to the proposed replacement of censuses with sample surveys, concerns whether we are better off assessing social conditions through frequent sample surveys, or whether it is necessary to aim for 100% census coverage (with consequent decreases in temporal granularity) to develop reliable representations of conditions.

Our empirical case study begins to assess the contribution of new data sources to building improved representations of income distributions. We take a lifestyles survey that enables us to create small area estimates of household incomes for a study of the City of Bristol, UK, defined as the census zones lying within a radial distance of approximately 9 km of the historic city centre. We use this analysis to begin to identify the relative priorities of better data, better specification of spatial dependence to draw inferences across space, and better understanding of local heterogeneity. The case study is also used to evaluate, in an inductive way, the scale at which it seems meaningful to draw inferences about hardship in our study area.

2. Data Considerations

'Lifestyles' datasets provide detailed socio-economic information about individuals' circumstances and preferences. The term 'lifestyles' is used as an umbrella term for a diverse range of survey data, which may be concatenated from disparate data sources using a common field (the residential address) to form a profile of individual households. They are mostly obtained from responses to postal consumer surveys or questionnaires and are widely used for direct marketing. Although they are collected by private sector organisations, they can be made available (in suitably anonymised form) as individual or household records.

The limitations of lifestyles datasets are well known in general, if not always well-quantified, terms (but see Longley and Harris, 1999). *Inter alia*, they almost invariably employ closed response questions, completion is voluntary (Harris, 1999), and some low income groups are likely to be under-enumerated. These limitations make lifestyles data less reliable than conventional census data, for example, and the operation of biases in their collection remains an under-researched area. Nevertheless, they are updated every year, response rates enable reasonably robust estimates to be made at small area levels, and they contain a direct question on income – which has never been asked as part of any UK Census of Population.

The limitations of lifestyles data for academic research and policy purposes can be understood in relation to the quite distinct rationale for their creation (Webber and Longley, 2002). First, most commercial organisations are interested in the estimation of a metric which is directly indicative of the level of household expenditure, rather than income. Their interest in income is thus as a proxy for expenditure. It is also worth distinguishing between two rather different applications. Direct marketers, who communicate with individuals rather than serve areas, desire estimates of expenditure which are optimally predictive at the person or household level but are not materially concerned whether there is any geographically systematic error in this estimate. Thus there are no inferential errors generated in 'one to one' marketing applications. By contrast retailers, who serve areas rather than individuals, are not particularly concerned whether their income estimation methods are accurate at the household level. Rather, like the academic or policy analyst, it is more important that whatever inferential errors there may be are not systematic at the area level. Thus the best model for direct marketers is not necessarily the best model for retailers.

There are additional inferential errors that arise in using 'non-traditional' data sources in spatial analysis. To the collector of lifestyle data, there are significantly different commercial returns from having a high income household respond than a low income household. Companies that use lifestyle lists to identify potential mail-drop targets are unlikely to select households with low incomes. For this reason most lifestyle operators have decided that it is more cost effective to target the blanket door drops of questionnaires to postcode sectors with higher rather than lower levels of affluence (Webber, personal communication). However, UK postcode sectors are the lowest areal units which can be leafleted by distributors, and this is likely to improve the representativeness of response – especially where, as in Bristol, high and low income groups continue to live in quite close proximity to one another. Thus while timely, small area income data provide a central ingredient of small area hardship measures, there is a clear need to cross validate small area income estimates with respect to external data sources. Some aggregate comparisons are reported by Longley and Harris (1999), but this does not accommodate small area variability in income measures.

The lifestyles survey used here comprise 51 882 household responses to a postal survey, of which 43 278 (83%) include income information. Each household record is georeferenced to the unit postcode level (equivalent to the full US Zip code), which enables matching of the records to individual census enumeration districts (blocks). The income question required the respondents to identify their approximate household income from seven bands, which were rescaled to take proportionate values within the 0-10 interval[i] (Harris, 1999). Income scores were aggregated by enumeration district (ED: the UK equivalent of a US census block) and divided by the total number of respondents in each of the 844 EDs in the study area, in order to obtain an income score (INC) for each ED used in the calculations. We focus here upon the relationship between the small area income measure and a number of conventional 1991 Census variables. A combination of *a priori* reasoning and statistical analysis led us to posit a relationship between income scores and the following indicator variables found in the 1991 Census:

- OLDPSN is the percentage of households per ED where residents are aged 65-74.

- 2534NK is the percentage of households per ED where residents are aged 25-34 and there are no children aged 0-15 years.
- WKWIFE is the percentage of households per ED where there are married females working.
- BIGACC is as the percentage of total households per ED with 7 or more rooms.
- COUNCL is the percentage of households per ED in council (public sector) rented property.
- HHTCA is the percentage of households per ED owning two cars.
- UNSKLD is the percentage of households per ED with unskilled workers.
- QUALML is the percentage of qualified male residents per total households in the ED.

Most of the variables are positively skewed, in part because of the incidence of low or zero values near to the historic centre of the city. The spatial distribution of these nine variables is shown in Figure 1, where maps are shaded according to quintile ranges. Some interesting spatial features of the data are made apparent by these maps. For instance, EDs with a large proportion of young childless residents (2534NK, top centre map in Figure 1) are concentrated in areas surrounding the historic centre of Bristol and generally coincide with areas of low proportion of households in council accommodation (COUNCL, central map in Figure 1). In contrast, the older population (OLDPSN, top left map) tends to locate towards the outskirts and particularly to the north of Bristol in areas that also appear to have a high percentage of the households living in council (public housing) accommodation. The spatial distribution of qualified male residents (QUALML, bottom centre in Figure 1) indicates that the 574 EDs that have the lowest proportion for this variable (less than 25% of the households in the ED) are largely coincident with areas where households live in spacious accommodation (BIGACC, middle left map in Figure 1). Unsurprisingly, most of the EDs with higher scores for the income score variable (INC, bottom right map) are in that same sector. EDs with a high proportion (30% or more) of multiple car ownership (HHTCAR, middle right in Figure 1) are also located in this area, as well as in EDs that are further away from the city centre. These observations bear out the common experience that areas with poor physical and social conditions broadly correspond with one another. However, the relationship between income and various deprivation indicators is not invariably straightforward.

[Figure 1 here]

3. A city wide relationship between income and the indicator variables

Linear regression analysis is the standard technique for formalising statistical associations between a dependent variable and a set of explanatory ones, and estimating the best fit between the predicted and observed values of the dependent variable. The standard regression equation is given by:

$$y_{Nx1} = X_{NxK}\beta_{Kx1} + \varepsilon_{Nx1} \quad \textbf{(1)}$$

where y is the dependent variable in vector form of length N ; X is a matrix of N observations and K explanatory variables; β a vector comprising K regression coefficients; and ε is a vector comprising random errors for each observation N . Ordinary least squares (OLS) provides the best linear unbiased estimators β , assuming that the random error term is uncorrelated and normally distributed with zero mean and constant variance. These assumptions allow statistical inferences to be drawn about the β estimates.

Table 1 shows the results of OLS regression[ii] using the lifestyles income score variable (INC): it exhibits a statistically significant positive relationship with six variables in the dataset (OLDPSN, 2534NK, WKWIFE, BIGACC, HHTCAR and QUALML) and a statistically significant negative relation with two more attributes (COUNCL and UNSKLD). Hence, the higher the proportion of households in the ED with older people or young childless residents, working wives, large accommodation, qualifications and at least two cars, *ceteris paribus*, the higher the predicted income score. Similarly, the higher the percentage of households in the ED living in council accommodation or having unskilled residents, *ceteris paribus*, the lower the predicted income score. The over-all regression specification generates a significant F statistic and the adjusted R^2 of 0.65 suggests a reasonable global fit.

[Table 1 here]

However, as in most geographical analysis, the assumption of uncorrelated normal errors with constant error variance is not realistic, given the evidence of spatial dependence in the data. Moreover, a problem commonly encountered when working with census data is the presence of multicollinearity or high correlation between observations of the explanatory variables – because many of the census variables are in effect measuring the same, or similar, constructs. Although there are no tests for directly quantifying the degree of correlation between variables in a regression, the Condition Number (Anselin, 1992) can be used as a diagnostic: our regression yields a value of 21.74 which, although high, does not have the effect of increasing the parameter variance estimates to the point that the t statistics are unable to attain significance.

However, a significant Jarque-Bera test statistic (value 11 131) indicates that the null hypothesis of normally distributed errors cannot be accepted. In such cases, tests on heteroskedasticity and spatial dependence must be interpreted with caution as they assume that this is the case. The Koenker-Bassett and White tests indicate the presence of heteroskedasticity in the residuals, which subsequent experiments with a heteroskedastic error specification were unable to rectify. However, tests to identify non-constant error variance are sensitive to the presence of spatial dependence (Anselin, 1990). A possible implication is thus that the heteroskedasticity tests indicate the presence of spatial autocorrelation in the residuals or in the dependent variable. Results of an estimator robust to heteroskedastic errors (see Anselin, 1992 for a discussion) are presented in Table 2. Inference in this case is based on the z-value[iii] and all explanatory variables can be seen to be significant, except for COUNCL and (marginally) WKWIFE. These various results provide circumstantial evidence that the model may be mis-specified in some way, yet it is not clear whether any mis-specification is attributable to spatial autocorrelation in the dependent variable and/or the residuals.

