Geographical & Environmental Modelling, Vol. 3, No. 1, 1999, 47–62

A Comparison of Random Coefficient Modelling and Geographically Weighted Regression for Spatially Non-stationary Regression Problems

CHRIS BRUNSDON, MURRAY AITKIN, STEWART FOTHERINGHAM & MARTIN CHARLTON

ABSTRACT *The problem of locally varying coefficients in geographical applications is considered. Two approaches to this are then discussed—geographically weighted regression and random coefficient models. The latter is considered in two forms: firstly, the case that only the intercept coefficient is random; and then the case in which all coefficients are random. All these techniques are applied to a data set derived from the 1991 UK Census of Population relating to limiting long-term illness.*

Introduction

In many geographical models, there are strong arguments that the relationships between variables are not fixed over space. Consider, for example, the case of a house price prediction model, or hedonic model. Such a model attempts to predict the market price of a house by considering the component parts of the house, such as the number of bathrooms, number of bedrooms, size of garden and so on. However, although it is likely that all these variables will influence the price of the house, the degree of influence may well vary geographically. For example, areas close to good schools may well be popular with house buyers having families with several small children. These buyers may also place greater premiums than average on a second bathroom, or a large number of bedrooms. Another example is that of local crime rates: the causes of crime may well differ from one place to another, so that observed correlations between potential 'explanatory' variables and crime rates may well differ from place to place. A similar situation could occur in epidemiology. A high (or low) incidence of some disease could be caused by different phenomena in different places.

Generally, considering processes in this way presents an interesting geographical approach. Rather than modelling a 'global' relationship between a set of variables, an approach that allows the nature of the model to change spatially could be used instead. For example, an often-used model is the general linear model (GLM). Here, the mean level of the ith response variable, $E(y_i)$, modelled as a function of the

Chris Brunsdon, Department of Town and Country Planning, University of Newcastle upon Tyne, Newcastle NE1 7RU. E-mail: chris.brunsdon@ncl.ac.uk.

1361-5939/99/010047-16

regression parameter vector, $\boldsymbol{\beta}$ and a vector of predictor variables, X_i; thus:

$$G[E(y_i)] = \boldsymbol{\beta}' X_i \tag{1}$$

G is referred to as the link function. In a typical geographical application, the index i would refer to a location, and X_i would refer to a set of predictor variables measured at that location. However, the regression parameter vector β is the same for all locations and therefore the relationship between the predictor and response variables remains the same everywhere. To model the phenomena discussed earlier, β would need to vary from place to place, giving a revised model:

$$G[E(y_i)] = \beta_i' X_i \tag{2}$$

If there was some way of estimating the vector β_i at each location, then spatially varying parameters of models could be mapped using a geographic information system (GIS). It is intended that this should provide greater insight into the geographical context of the process being modelled through a graphical representation of the changing geographical characteristics of the relationships. Unfortunately, without further information, there are some problems with model identifiability. If the n β_i values are allowed to take any values, then for most data sets there will be an infinite number of β_i's solving equation (2) with $E(y_i) = y_i$. To tackle this problem, some constraining assumptions must be placed on the nature of the β_i's.

Two very different approaches to this are random coefficients modelling (RCM) and geographically weighted regression (GWR). In RCM, the β_i are considered as random quantities. The problem of model calibration then becomes a problem of estimating the probability distribution of the β_i's, and then of using Bayes' theorem to provide estimates of the probability distribution of each β_i given X_i and y_i. This in turn will lead to a point estimate of each β_i.

In GWR (Brunsdon *et al.*, 1996), the β_i are not assumed to be random, but are assumed to be a function of the coordinates in geographical space of the ith observation, (u_i, v_i). If we are dealing in zonal data, then (u_i, v_i) should represent the centroid of the ith zone. The regression model now becomes

$$G[E(y_i)] = \boldsymbol{\beta}(u_i, v_i)' X_i \tag{3}$$

Thus, β can be regarded as a vector of functions of two variables, say $[f_1(u, v), f_2(u, v), \ldots f_m(u, v)]'$. GWR uses kernel-based techniques to obtain estimates of $f_1, f_2 \ldots f_m$ given the u_i's, v_i's, X_i's and y_i's. To obtain the β_i estimates here, one uses the values of these function estimates at (u_i, v_i).

Despite the clear differences between the two approaches, there are some common qualities. Most notably, both approaches are in part non-parametric. In RCM, no assumptions are made about the probability distribution of the random β_i's. In GWR, no assumptions (apart from continuity) are made about $f_1, f_2 \ldots f_m$. Thus, both approaches allow reasonable flexibility in modelling although in different ways. The major difference is perhaps the fact that GWR takes spatial location explicitly into account, while RCM does not.

The aim of this paper is to contrast the two methods by using them to analyze the same data set. This data set is derived from the OPCS 1991 Census of Great Britain and Northern Ireland, and was drawn up to investigate the relationship between limiting long-term illness (LLTI) and a number of social factors. In the next

section, the data will be discussed. Following this, GWR and RCM will be discussed in more detail; in particular, two variants of RCM will be considered—one in which only the intercept term varies randomly, and one in which all terms do. Following this, the results of the analyses will be reported. The paper will conclude with a discussion of these results.

The LLTI Data Set

A set of predictor variables was chosen to represent various hypothesized contributory factors to the prevalence of LLTI. These are listed in the following. These are based on the chapter by Rees (Rees, 1995) in the *Census User's Handbook* for 1991.