[Table 2 here]

These estimates consider each individual area independently of the values of its neighbours. As such, it is difficult to develop a picture of the scale at which hardship, or any of its indicators should be conceived and measured. The results from this regression technique can only be interpreted as yielding average parameter values for the study area as a whole (Fotheringham *et al.*, 2001; Harris, 1999), that is as presenting a *global* characterisation of the prevailing relationship. They cannot account for spatial variation or *local* differences in the data – and provide no indication of whether each indicator variable is specified at an appropriate scale. The statistical tests in our own analysis suggest a need to investigate possible spatial dependence in the indicator variables, above and beyond the need to investigate the income variable from the lifestyles data.

4. Measuring spatial association

4.1. Global Indicators of Spatial Association: Moran's I

Moran's I measures the spatial dependence or autocorrelation between values of a variable. It is structured as a measure of covariance or it may be used to test hypotheses concerning the similarity of specified values of a variable (Getis and Ord, 1996). Formally:

$$I = \left(\frac{N}{\sum_i \sum_j w_{ij}} \right) \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2} \qquad \textbf{(2)}$$

where i and j refer to the spatial units of which there are N ; $z_i = x_i - \bar{x}$, where $\bar{x}$ is the mean of x or the attribute being measured; and w_{ij} is the degree of connection or potential spatial interaction between zones i and j .

In an area of N spatial units, there are $N(N-1)$ possible interactions between each unit and all the rest. In our case, the study area comprised 844 EDs which would potentially represent 844*843 = 711,492 interactions, if unconstrained in any way by contiguity or adjacency. Hence, there is a need to impose structure on the nature and extent of the possible spatial interactions. This is normally achieved by defining a neighbourhood for each unit by means of a spatial weights matrix. These weights "are usually determined either by continuous inverse distance measurements or by binary definitions of whether or not the two zones are contiguous" (Fotheringham *et al.*, 2000: 202). Here we will assign weights defined according to the contiguity criterion. The neighbourhood matrices used are of various orders where the order is an indication of the region considered to be neighbouring each ED. [iv]

The expected value of Moran's I indicates the value that would be obtained if there were no spatial autocorrelation in the data and it is defined as:

$$E[I] = -1/(N-1) \qquad \textbf{(3)}$$

or –0.001186 in our case of 844 ED observations. Values greater than $E[I]$ indicate positive spatial autocorrelation or similar values—either high or low—clustered together. A Moran's I coefficient smaller than its expected value would indicate negative autocorrelation or dispersion of similar values (Fotheringham *et al.*, 2000; Longley *et al.*, 2001: 100-3). For inferential purposes, a standardized value of I is used such that:

$$z = (I - E[I])/SD[I] \tag{4}$$

where z is the standardised value of I ; $E[I]$ and $SD[I]$ are the theoretical mean and standard deviation of I , respectively.[v] Table 3 shows the results for the standardised global Moran's I for nine variables.[vi] All values are highly significant and greater than the mean, indicating positive autocorrelation: that is, similar values (high or low) are more spatially clustered than would be expected by chance (Anselin, 1992). For spatial weights of higher order, Moran's *I* shows diminishing values. Note for instance the values of the statistic for the INC variable that indicates decreasing evidence of spatial autocorrelation between distant neighbours.[vii] All lags up to and including lag 5 are significant, which might be taken to suggest some very coarse patterns of spatial autocorrelation. However, the global Moran's *I* statistic only provides an average measure of spatial dependence across the entire study region. Hence, the evidence of spatial autocorrelation in higher order neighbourhoods may in fact subsume still stronger, but more local variation in the data. Such patterning would certainly be consistent with the substantive setting described in the introduction, in which zones of affluence and deprivation juxtapose within small areas.

[Table 3 here]

4.2 Local Indicators of Spatial Association: Local Moran's I

Local Indicators of Spatial Association (LISA) measure the degree of spatial dependence between locations. They identify the association between a single value of a variable at one location and its neighbours, which are defined according to their degree of contiguity, as in the previous sub-section. However, unlike the global statistics presented above, LISA account for non-stationarity across space. Of course, the definition of the neighbourhood remains shackled to the zonal system and is subject to the modifiable area unit problem (Openshaw, 1984). If the shapes and configuration of the Eds (census blocks) were changed, the individuals considered to be in adjacent zones of the various topological orders would also change, as would the values of the local statistics. However, there is reason to anticipate that ED zonations are not entirely random, i.e. they bear a correspondence to social structure and built form, and so it is likely that the ED zonation and contiguity rule will create robust results.

LISA are well suited for identifying the existence of *hot spots* or local spatial clusters, assessing assumptions of spatial stationarity and identifying spatial lags beyond which no discernible association can be obtained (Getis and Ord, 1996). For Anselin (1995) LISA have two characteristics: the LISA for each observation provides an indication of the extent of significant spatial clustering of similar values around an observation; and the sum of the LISA for all observations is proportional to a global indicator of spatial association. The first of these characteristics make them a useful inductive device for ascertaining the scale of 'pockets' or 'neighbourhoods' of hardship. Local Moran's *I* is one such statistic (Anselin, 1996) and is formally expressed as:

$$I_i = \left(z_i \Big/ \sum_i z_i^2 \right) \sum_j w_{ij} z_j \qquad \textbf{(5)}$$

where the subscript i refers to the location for which the statistic is estimated and j to any other zone; $z_i = x_i - \bar{x}$, where $\bar{x}$ is the mean of x or the attribute being measured; and w_{ij} is the degree of connection or potential spatial interaction between zones i and j. The expected value of the local Moran's *I* is defined as:

$$E[I_i] = -w_i/(N-1) \qquad \textbf{(6)}$$

The interpretation of I_i as an indication of local stability can be better understood from its relation to the global statistic as the average of all I_i values is equal to the global Moran's *I* scaled to a factor of proportionality (Anselin, 1992). Since I_i varies by location i, it is more easily interpreted visually by colour coding of each enumeration district. Figure 2 presents the normalised I_i values for the first order spatial weights matrix of each of the nine variables. Hardship and deprivation are usually defined as occurring in areas where poor physical and social conditions interact, yet there is only limited uniformity in the patterns revealed by the LISAs. Note for instance the presence of 'patches' of positive spatial dependence for BIGACC (middle left map in Figure 2), HHTCAR (middle right map), QUALML (bottom centre map) and INC (bottom right map) to the northwest of Bristol and an indication of negative spatial autocorrelation in COUNCL (middle centre map) for the same region. Other hotspots of positive spatial autocorrelation are apparent to the west for HHTCAR (middle right map), BIGACC (middle left map) and INC (bottom right map). Areas with similar values of 2534NK (top centre map) are concentrated around and to the northwest of the city centre where BIGACC and COUNCL also show less positive or even negative spatial autocorrelation patterns. Similarly, QUALML shows distinctive hotspots to the south of Bristol.

[Figure 2 here]

These results illustrate how scale effects differ amongst the constituents of the global regression. Thus, for example, the high local patterns of negative spatial autocorrelation in the COUNCL variable reflect the clustering of many housing units in public sector estates during construction, while the more gradual variation in the characteristics of currently resident households depicts the outcome of (typically) half a century of filtering of tenant characteristics and the effects of tenant 'right to buy' initiatives since the late 1970s. The values of the 2534NK and QUALML variables exhibit intermediate patterns.

5. Specifying Local Relations

These observed scale effects suggest a need to accommodate geographical variability in the regression specification. There is a long tradition in spatial analysis that has attempted to understand and specify local relations in multivariate data. For instance, Casetti's (1972) *expansion method* attempts to measure trends in relations over space by making the parameters of a global model a function of some other attribute such as location. In some circumstances, however, this technique can obscure significant local variation (Fotheringham *et al*., 2002: 16-17). Since the results in the previous sections give an indication of possible significant spatial variation in the data, the method was not considered for its investigation. The *spatial adaptive filtering* method proposes a model to investigate local and regional effects by estimating spatially varying parameters. However, if there are large differences between the true parameters of neighbouring observations, "the algorithm will tend to smooth out these changes rather than recognize sudden drops or increases" (Olligschlaeger, 1997). Hence, it tends to produce regression parameters that 'drift' slowly across geographical space (Fotheringham and Brunsdon, 1999). In the *random coefficient model*, parameters are allowed to vary randomly, as opposed to the classical linear regression model where they are assumed to be the same for all locations and cases, or the spatial expansion method where they are smooth. This technique however is not inherently geographical as coefficients can be drawn from very different distributions, even if the corresponding cases are in close proximity.[viii]

Two methods are used here to explore the local variability of the relations in the dataset, namely spatial autoregressive models and geographically weighted regression (GWR). The former recognises that spatial data may not be independent, and that this can have a number of effects on the optimality and/or efficiency of traditional linear regression estimates. If present, the technique attempts to model the spatial dependency in the dependent variable, in the error terms, or in both. Although spatially autoregressive models provide a good means of detecting local relations, they can only be accommodated in a set of global parameter estimates. GWR allows for local rather than global parameters to be estimated, and thus provides a better way of accommodating spatial heterogeneity in the local geography of high and low income in Bristol. Both techniques can be interpreted as localised versions of traditional global techniques.

5.1. Spatially Autoregressive Models

LISAs provide a useful means of identifying the extent of autocorrelated zones, because the statistic is sensitive to the definition of the neighbourhood that is used. They also provide a clearer indication of the intensity of the relation than is provided by a global autocorrelation statistic alone. This can be used as a starting point for direct specification of spatial dependence by allowing values of the dependent variable and/or the error term at a location to be correlated with observations at other locations. If the dependent variable is spatially autocorrelated with the values of neighbouring locations, then its spatial dependence can be formally modelled as a spatial autoregressive model:

$$y_{Nx1} = X_{NxK} \beta_{Kx1} + \rho (Wy)_{Nx1} + \varepsilon_{Nx1} \tag{7}$$

where y, X and β are as in Equation (1); the subscript N denotes the number of observations and K the number of explanatory variables; Wy is a vector of spatial lags for the dependent variable and ρ the corresponding regression coefficient, and ε is a vector of random and independently distributed errors.

If spatial autocorrelation is shown to be present, OLS estimates would be biased and inferences drawn using the aspatial regression model (Equation (1)) would be incorrect. Moreover, if there is spatial dependence in the error term, OLS estimators are not efficient and, although parameter estimates remain unbiased, inferences based on t and F significance tests are misleading, as is interpretation of the R^2 goodness-of-fit measure (Anselin, 1992).

In our case study, there was no evidence of spatial dependence in the error term although the Lagrange Multiplier (LM) diagnostics used to identify it are only valid under the normality assumption. Therefore, this evidence must be treated with care as the Jarque-Bera test showed the residuals not to be normally distributed. However, robust LM diagnostics could not identify this form of dependency which was therefore not considered to be present in our case study.

An instrumental variables approach to estimation was taken, because of the likely non-normal error distribution. Instrumental variable estimation is based on the principle that a set of instruments exists which is correlated with the original explaining variables but which is uncorrelated with the error term (Anselin, 1992). Instruments that fulfil this requirement will result in a consistent estimate of the parameters. The instruments are used to construct a proxy, in this case, of the spatial lag or the term Wy in Equation (7). In our case study, the first order lags of the independent variables are included as the instruments (see Anselin (1988) for a discussion) in order to accommodate the presence of spatial dependence. This occurs because if the error values were non-autocorrelated, as the LM diagnostics suggest, then lags of the explanatory variables would be uncorrelated with the error term and "at the same time probably highly correlated with the explanatory variable" (Judge *et al.*, 1980: 534). The instruments are then regressed against the independent variables in a standard OLS regression making it equivalent to a two-stage least squares method Table 4 shows the results.

[Table 4 here]

Apart from not requiring the assumption of normally distributed errors, instrumental variables estimation has been used widely in the econometric literature to account for the possible endogeneity between income and the socio-economic independent variables where the direction of causality in the relations between them is not straightforward (see for instance Mincer, 1974). For instance, it may be the case that being unskilled causes a lower income but the opposite relation may also hold.

COUNCL as well as the CONSTANT were excluded from this regression as they were found not to be statistically significant in the instrumental variables estimation. The ρ parameter associated with the spatial lag of the dependent variable, or the instruments in this case, is highly significant as well as the seven other explanatory variables included. All parameters, including the spatial lag or instrumental variable, are significant with the expected signs. The pseudo-R^2 of 0.7 reported in Table 4 gives an indication, though not comparable to the OLS results, of the goodness of fit of the model. Figure 3 shows the residuals for the estimation, which do not seem to exhibit any particular pattern.

[Figure 3 here]

Instrumental variables estimates can be interpreted as weighted averages of individual-specific causal effects. Although this model takes into account the relation between locations and allows us to account for the dependency in the data, they are not intended to estimate how that dependency may vary across space. This is explored using geographically weighted regression (GWR) in the following section.

5.2. Geographically Weighted Regression (GWR)

GWR is a technique that "extends the traditional regression framework by allowing local rather than global parameters to be estimated" (Fotheringham *et al.*, 2001: 51). As such, in the present case study, it provides an important method of identifying spatial heterogeneity in each of the predictors of income. Following Brusdon *et al.* (1996), GWR can formally be expressed as:

$$y_i = \beta_{0i} + \sum_K \beta_{Ki} x_{Ki} + \varepsilon_i \qquad \textbf{(8)}$$

where y_i is the observation of the dependent variable at location i and β_{Ki} is the value of the parameter for the corresponding explaining variable at point i. Hence, in this model, a continuous surface of parameter values is estimated under the assumption that locations nearer to i will have more influence on the estimation of the parameter $\overset{\wedge}{\beta_i}$ for that location (Fotheringham *et al.*, 2000). This is formally expressed in matrix form as:

$$\overset{\wedge}{\beta_i} = \left(X^t W_i X\right)^{-1} X^t W_i y \qquad \textbf{(9)}$$

where the n by n weights matrix W_i has diagonal elements $\left(w_{i1}, w_{i2}, \ldots, w_{in}\right)$ that denote the weighting of observed data on the calibration of the model around point i and off-diagonal elements equal to 0. The weights are defined as continuous functions of distance. Hence, they vary with i and are greater the closer an observed data point is to the calibration point (Fotheringham *et al.*, 2001; 2002). A function of the form

$$w_{ij} = \begin{cases} \left(1 - d_{ij}^{\ 2} / h^2\right)^2 & when \quad d_{ij} \leq h \\ 0 & when \quad d_{ij} > h \end{cases} \qquad \textbf{(10)}$$

was used for all estimations reported here where d_{ij} is the distance between points i and j and h denotes the bandwidth. In principle, GWR may be applied at any geographic scale of measurement, although in practice the availability of suitably anonymised data has restricted many applications to the sub-regional or coarser scales (e.g. Fotheringham *et al.*, 1996; Fotheringham and Brunsdon, 1999). The bandwidth, *h*, is increased where data points are more widely spaced, a desirable property in our case study where the data points are ED centroids, which vary greatly in their density across the City of Bristol. Note however that the selection of the continuous function does not appear to have an effect on the results (see Fotheringham *et al.*, 1998) whereas the selection of the bandwidth is crucial: a large bandwidth would produce parameters with little spatial variation or greater smoothing and a small bandwidth would produce parameter estimates with increased variance or large local variation. Bandwidth selection can follow a number of criteria. In this paper, the bandwidth that minimised an AIC criterion was used (see Charlton *et al.* 2002, for details).[ix]

The OLS estimation in Equation (1) indicated that, on average for the whole region, high levels of the income score variable are positively related to high levels of OLDPSN, 2534NK, WKWIFE, BIGACC, HHTCAR and QUALML, and negatively with high levels of COUNCL and UNSKLD. GWR extends this aspatial regression framework by allowing the estimation of local parameters instead of global ones. These can be more easily inspected visually when represented as maps like those in Figure 4 which illustrate the variation of the parameters by location for the eight variables in the regression. These maps also show how the relation between the explanatory and the dependent variable can differ by location. Interesting relations include the areas of negative parameters of WKWIFE (top right map in Figure 4). Although the ‘global’ relation between this variable and INC for the whole region is positive, this relation clearly varies across space with the stronger relation being in the southeast (darker areas). A similar situation occurs with the coefficients for all other variables which have the expected signs but their spatial patterns indicate that the relation of the variables to INC is less positive in the lighter shaded or white areas. The intercept terms also show a clear spatial pattern with lower positive values to the southeast and higher values located at the north western edges of Bristol. Taken together, these maps suggest considerable heterogeneity in the contribution of the predictor variables to accounting for the geography of income variation. While they some of the mapped variables show clear sectoral patterns, the differences between the maps suggest the geography of surrogate indicator measures may compromise their use in policy applications.

[Figure 4 here]

The values represented in the maps in Figure 4 do not take into account the standard error of the parameter estimates (Fotheringham *et al.*, 2001: 53). Hence, dividing the local estimates by their corresponding standard errors permits the mapping of a pseudo t-statistics (Figure 5) that can indicate the significance of the parameters described above (darker greys or black indicate the highest significance). In this case, areas of larger parameter values (either positive or negative) in Figure 4 generally coincide with corresponding large t-values in Figure 5. This observation does not always hold as in the case of COUNCL (middle centre map). Thus the magnitude as well as the significance of the relationships between the explanatory variables and the income score are not spatially invariant. For instance, QUALML (bottom centre map in Figure 4) shows a more negative relationship to the west and northwest of the study area where the pseudo t-statistics are also higher, while the opposite relation occurs to the east. HHTCAR (middle right maps) shows a strong positive relation with INC around the centre and along a strip running from the north west to the south east of Bristol. 2534NK (top centre maps) shows a strong positive relation to INC across a similar strip as that just described. Conversely, the negative dependency relation of INC to the COUNCL (middle centre) appears to be stronger and more significant in areas further from the historic city centre. The parameters of BIGACC (middle left) are of highest magnitude, indicating a stronger statistical relationship to INC, to the northeast, but the t-values are high throughout most of the study area, with the exception of a patch to the north west of the centre.