- LLTI: The percentage of persons in households in each ward where a member of the household has some LLTI. This is the response variable. Note that to control for different age profiles in areas, this is only computed for people aged 45–65—an age category that is perhaps most likely to suffer from LLTI as a result of working in the extractive industries.
- CROWD: This is the proportion of households in each census ward having an average of more than one person per room. This is an attempt to measure the level of cramped housing conditions in each ward.
- DENSITY: This is the housing density of each ward, measured in millions per square kilometer. This is intended to measure 'rurality' of areas. Note the differences between this and the previous variable—a remote village with poor housing conditions may well score low in this variable but high in the previous.
- UNEMP: The proportion of unemployment persons in an area. This is generally regarded as a measure of economic well-being for an area.
- SCLASS1: The proportion of heads of households whose jobs are classed in social class 1 in the census. These are professional and managerial occupations. While the previous variable measures general well-being, this measures affluence.
- SPFAM: The proportion of single-parent families in an area. This is an attempt to measure the nature of household composition in areas.

The dependent variable here is assumed to be LLTI and the remaining variables are used as predictors. In the statistical analysis package SAS, the ratios were calculated as follows (census cells are addressed as CNNnnnn where NN is the table number and nnnn is the cell number within that table.):

```
LLTI    = 10000*((C120019 + C120022 + C120025)/
          (C020133 + C020144 + C020155 + C020166));
UNEMP   = 10000*(C080134/C080020);
CROWD = 10000*(C230001 - C230002)/C230001;
SPFAM   = 10000*(C400046/C400001);
SCLASS1 = 10000*(C900007/C900002);
DENSITY = 10000*(C230001/AREA);
```

All data are taken from the 1991 UK census local base statistics at ward level.

Random Intercept Models

The analysis presented here is based on a generalization of the binomial logit model, in which the number of people with LLTI in the area (UK Census Ward) is treated

as binomial with the number of 'trials' equal to the population size of the area, and regression model linear in the five explanatory variables X. So the regression model is $\eta_i = \boldsymbol{\beta}' X_i$ for the ith area.

In this form the model is not a two-level variance component model, as there are no individual-level variables, only aggregated individual and district-level variables.

Over the 595 wards in the study, this model is much too rigid, as we may expect that the logistic model cannot be a full specification over the whole region of all the important variation in the proportion of LLTI. Many relevant variables may have been omitted from the model, leading to misspecification of the simple logit model.

A simple way of representing this misspecification is by including a random intercept z in the model, with an unspecified distribution. We may think of this as representing all the omitted variables Z from the model, weighted by their respective regression coefficients, so that we have omitted a term $\gamma' Z$ from the model. But since the distribution of the unobserved variable Z is unknown, this is entirely equivalent to the omission of a single variable z with an unknown distribution, appearing as a random intercept term in the model. The model then becomes $\eta_i = z_i + \beta' X_i$. Such a model is known by statisticians as an 'overdispersed' or 'unobserved heterogeneity' model, as the random intercept induces additional variation in the response beyond that in the 'core' binomial logit model.

Recent advances in computational modelling (Aitkin, 1996) allow us to fit such models easily by (non-parametric) maximum likelihood (NPML), simultaneously estimating the regression coefficients $\boldsymbol{\beta}$ for the observed explanatory variables X and the distribution of the random intercept term as well. This distribution is estimated as a discrete distribution with masses π_k on a finite number of mass points z_k; the overall distribution of the response is therefore a mixed binomial distribution.

Random Coefficient Models

The foregoing analysis can be extended to general random coefficient models. The simple overdispersed model assumes that all regression coefficients except the intercept are constant over wards. But this can be relaxed and the model extended to a quite general random regression coefficient model in which some or all the explanatory variables have slopes that vary across districts.

Consider a simple example with a variable x_{1i} whose coefficient β_1 varies across wards. We index it by $\beta_{1i} = \beta_1 + u_i$, where u_i represents variation about a 'mean' β_1. The regression coefficients β_2 of the remaining variables X_2 are fixed. Then conditional on u_i and z_i, the regression model is

$$\eta_i = \beta_1' x_{1i} + \beta_2' X_{2i} + z_i + u_i' x_{1i}$$

while marginally z_i and u_i have an unknown joint distribution $\pi(z, u)$.

By estimating the joint distribution of z_i and u_i non-parametrically, we obtain the NPML as a discrete distribution on a finite number of points in the (z, u) plane, with an estimated mass π_k and estimated mass points z_k and u_k in the kth component.

We follow this approach in the NPML analysis of the logistic model over the 595 wards. All the regression coefficients are allowed to be random over the wards, and we allow 10 components in the mixture—that is, the random slopes are defined by at most 10 distinct points in the six-dimensional space of the regression coefficients and intercept. This restriction to 10 is rather arbitrary, and more components in the mixture could be allowed.

To provide a 'fitted' model for each district, we use the general properties of empirical Bayes shrinkage, to 'shrink' the regression coefficients of the individual within-ward regressions towards the common mean of the coefficients over wards, so providing a more stable prediction for each ward, especially those with small population bases.

The fitted values of the LLTI proportions in this analysis are based on the posterior ('shrunken') means of the regression coefficients for each ward. Note that there is no connection between adjacent wards; this leads to a greater appearance of randomness, or variation, in the fitted proportions between wards than for locally weighted regression or other methods which build in dependence between adjacent or neighbouring districts.