[Figure 5 here]

Figure 6 shows the local values of R^2, a goodness-of-fit measure that can “informally depict the accuracy with which the model replicates the observed values [of the income score variable] in the vicinity of the point for which the model is calibrated” (Fotheringham *et al.*, 2000: 125). The map indicates that there is some variation in the R^2 statistic: however, the statistic ranges from moderate levels (c. 0.67) to high values (up to 0.92), with the highest values occurring to the north of the study area. The map of standardised residuals to the right of Figure 6 illustrates that they have no particular spatial pattern.

[Figure 6 here]

Finally, GWR makes it possible to inspect the spatial variability of each parameter by calculating a Monte Carlo test of significance (Fotheringham *et al.*, 2000). The results presented in Table 8 indicate that only UNSKLD and marginally QUALML appear to vary significantly across space. The spatial variation in the remaining variables cannot be said to be significant. This would indicate that the distribution of unskilled workers and qualified males shows spatial non-stationarity.

6. Consolidation and assessment

Better measurement usually precedes the development of better theory in both science and social science. In the context of urban geography, ‘better’ measures need to be more timely, more relevant and highly disaggregate if generalisation is not to be blind to pattern, reductionist in classification and oblivious to rapid change. The ongoing revolution in the provision and handling of socio-economic information is improving the supply of geographic data, yet a challenge to researchers is to better understand the sources and operation of biases in data collection. In this paper we have used analysis techniques that are sensitive to context to begin to assess the inherent differences and spatial mismatches between conventional indicators of income from the UK Census and direct income measures from a lifestyles survey. We believe that extending the interests of urban geographers from census analysis towards work with direct, timely, spatially disaggregate indicators is key to developing the data foundations to a new, data rich and relevant urban geography. The issues of using lifestyles data are fraught with problems, not least because some users of such surveys have little interest in the vagaries of spatial heterogeneity and non-stationarity. However, just as retailers use such data to supplement conventional geodemographic indicators, there is a need to tease out the relationship between these new pertinent measures and conventional indicators.

The problems of inference in urban geography are not just of technique, but of generalised, timely, spatially disaggregate data. This paper has begun to address the advantages of using new sources of digital data for estimating models that allow statistical inference and a better understanding of their processes in space. The range of spatial models permitted the identification of spatial patterns of interest beyond averaged relations between dependent and explanatory variables. It may appear as intuitive to argue that variables can better explain the variability in the level of income in some EDs rather than others, or in statistical terms, that the significance of their relation to the dependant variable can vary across space. Yet this confirmation of local heterogeneity and spatial dependence can also be interpreted as providing evidence of local misspecification of multifaceted concepts such as hardship. Use of a range of techniques points both towards interesting local spatial patterns and at problems in the specification of the models, both of which encourage iterative refinement and better understanding of local heterogeneity. This also underscores the importance of including spatial effects when they exist as a means to account for local variability, and the complexity of specifying its manifestations at finer scales. There is a sense here of the receding horizon – whilst the economist might adhere to the Holy Grail quest for a fully specified model, we incline to the geographer's view that we will never completely specify the attributes of place. Our case study illustrates the ways in which surrogate measures of income (such as car ownership and ownership of consumer durables) may be affected by the existing physical and socio-economic infrastructure.

Inevitably analysis such as this raises as many questions as it begins to answer, but a task of urban geography should be to contain such questions to clearly defined issues, as a precursor to making their treatment more routine. Accumulated experience should be able to tell us about the effects of different contiguity/proximity definitions, the weighting of contiguity relations according to the over-all similarities of adjacent zones, the vagaries of data linkage mechanisms, differential weighting of sample surveys and the likely biases in data from different sources. In a similar manner, it should be possible to inform the specifically geographical aspects of analysis with the results of accumulated experience. In the context of UK studies of hardship and deprivation, these might include: regional differences (e.g. the result of average rooms per dwellings in Scotland being consistently low the national average irrespective of demographics); factors pertaining to accessibility (e.g. the importance of proximity to town centres and to public transport networks when representing the effects of car ownership); and ecological factors (e.g. studies of voting behaviour suggest that variables from a coarser levels of granularity have incremental impacts because of spillover effects: Webber and Farr 2001). It may be that different methods would be appropriate in these different circumstances. Taken together, continuing research needs to consider the extent to which these techniques are necessary because of the unreliability of the data that is being predicted or because or localised distortions in the independent variables.

In a wider sense, there is an important role for geographers in relation not just to policy but the broader remit of business and service planning. Just as the importance of car and freezer ownership is not spatially invariant, so too the range of lifestyles variables now used by retailers have different connotations in different settings. In the UK, for example, it has been suggested that rural areas have higher ownership of houses, cars and household equipment than do inner city areas, reflecting the lower level of competition for land in rural areas (Webber, personal communication). By contrast, inner city residents may spend a much higher proportion of their disposable income on experiences (cinema, theatre, eating out, going on holiday, going to the pub), in part because of the greater range of commercial opportunities for spending money on leisure experiences that are available to them. This is an important and developing area of research and, like the creation of policy-specific hardship measures, resolution of uncertainty is as much concerned with issues of application as with issues of science (Zhang and Goodchild, 2002). Despite these limitations, new sources of data combined with appropriate spatial analysis techniques can reveal information and insight into spatial relations that may be little understood. Hopefully this exercise has shown the potential richness and depth of analysis that can be attained.

Notes

[1.]The household income bands and their values as rescaled by Harris (1999) are: 1) household income under £5,000 (rescaled to 1.2 to accommodate social security thresholds); 2) between £5,000 and £9,999 (rescaled to 2); 3) between £10,000 and £14,999 (rescaled to 3); 4) between £15,000 and £19,999 (rescaled to 4); 5) between £20,000 and £29,000 (rescaled to 5.5); 6) between £30,000 and £39,999 (rescaled to 7.5); 7) household income over £40,000 (rescaled to 10).

[2] Estimations in this section were performed using SpaceStat 1.8 (Anselin, 1992) which, in addition to calculation of spatial statistics, is also a convenient environment for performing aspatial multiple regression.

[3] The significance of individual coefficients is based on the standard normal distribution because the estimation method is based on asymptotic considerations. Hence the z-value column of Table 2, which is actually an asymptotic t-test.

[4] For instance, in the so-called queen's case, a first order adjacency neighbourhood matrix would consider as neighbours of an area all other units that share a border or a vertex with it. By definition a location or unit is not contiguous to itself. Higher order adjacency matrices follow a recursive definition. Hence, a second order adjacency matrix would take as second order neighbours of a location all of the first order neighbours of its first order neighbours.

[5] The expressions for the mean and standard deviation of *I* vary depending on the assumptions made about the data and the nature of the spatial autocorrelation. The approach taken here is to assume that, asymptotically, Moran's *I* follows a standard normal distribution.

[6] Estimations in sections 4.1 to 4.3 were calculated using SpaceStat 1.8.

[7] Note that the expected value is the same for all spatial weights as it is only dependent on the number of observations, i.e. -0.001 = 1/844. Conversely, the standard deviation (column SD in Table 3) is a function of the spatial weights matrix and so it is different for each one, with the SD decreasing for higher order matrices.

[8] See Fotheringham and Brunsdon (1999) and Brunsdon *et al*., (1999) for a discussion on how this technique can be used to explore local variability in some cases.

[9] GWR version 2.0, the software used to estimate the models in this section, allows the selection from a number of criteria for determining the bandwidth.

Acknowledgements

The research was funded under the (UK) National Environment Research Council, Urban Regeneration and the Environment Programme (http://urgent.nerc.ac.uk/), grant GST/02/2241. The authors are grateful to a commercial data vendor for supplying the lifestyles data. Thanks are also due to Richard Webber and Mike Goodchild for very helpful comments on a draft version of this paper and to Richard Harris for the discussions about data issues and some of the results.

References

Anselin, L. (1988). *Spatial Econometrics, Methods and Models*. Dordrecht: Kluwer Academic.

Anselin, L. (1990). Spatial Dependence and Spatial Structural Instability in Applied Regression Analysis. *Journal of Regional Science*, 30, 185-207.

Anselin, L. (1992). *SpaceStat Tutorial: A Workbook for using SpaceStat in the Analysis of Spatial Data*, University of Illinois, Urbana-Champaign, 250.

Anselin, L. (1995). Local Indicators of Spatial Association - LISA. *Geographical Analysis*, 27, 93-115.