Geographically Weighted Regression

The problem here is to provide estimates of $\boldsymbol{\beta}(u_i, v_i)$, for each location i. This is achieved by considering data for places in the vicinity of the point location (u_i, v_i). For example, if one drew a circle of some radius, say r, around one particular (u_i, v_i), and calibrated an ordinary least squares regression model only on the basis of observations within this circle, then the $\boldsymbol{\beta}$ obtained could be thought of as an estimate of the regression parameter vector in the vicinity of (u_i, v_i). In short, it is an estimate of $\boldsymbol{\beta}(u_i, v_i)$, or β_i using the notation of equation (2). By evaluating β_i for each (u_i, v_i), it is possible to obtain a set of estimates of spatially varying parameters without specifying a functional form for the spatial variation. In a sense, this technique 'lets the data speak for itself' when providing estimates of each β_i. With some modifications, this is the underlying concept of GWR.

An initial consideration of this technique may raise issues relating to the notion of the 'circle of inclusion' of observations around each (u_i, v_i). This circle has been specified to have radius r, but what value should r take? Another issue that may be raised is the binary nature of inclusion of observations in the regression model calibration. An observation whose distance from (u_i, v_i) falls just below r will be included in the model, whereas one whose distance just exceeds this quantity will be excluded. It seems unnatural that the spatial association between the variables ends so abruptly.

The regression model centred around each (u_i, v_i) could be thought of as a weighted regression, with observations in the circle of inclusion weighted as unity, and other observations weighted as zero. Thus, for a given (u_i, v_i), the weight α_{ik} given to observation k would be

$$\alpha_{ik} = \begin{cases} 1, & \text{if } d_{ik} < r \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where d_{ik} is the distance between the locations of observations i and k. However, there is no reason to restrict the weighting function to a step function in this way. It is also possible to relate d_{ik} to α_{ik} with a continuous function. For example, a Gaussian distance decay-based weighting would be achieved by

$$\alpha_{ik} = \exp(-d_{ik}^2/2h^2)$$

Here, the value of the weight would decay gradually with distance, to the extent that when $d_{ik} = h$ the weighting would be about 0.05.

Functions such as this will be referred to as kernel functions or kernels, and denoted by the letter K as in $\alpha ik = K(d_{ik})$. In each case, the constant h provides some control of the range of 'circle of influence' of the geographical data, as r does in the basic step-function example, but the degree of weighting decays with distance rather than suddenly dropping to zero when a certain distance is reached. Generally, desirable features of a kernel function K are

- $K(0) = 1$.
- $\lim_{d \to \infty} K(d) = 0$.
- K is a monotone decreasing function for positive real numbers.

It is also interesting to note that although we have restricted our attention to estimating β_i values at points in space corresponding to the geographical locations of the observations, the same methodology can be applied to any point.

Another important issue in GWR is the choice of h—sometimes referred to as the kernel bandwidth. As stated earlier, this can greatly affect the properties of the β estimates. Following the advice of Silverman (1986) when considering kernel density estimates, there are occasions when subjective choice lends itself well to the problem in hand. If one has strong theoretically based prior beliefs about the value of h in a given situation, then it seems reasonable to make use of them.

However, there are many situations in which no such theoretical understandings exist, and in these cases some form of automatic, data-led choice of h may be more appropriate. One method suggested here is that of least squares cross-validation. A common calibration technique is that of least squares. Suppose for a pre-specified kernel function, the predicted value of y_i from GWR is denoted (as a function of h) by $\hat{y}_i(h)$. The sum of squared errors may then be written as

$$\mathrm{SS}(h) = \sum_i [y_i - \hat{y}_i(h)]^2 \tag{5}$$

A logical choice may then be to find h minimizing equation (5). However, at this stage, a problem is encountered. As $h \to 0$, $\hat{y}_i(h) \to y_i$, so that equation (5) is minimized when $h = 0$. To see why this is the case, note that for all K functions, $\alpha_{ii} = 1$, and that if $i \neq k$, then $h \to 0 \Rightarrow \alpha_{ik} \to 0$, so that the weighted regression is dominated by the term for observation i. This suggests that an unmodified least squares automatic choice of h would always suggest $h = 0$, or possibly result in computational errors. This problem can be avoided if, for each i, a GWR estimate of y_i is obtained by omitting the ith observation from the model. This is equivalent to replacing the kernel function K by a modified function K^* such that

$$K^*(0) = 0$$
$$K^*(d) = K(d), \qquad \text{if } d \neq 0$$

If the modified GWR estimate of y_i is denoted by $\tilde{y}_i(h)$, then the cross-validated sum of squared errors is denoted by

$$\mathrm{CVSS}(h) = \sum_i [y_i - \tilde{y}_i(h)]^2 \tag{6}$$

Choosing h to minimize equation (6) provides a method for choosing h automatically that does not suffer from the problems encountered by working with equation (5).

Table 1. Correlation matrix for the LLTI data

Variable	DENSITY	LLTI	SCLASS1	SPFAM	UNEMP
CROWD	0.249	0.371	−0.501	0.382	0.557
DENSITY		0.271	−0.185	0.149	0.480
LLTI			−0.458	0.237	0.711
SCLASS1				−0.249	−0.438
SPFAM					0.374

Results

For the LLTI data, both the random coefficient model and the GWR approach were used to obtain estimates of β_i. Since each β_i is a six-dimensional vector, each analysis yields six maps, one for each element of β_i over the 595 census wards. Firstly, however, one should consider the results of fitting the global model shown in equation (1).

A table of correlations between the variables is shown in Table 1. From this, it may be seen that although the predictor variables are far from independent, there are no correlations sufficiently high to cause problems with collinearity. The largest correlation is between LLTI and male unemployment—perhaps this is hardly surprising, as several studies have already highlighted linkages between deprivation and health problems (for example, see Townshend *et al.*, 1988).