Anselin, L. (1996). The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association. In: M. Fischer, H. J. Scholten and D. J. Unwin (Ed), *Spatial Analytical Perspectives on GIS*, London: Taylor and Francis. 4, 111-125.

Ball, M. (1994). The 1980s property boom. *Environment and Planning A*, 26, 671-95.

Brunsdon, C., M. Aitkin, A. S. Fotheringham and M. E. Charlton (1999). A Comparison of Random Coefficient Modeling and Geographically Weighted Regression for Spatially Non-Stationary Regression Problems. *Geographical and Environmental Modeling*, 3(1), 47-62.

Brunsdon, C., A. S. Fotheringham and M. E. Charlton (1996). Geographically Weighted Regression: A Method for Exploring Spatial Non-Stationarity. *Geographical Analysis*, **28**, 281-298.

Campbell H. (1999). Institutional consequences of the use of GIS. In *Geographical Information Systems: Principles, Techniques, Management and Applications* (2nd Edn.) Eds P A Longley, M F Goodchild, D J Maguire, D W Rhind. New York: Wiley: 621-31

Casetti, E. (1972). Generating Models by the Expansion Method: Applications to Geographic Research. *Geographical Analysis*, **4**, 81-91.

Ceccato, V., R. Haining and P. Signoretta (2002). Exploring Offence Statistics in Stockholm City Using Spatial Analysis Tools. *Annals of the Association of American Geographers*, **92**(1), 29-51

Champion, T., C. Wong, A. Rooke, D. Dorling, M. Coombes and C. Brunsdon (1996). *The Population of Britain in the 1990s: a Social and Economic Atlas.* Oxford: Oxford University Press.

Charlton, M. E., A. S. Fotheringham and C. Brunsdon (2002). *Geographically Wieghted Regression Version 2.x User Manual and Installation Guide*. Newcastle: University of Newcastle, 39.

Dale, A. and A. Teague (2002). Microdata from the census: samples of anonymised records. In *The Census Data System*. Eds. P. Rees, D. Martin, P. Williamson. Chichester: Wiley: 203-12

Donnay, J. P., M. J. Barnsley and P. A. Longley (2001: Eds). *Remote Sensing and Urban Analysis.* London: Taylor and Francis

Fotheringham, A. S., and C. Brunsdon (1999). Local Forms of Spatial Analysis. *Geographical Analysis*, **31**: 340-58.

Fotheringham, A. S., C. Brunsdon and M. E. Charlton (1998). Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial data Analysis. *Environment and Planning A*, **30**, 1905-1927.

Fotheringham, A. S., C. Brunsdon and M. E. Charlton (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis.* London: SAGE Publications Ltd, 270.

Fotheringham, A. S., C. Brinsdon and M. E. Charlton (2002). *Geographically Weighted Regression: the Analysis of Spatialy Varying Relationships*. Chichester: Wiley.

Fotheringham, A. S., M. E. Charlton, and C. Brunsdon (1996). The Geography of Parameter Space: An Investigation into Spatial Non-Stationarity. *International Journal of Geographic Information Systems*, **10**: 605-27

Fotheringham, A. S., M. E. Charlton and C. Brunsdon (2001). Spatial Variations in School Performance: A Local Analysis Using Geographically Weighted Regression. *Geographical and Environmental Modelling*, **5**, 43-66.

Getis, A. and J. K. Ord (1996). Local Spatial Statistics: An Overview. In: Longley, P. A. and M. Batty (Ed), *Spatial Analysis: Modelling in a GIS Environment*, Cambridge: GeoInformation International, 261-277.

Goodchild, M. F. and P. A. Longley (1999). Modern Geographic Information Systems and Model Linking. In A. Stein, F. W. T. Penning De Vries (Eds) *Data and Models In Action: Methodological Issues In Production Ecology* (Current Issues In Production Ecology Vol. 5). Dordrecht: Kluwer: 103-118

Gordon D. and C. Pantazis (1995). *Breadline Britain in the 1990s*. York: Joseph Rowntree Foundation.

Gordon D., L. Adelman, K. Ashworth, J. Bradshaw, R. Levitas, S. Middleton, C. Pantazis, D. Patsios, S. Payne, P. Townsend and J. Williams (2000). *Poverty and Social Exclusion in Britain*. York: Joseph Rowntree Foundation

Hall P. and U. Pfeiffer (2000). *Urban Future 21: a Global Agenda for Twenty-First Century Cities*. London: E & FN Spon

Harris, R. J. (1999). Geodemographics and the Analysis of Urban Lifestyles. School of Geographical Sciences: University of Bristol, 383pp. (Unpublished).

Harris, R. J. and P. A. Longley (2000). New Data and Approaches for Urban Analysis: Modelling Residential Densities. *Transactions in GIS*, **4**, 217-234.

Harris R. J. and P. A. Longley (2002). Creating Small Area Measures of Urban Deprivation. *Environment and Planning A* 34: 1073-93

Judge, G. G., R. C. Hill, W. E. Griffiths, H. Lutkepohl and T. C. Lee (1982). *An Introduction to the Theory and Practice of Econometrics.* New York: Wiley.

Johnston R. J. (1999). Geography and GIS. In *Geographical Information Systems: Principles, Techniques, Management and Applications* (2nd Edn.) Eds P A Longley, M F Goodchild, D J Maguire, D W Rhind. New York: Wiley: 39-47

Lee P. (1999). Where Are the Deprived? Measuring Deprivation in Cities and Regions. In *Statistics In Society: the Arithmetic Of Politics* Eds D Dorling, S Simpson London: Arnold: 172-180

Lemmen C. and P. van Oosterom (2002). Cadastral Systems II**.** *Computers, Environment and Urban Systems* 25: 355-60

Longley, P. A. (2002). Will Developments In Urban Remote Sensing and GIS Lead To 'Better' Urban Geography? *Progress In Human Geography* 26: 231-9

Longley, P. A. and R. J. Harris (1999). Towards a new digital data infrastructure for urban analysis and modelling. *Environment and Planning B Planning and Design,* 26: 855-78.

Longley, P. A., M. F. Goodchild, D. J. Maguire, D. W. Rhind (2001). *Geographic Information Systems and Science*. Chichester: Wiley

Longley, P. A. and T. V. Mesev (2002). Measurement of Density Gradients and Space-Filling in Urban Systems. *Papers in Regional Science* 81: 1-28

Mincer, J. (1974). *Schooling, Experience and Earnings*. Columbia University Press: New York.

Morad, M. (2002). British Standard 7666 as a Framework for Geocoding Land and Property Information the UK**.** *Computers, Environment and Urban Systems* 25: 483-92

Olligschlaeger, A. M. (1997). Spatial Analysis of Crime Using GIS-Based Data: Weighted Spatial Adaptive Filtering and Chaotic Cellular Forecasting with Applications to Street Level Drug Markets. PhD Thesis, http://www-2.cs.cmu.edu/~olli/dissertation.html, Unpublished.

Openshaw, S. (1984). The Modifiable Area Unit Problem. *CATMOG 38*, Norwich: GeoInformation.

Rees, P. (1998). What do you want from the 2001 Census? Results of an ESRC/JISC Survey of User Views. *Environment and Planning A* 30: 1775-96

Rees, P. and D. Martin (2002). The debate about census geography. In *The Census Data System*. Eds. P. Rees, D. Martin, P. Williamson. Chichester: Wiley: 27-36

Rose, D. and D. Pevalin (2002). *A Researcher's Guide to the National Statistics Socio-Economic Classification.* London: Sage

Senior, M., H. Williams and G. Higgs (2000). Urban-Rural Mortality Differentials: controlling for material deprivation, *Social Science and Medicine,* 51, 289-305.

Swetnam, R. D., C. I. Tindall, J. M. Cook, S. J. Pepler, R. P. Shaw (2002). Collation, management and dissemination of environmental research relating to urban areas in the UK. The approach used within the Natural Environment Research Council's URGENT Programme. *Computers, Environment and Urban Systems*, 26, 63-84.

Webber, R. and M. Farr (2001). MOSAIC: from an area classification system to individual classification. *Journal of Targeting, Measurement and Analysis for Marketing*, 10 (1)

Webber, R. and P. A. Longley (2002). Similarity and Proximity: Conflicting or Potentially Complementary Approaches to the Spatial Measurement of Social Phenomena. In Longley, P. A. and M. Batty (Eds) *Advanced Spatial Analysis*, Redlands, ESRI Press, under review.

White, M. (2002). Clearer poverty definition 'vital'. *The Guardian* 27 August: 2.

Williams, C. and J. Windebank (2001). Reconceptualising paid informal exchange: some lessons from English cities. *Environment and Planning A*, 33, 121-40.

Zhang, J. and M. F. Goodchild (2002). *Uncertainty in Geographical Information*. London: Taylor and Francis.

Table 1: The results of OLS regression.