Next, the regression model itself will be considered. Results for this are given in Table 2. From this, it can be seen that, at least globally, every variable has a statistically significant coefficient except SPFAM. The general fit of the model, as measured by the R^2 statistic, is good. Perhaps the most surprising results are the coefficients for DENSITY and CROWD, which are both negative. However, although it is possible that cramped housing and urban environments contribute to ill health, it is also the case that once people have become ill they may tend to move away from the worst environments. Also considering the relationship demographically, many elderly people are likely to suffer from LLTI, but these people are less likely to live in crowded households, as any children they have are likely to have left home. They are also likely to retire to rural areas in some cases, and so this demographic factor is likely to influence the coefficients for the CROWD and DENSITY variables.

The results of calibrating a GWR model for the LLTI data are now discussed. Firstly, spatial referencing must be considered. Data observations here correspond to census wards, the second smallest aggregational unit used in the 1991 census. After using Newton's method to minimize expression (6) with respect to h, a value of 17.1 km was chosen. The results of fitting this GWR are given in Figures 1–6.

Firstly, consider the map for the density coefficient. This is particularly interesting,

Table 2. Results of global regression model ($R^2 = 0.547$)

Variable	β Coefficient	SE (β)	Significance
CONSTANT	13.52	0.841	$p < 0.01$
CROWD	−7.592	2.48	$p < 0.01$
DENSITY	−4.506	1.62	$p < 0.01$
SCLASS1	−17.40	2.634	$p < 0.01$
SPFAM	−1.700	1.452	NS
UNEMP	46.07	2.411	$p < 0.01$

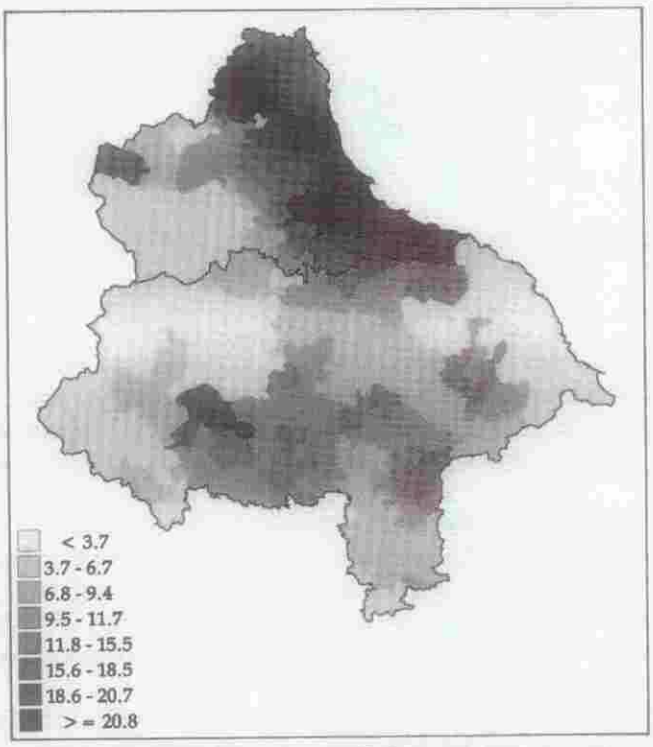

Figure 1. GWR results (coefficient: constant term).

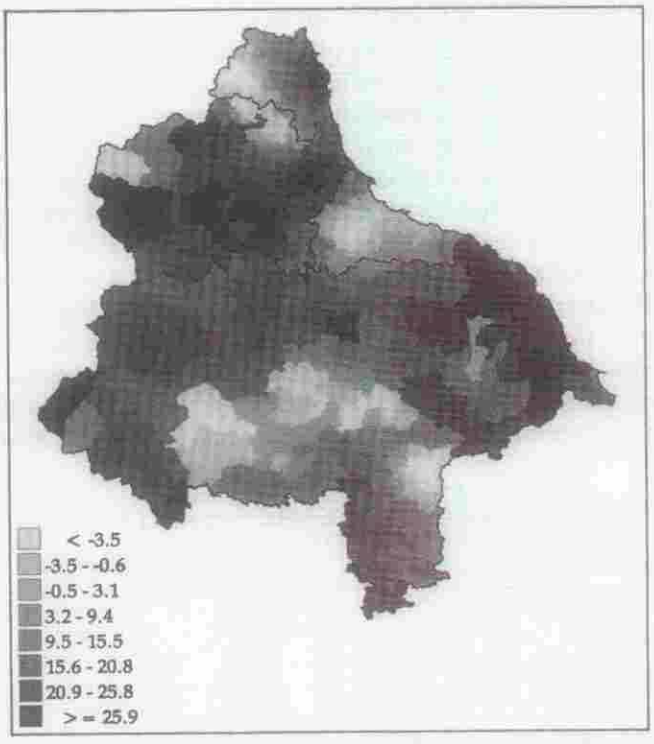

Figure 2. GWR results (coefficient: crowded households).