OBS	844	DF	835
R^2	0.6546	Adj- R^2	0.6513
F-test	197.85	Prob	0.0000

VARIABLE	**COEFF**	**SD**	**t-value**
CONSTANT	1.470	0.170	8.639
OLDPSN	0.029	0.006	4.517
2534NK	0.022	0.005	4.033
WKWIFE	0.016	0.003	5.238
BIGACC	0.021	0.003	7.352
COUNCL	-0.004	0.001	-3.086
HHTCAR	0.035	0.005	6.893
UNSKLD	-0.012	0.005	-2.736
QUALML	0.012	0.002	6.573

Table 2: The results of robust OLS estimation

OBS	844	DF	835
R^2	0.6546	Adj- R^2	0.6513
F-test	197.85	Prob	0.0000

VARIABLE	COEFF	SD	z-value
CONSTANT	1.470	0.456	3.225
OLDPSN	0.029	0.009	3.212
2534NK	0.022	0.007	3.039
WKWIFE	0.016	0.006	2.741
BIGACC	0.021	0.003	6.310
COUNCL	-0.004	0.002	-2.209
HHTCAR	0.035	0.008	4.442
UNSKLD	-0.012	0.004	-3.293
QUALML	0.012	0.003	4.819

Table 3: Moran Statistics for all variables at different lags.

Variables[x]	Neighbourhood	I	SD	Z-VALUE
OLDPSN	First Order	0.441	0.021	21.536
	Second Order	0.259	0.014	19.211
	Third Order	0.200	0.010	19.225
	Fourth Order	0.166	0.009	19.101
	Fifth Order	0.102	0.008	13.163
2534NK	First Order	0.654	0.021	31.894
	Second Order	0.557	0.014	41.319
	Third Order	0.453	0.010	43.333
	Fourth Order	0.376	0.009	43.002
	Fifth Order	0.284	0.008	36.192
WKWIFE	First Order	0.250	0.020	12.281
	Second Order	0.160	0.013	11.983
	Third Order	0.088	0.010	8.500
	Fourth Order	0.050	0.009	5.907
	Fifth Order	0.028	0.008	3.689
BIGACC	First Order	0.615	0.020	30.071
	Second Order	0.467	0.014	34.657
	Third Order	0.358	0.010	34.317
	Fourth Order	0.251	0.009	28.789
	Fifth Order	0.168	0.008	21.530
COUNCL	First Order	0.552	0.021	26.951
	Second Order	0.322	0.014	23.902
	Third Order	0.180	0.010	17.260
	Fourth Order	0.104	0.009	11.947
	Fifth Order	0.047	0.008	6.071
HHTCAR	First Order	0.591	0.021	28.842
	Second Order	0.413	0.014	30.677
	Third Order	0.255	0.010	24.418
	Fourth Order	0.155	0.009	17.833
	Fifth Order	0.086	0.008	11.103
UNSKLD	First Order	0.242	0.021	11.844
	Second Order	0.176	0.014	13.145
	Third Order	0.113	0.010	10.892
	Fourth Order	0.071	0.009	8.249
	Fifth Order	0.069	0.008	8.873
QUALML	First Order	0.657	0.021	32.064
	Second Order	0.595	0.014	44.066
	Third Order	0.527	0.010	50.406
	Fourth Order	0.435	0.009	49.677
	Fifth Order	0.361	0.008	45.924
INC	First Order	0.510	0.021	24.917
	Second Order	0.385	0.014	28.597
	Third Order	0.302	0.010	28.988
	Fourth Order	0.232	0.009	26.576
	Fifth Order	0.173	0.008	22.174

Table 4: Instrumental variables estimates.

Dependent Variable	INC	**OBS**	844
VARS	5	**DF**	838
R^2	0.7002	**Sq. Corr.**	0.6615

VARIABLE	**COEFF**	**SD**	**z-value**
W_INC	0.243	0.060	4.031
OLDPSN	0.034	0.007	5.257
2534NK	0.024	0.005	4.472
WKWIFE	0.025	0.003	7.799
BIGACC	0.017	0.004	4.805
HHTCAR	0.034	0.006	6.219
UNSKLD	-0.010	0.003	-3.048
QUALML	0.007	0.002	2.830

Diagnostics for spatial dependence

Lagrange Multiplier (error)	**DF**	**VALUE**
First order standardized weights	1	10.6036*
Second order standardized weights	1	0.4375
Third order standardized weights	1	0.9244
Fourth order standardized weights	1	0.7320
Fifth order standardized weights	1	0.2431

Table 8: Monte Carlo Test

Parameter	P-value
CONSTANT	0.15
OLDPSN	0.49
2534NK	0.70
WKWIFE	0.15
BIGACC	0.43
COUNCL	0.14
HHTCAR	0.77
UNSKLD	0.00
QUALML	0.08

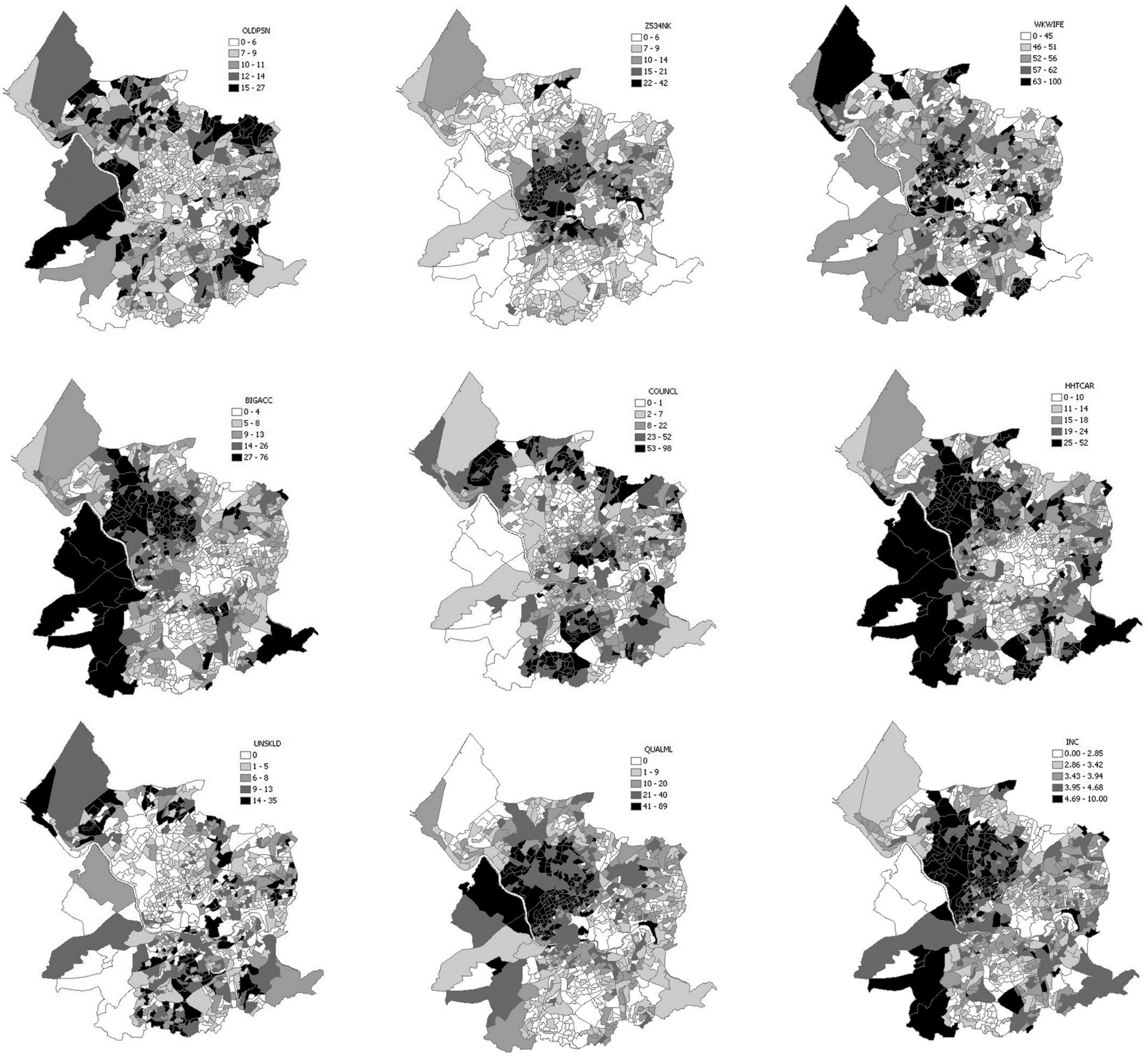

Figure 1: Spatial distribution by quintile ranges for: old residents (OLDPSN: top left); young childless households (2534NK: top centre); working wives (WKWIFE: top right); large accommodation (BIGACC: middle left); residents in council property (COUNCL: middle centre); multiple car ownership (HHTCA: middle right); unskilled workers (UNSKLD: bottom left); qualified male residents (QUALML: bottom centre); income score (INC: bottom right).