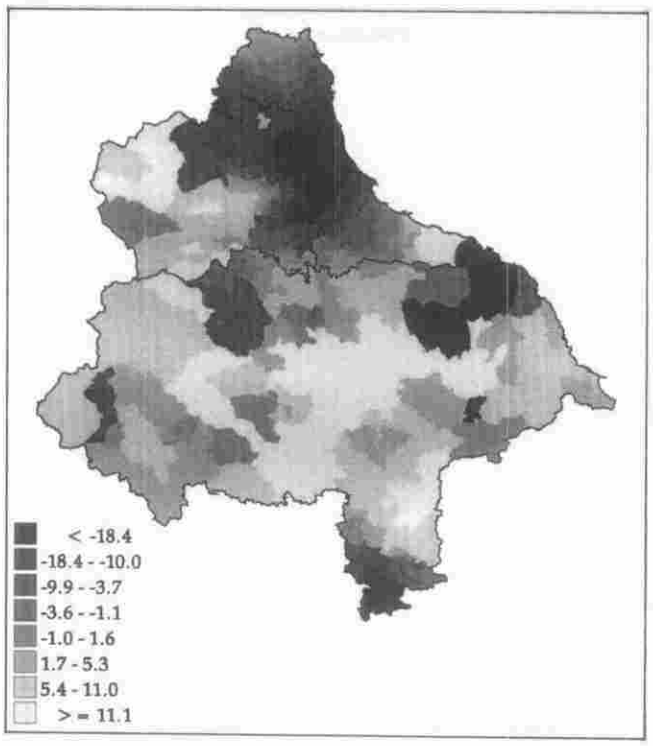

Figure 3. GWR results (coefficient: housing density).

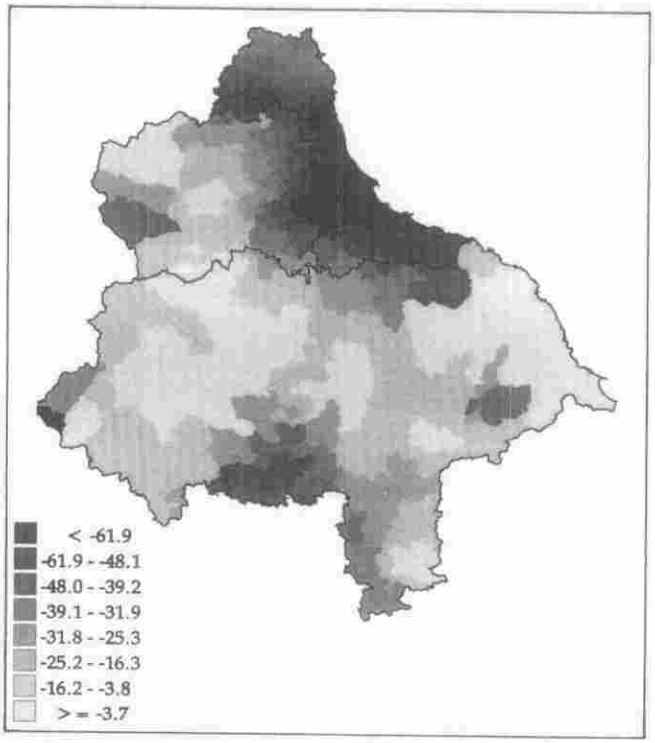

Figure 4. GWR results (coefficient: social class 1).

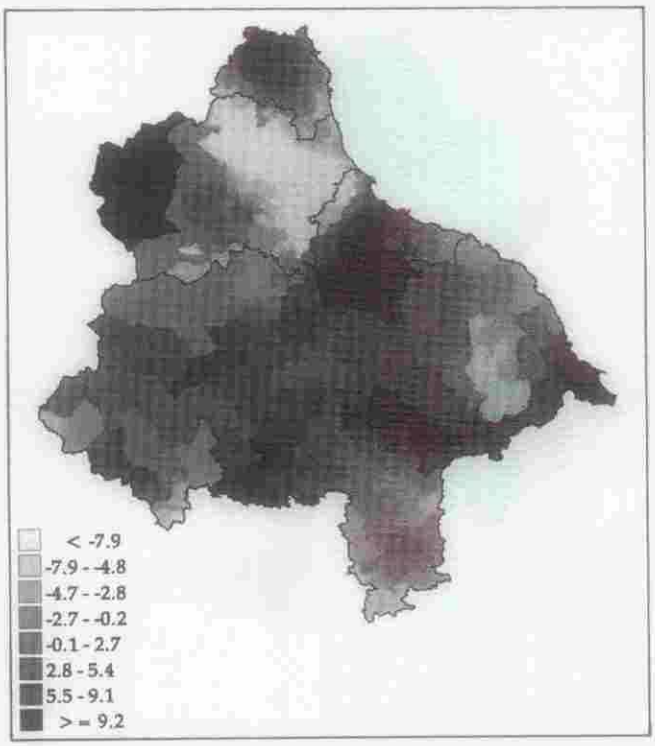

Figure 5. GWR results (coefficient: single-parent families).

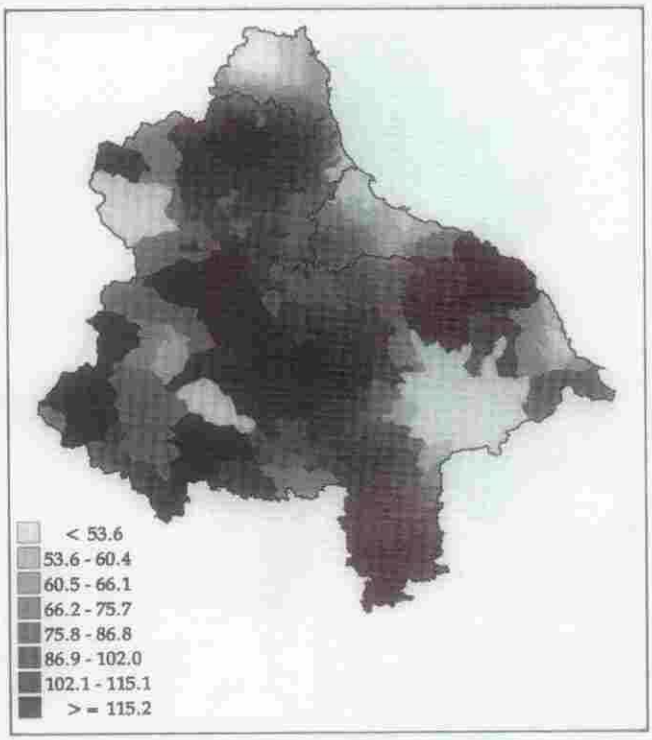

Figure 6. GWR results (coefficient: unemployment).

as the coefficient takes both positive and negative values. In particular, in an area to the north-east, there is a strong negative relationship (so that high density implies low LLTI), whereas in most other areas the relationship is positive. The latter is perhaps what one would intuitively expect to see—urban areas having a greater prevalence of long-term illness. But, one has to consider that in the Durham area there is a strong tradition of employment in the coal mining industry. Coal mining communities exist typically in pit villages—which are relatively sparsely populated compared to urban areas. It is also well known that a number of illnesses are associated with working in coal mines. In contrast, the main city in this area is Durham City, and much employment within the city is associated with its university, cathedral, shops and tourism industry. There are considerably fewer illnesses associated with occupations of this sort. In the context of this locality, a negative association between population density and LLTI is not surprising. However, when one looks further south on the map to the county of North Yorkshire, the areas having lower housing density are associated with more traditional rural communities, and so the more usual positive association between housing density and illness is seen.

Next, the maps of results for the RCMs are considered. In this instance, the non-parametric distribution estimators for the coefficients were based on 10 mass points—the maximum possible when using the GLIM macros. The shrinkage estimates of the ward-based coefficients are shown in Figures 7–13. Here, the mapped patterns exhibit more 'noise' than their GWR counterparts. However, this is hardly surprising given the spatial smoothing inherent in the GWR calibration process. In some cases, however, there are notable similarities between the two processes. For example, the estimate of the intercept term in the random intercept model has high

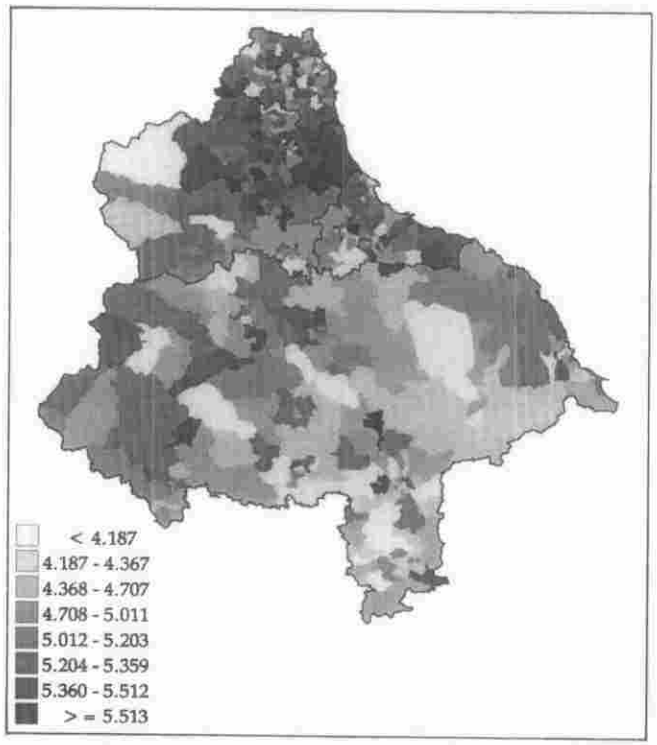

Figure 7. RCM results (random coefficient: unemployment).

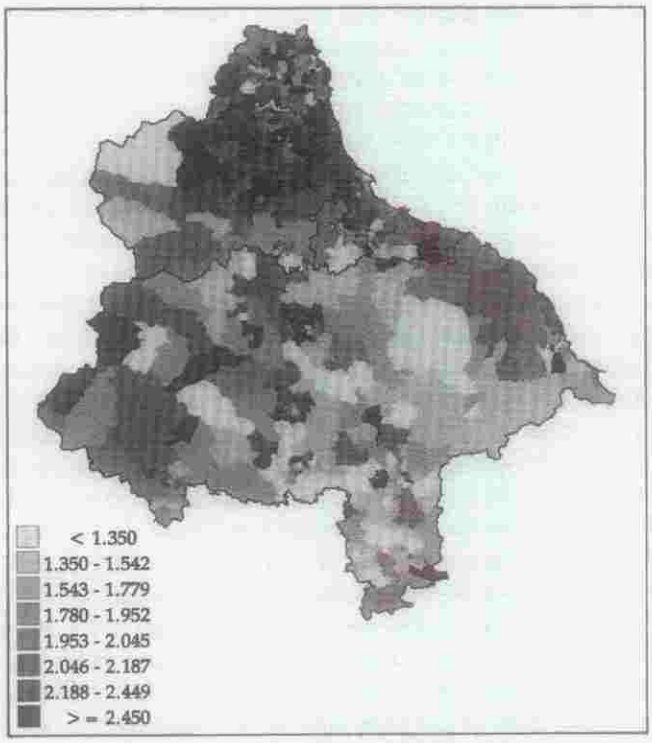

Figure 8. RCM results (random coefficient: crowding).