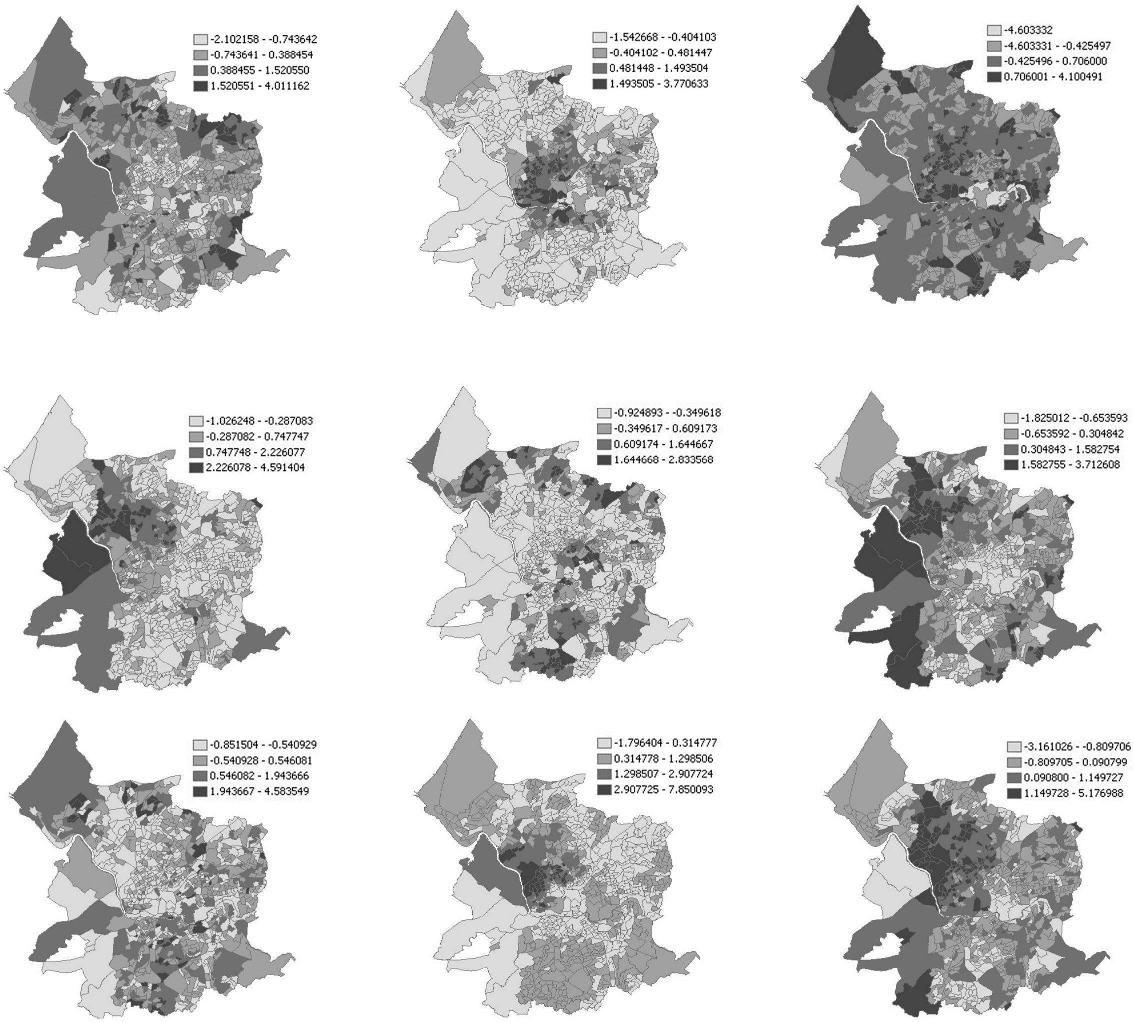

Figure 2: Local Moran's I statistic for: old residents (OLDPSN: top left); young childless households (2534NK: top centre); working wives (WKWIFE: top right); large accommodation (BIGACC: middle left); residents in council property (COUNCL: middle centre); multiple car ownership (HHTCA: middle right); unskilled workers (UNSKLD: bottom left); qualified male residents (QUALML: bottom centre); income score (INC: bottom right).

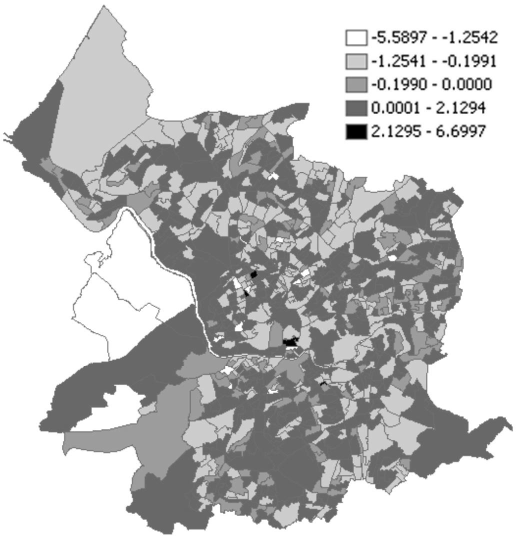

Figure 3: Residuals from the Instrumental Variables estimation.

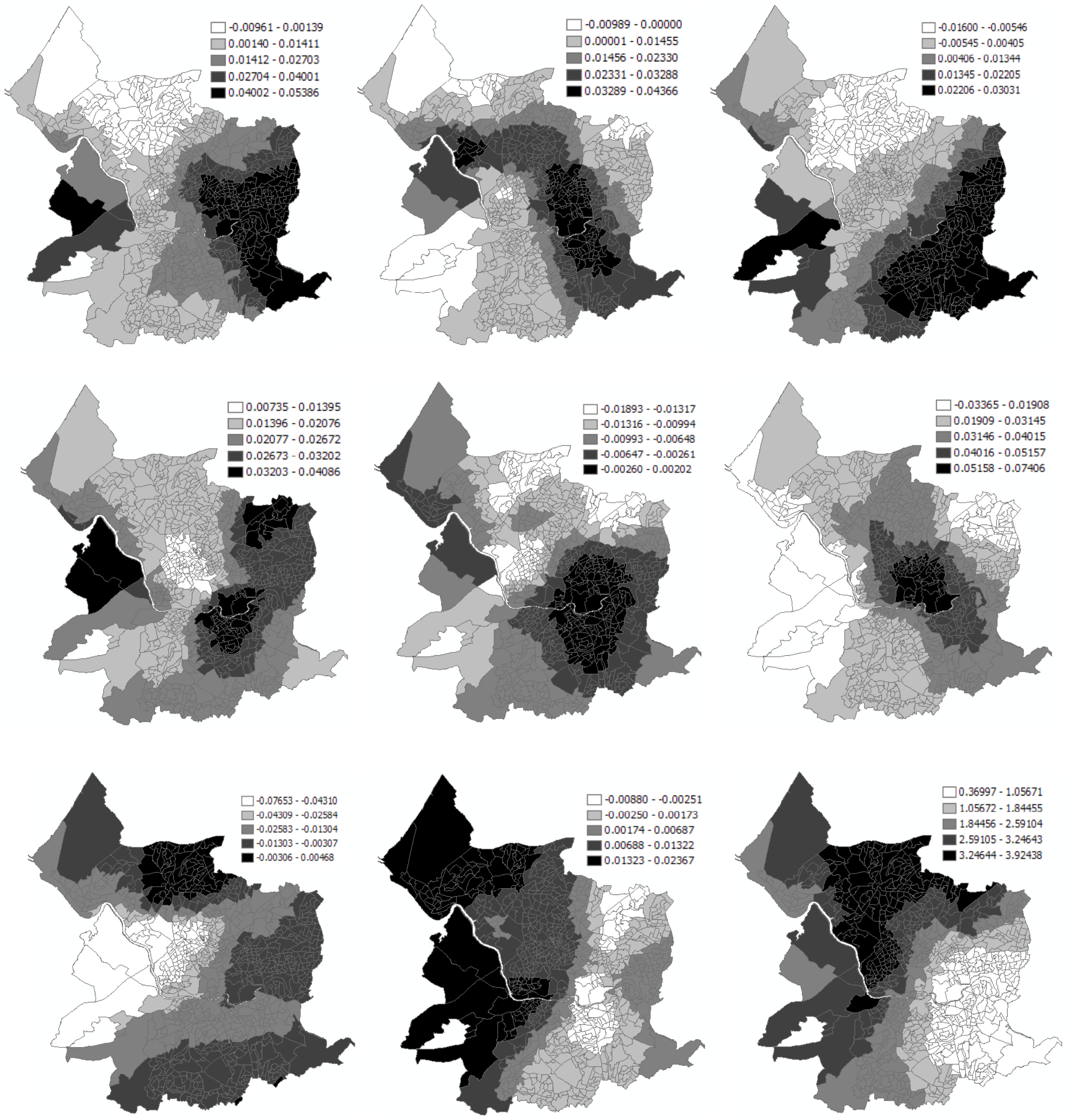

Figure 4: GWR parameter variation across the study area for: old residents (OLDPSN: top left); young childless households (2534NK: top centre); working wives (WKWIFE: top right); large accommodation (BIGACC: middle left); residents in council property (COUNCL: middle centre); multiple car ownership (HHTCA: middle right); unskilled workers (UNSKLD: bottom left); qualified male residents (QUALML: bottom centre); and constant (β_{0i} : bottom right).