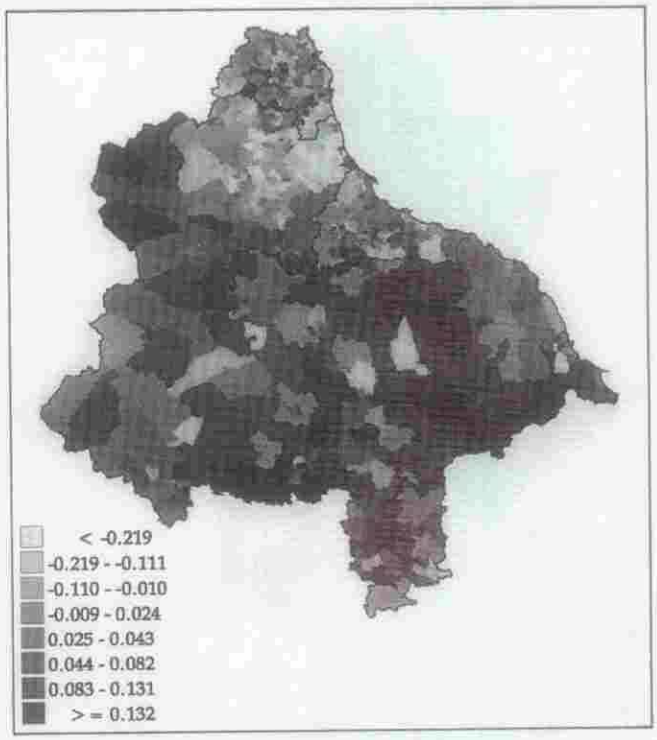

Figure 9. RCM results (random coefficient: single parents).

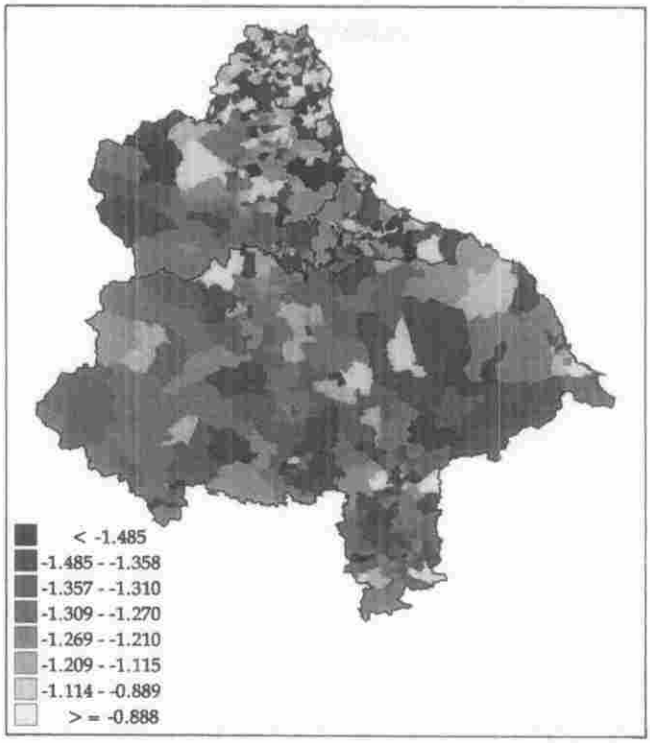

Figure 10. RCM results (random coefficient: social class 1).

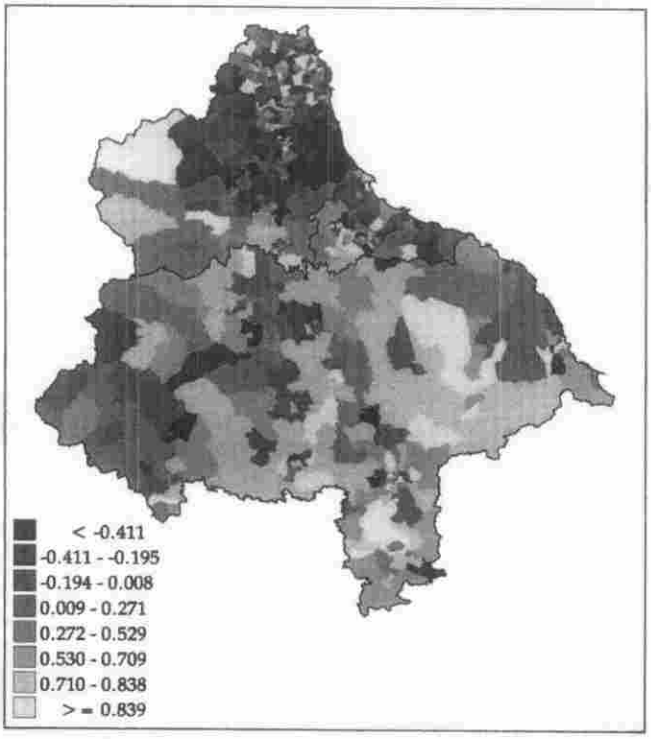

Figure 11. RCM results (random coefficient: housing density).

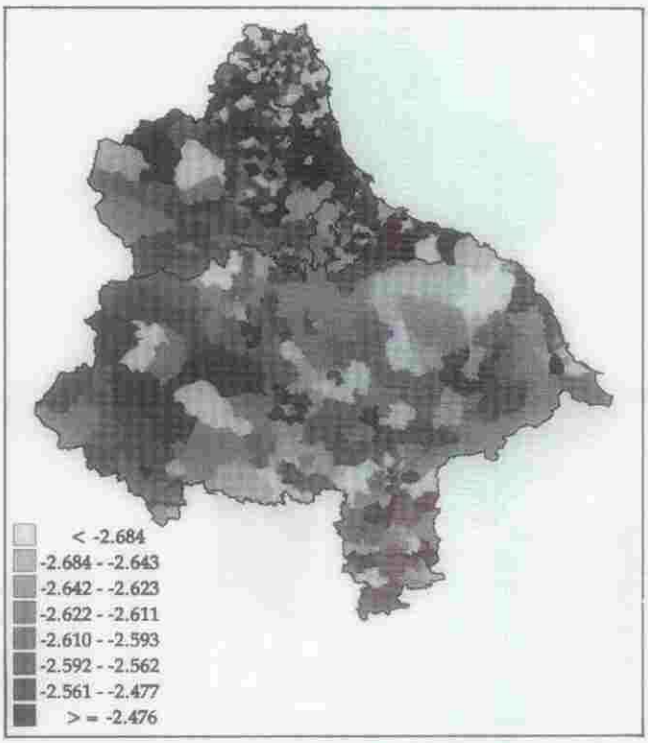

Figure 12. RCM results (random coefficient: intercept).

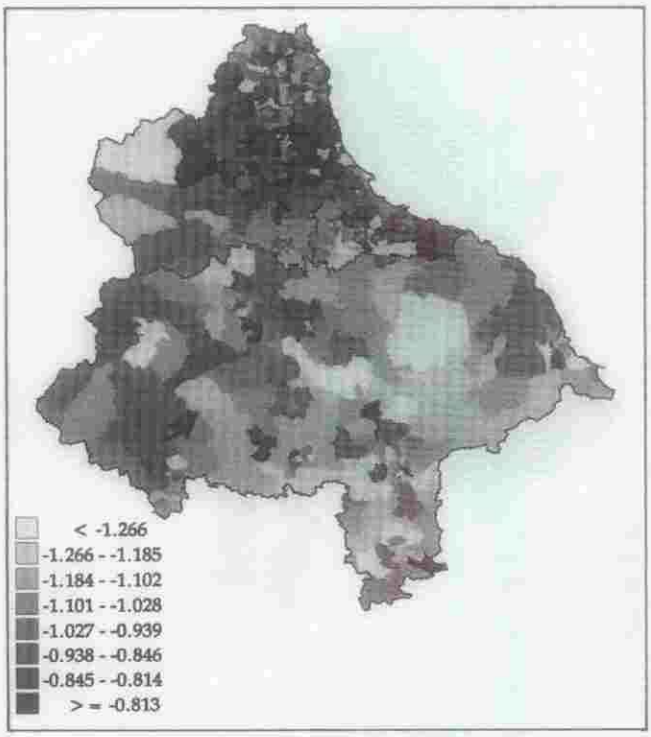

Figure 13. Random intercept model result (fixed model: intercept).

values in urban areas, as does the intercept estimate from GWR. However, the RCM model also exhibits some much lower values associated with other urban wards. In this sense, there are some notable differences between the two models.

Conclusions

In this paper, two very different approaches to modelling variability in regression parameters are considered. Both lend themselves to mapping—and are, therefore, GIS-able in accordance with Openshaw's requirement (Openshaw, 1994). However, the two analyses yield very different sets of results. GWR provides estimates using a mechanism that is essentially based on spatial smoothing, whereas the random coefficient model makes no spatial assumptions. The estimates in this case are 'shrunk' towards a global mean value, but no special attempt is made to reduce spatial 'roughness' in any way. The results of the two analyses reflect this difference in estimation procedure. Both sets of RCM maps appear less smooth than the corresponding GWR maps.

Which of the two approaches should be used? To answer this question, one needs to decide whether equation (2) or (3) best reflects the process used to generate the data. If the latter is most like the 'true' process, then RCM will allow noise in the model to introduce roughness into the local coefficient estimates. If the former is a better reflection of reality, the smoothing process in GWR will give unrealistically smooth estimates of the coefficient estimates. Thus, in the first instance, GWR is the better approach, while in the second some form of RCM would be better. Of course, if in reality there are no local effects, then a global regression model would provide the best approach.

Unfortunately, without prior knowledge, a choice is difficult to make. Perhaps what is needed is some means of comparing the competing models for a given data set—something equivalent to a Mallows' C_p statistic (Mallows, 1973) or an Akaike information criterion (Akaike, 1973). In addition to this, perhaps some understanding of how each of the techniques behaves when applied in situations where one of the competing models prevails would be useful. For example, when there is no spatial pattern in the coefficients in reality, what kind of patterns does GWR generate? Similarly, if there is a strong geographical pattern, how likely is it that an RCM-based approach will detect it? Questions of this type can perhaps be investigated using simulation techniques.

In summary, this study has been useful in identifying at least three possible approaches to the identification of local variation in regression models. However, an important question that has been uncovered is that of choosing which method is the most appropriate. It is perhaps of extreme importance to quantitative geographers that locally varying models may be formulated, but it should also be noted that there is a broad spectrum of such models and that any advances in the specification and calibration of such models should be accompanied by parallel research in the areas of model choice and model validation in the context of competing explanations.

References

Aitkin, M. (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6, 251–262.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In: B. Petrov & F. Csaki, Eds, *2nd Symposium on Information Theory*. Budapest: Akademiai Kiado, pp. 267–281.

Brunsdon, C., Fotheringham, A. & Charlton, M. (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28, 281–289.

Mallows, C. (1973) Some comments on c_p. *Technometrics*, 15, 661–667.

Openshaw, S. (1994) Two exploratory space–time attribute pattern analysers relevant to GIS. In: A.S. Fotheringham & P.A. Rogerson, Eds, *Spatial Analysis and GIS*, London: Taylor & Francis, pp. 83–104.

Rees, P. (1995) Putting the census on the researcher's desk. In: S. Openshaw, Ed., *Census Users' Handbook*, Cambridge, UK: GeoInformation International, pp. 27–82.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Townsend, P., Phillimore, P. & Beattie, A. (1988) *Deprivation and Ill Health: Inequality and the North*. London: Croom Helm.