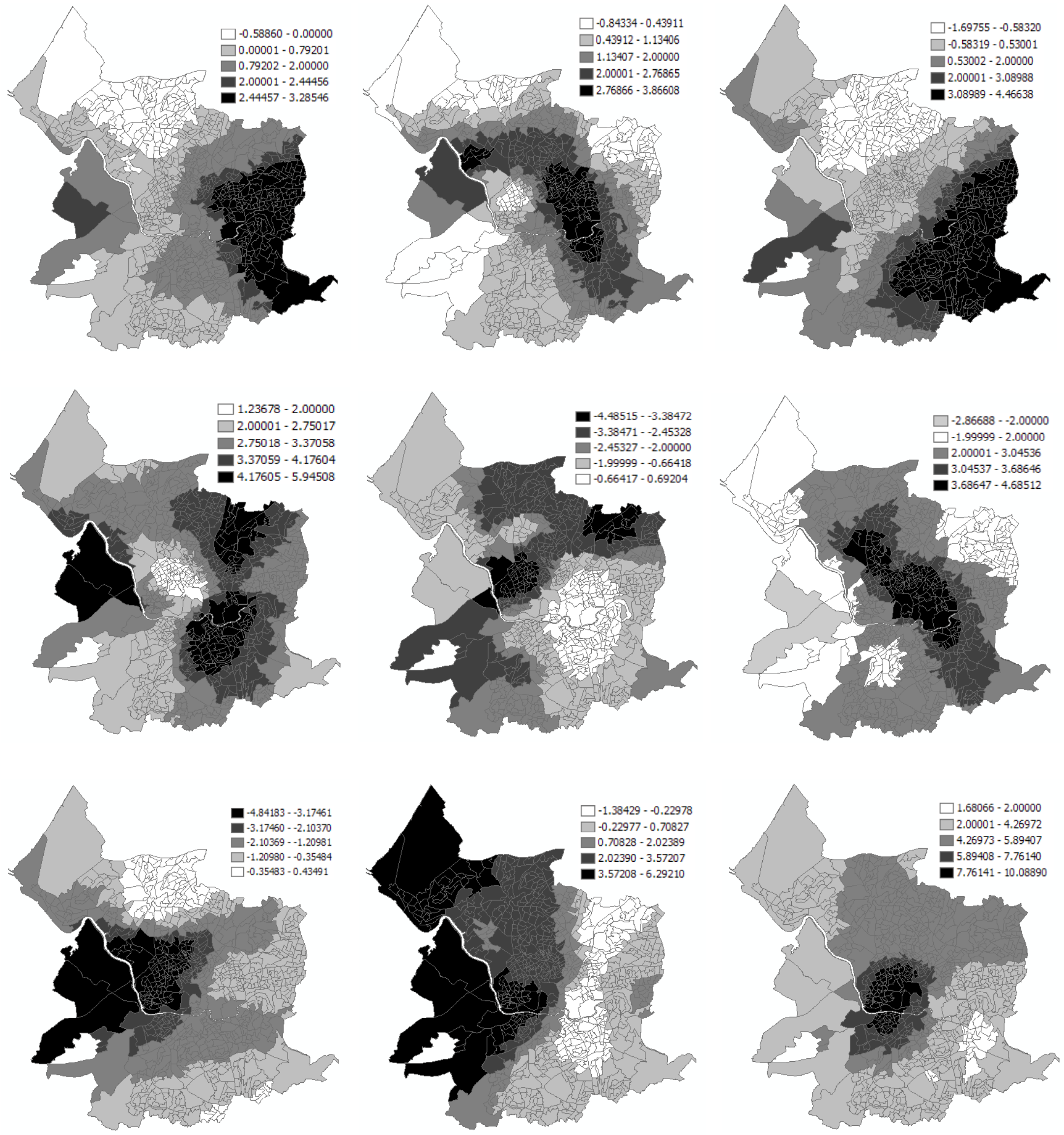

Figure 5: Parameter t-values for: old residents (OLDPSN: top left); young childless households (2534NK: top centre); working wives (WKWIFE: top left); large accommodation (BIGACC: middle left); residents in council property (COUNCL: middle centre); multiple car ownership (HHTCA: middle right); unskilled workers (UNSKLD: bottom left); qualified male residents (QUALML: bottom centre); and constant (β_{0i} : bottom right).

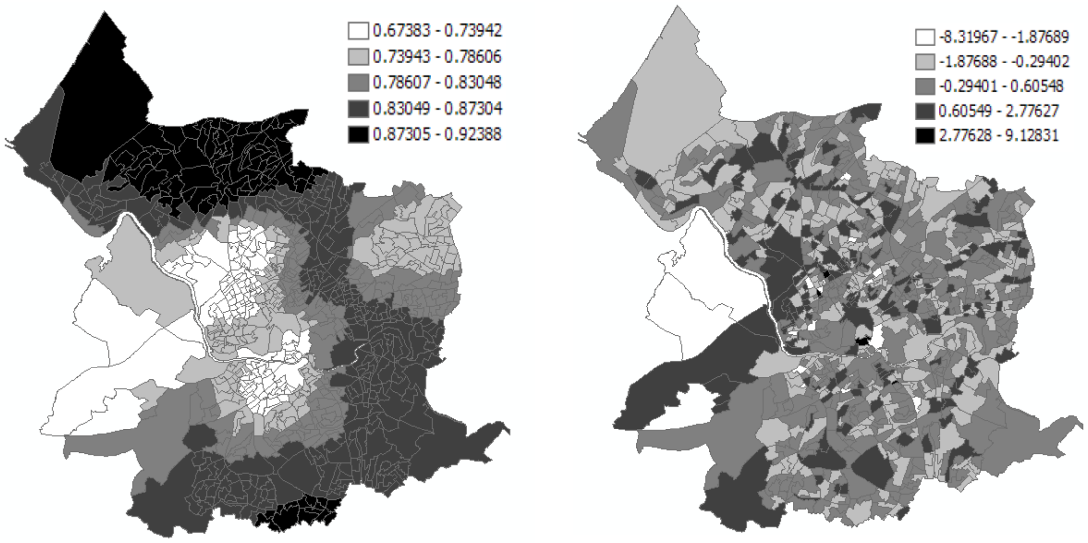

Figure 6: R-squared (left) and residuals (right) for the GWR regression

[i] The household income bands and their values as rescaled by Harris (1999) are: 1) household income under £5,000 (rescaled to 1.2 to accommodate social security thresholds); 2) between £5,000 and £9,999 (rescaled to 2); 3) between £10,000 and £14,999 (rescaled to 3); 4) between £15,000 and £19,999 (rescaled to 4); 5) between £20,000 and £29,000 (rescaled to 5.5); 6) between £30,000 and £39,999 (rescaled to 7.5); 7) household income over £40,000 (rescaled to 10).

[ii] Estimations in this section were performed using SpaceStat 1.8 (Anselin, 1992) which, in addition to calculation of spatial statistics, is also a convenient environment for performing aspatial multiple regression.

[iii] The significance of individual coefficients is based on the standard normal distribution because the estimation method is based on asymptotic considerations. Hence the z-value column of Table 2, which is actually an asymptotic t-test.

[iv] For instance, in the so-called queen's case, a first order adjacency neighbourhood matrix would consider as neighbours of an area all other units that share a border or a vertex with it. By definition a location or unit is not contiguous to itself. Higher order adjacency matrices follow a recursive definition. Hence, a second order adjacency matrix would take as second order neighbours of a location all of the first order neighbours of its first order neighbours.

[v] The expressions for the mean and standard deviation of *I* vary depending on the assumptions made about the data and the nature of the spatial autocorrelation. The approach taken here is to assume that, asymptotically, Moran's *I* follows a standard normal distribution.

[vi] Estimations in sections 4.1 to 4.3 were calculated using SpaceStat 1.8.

[vii] Note that the expected value is the same for all spatial weights as it is only dependent on the number of observations, i.e. -0.001 = 1/844. Conversely, the standard deviation (column SD in Table 3) is a function of the spatial weights matrix and so it is different for each one, with the SD decreasing for higher order matrices.

[viii] See Fotheringham and Brunsdon (1999) and Brunsdon *et al*., (1999) for a discussion on how this technique can be used to explore local variability in some cases.

[ix] GWR version 2.0, the software used to estimate the models in this section, allows the selection from a number of criteria for determining the bandwidth.

[x] The variables reported here are the dependent variable in later models (INC) and four socio-economic attributes from the Census dataset found to be statistically significant in explaining its variability out 50 variables originally considered.