Spatial Regression and Geostatistical Models

In: Quantitative Geography

By: A. Stewart Fotheringham, Chris Brunsdon & Martin Charlton

Pub. Date: 2011
Access Date: September 8, 2016
Publishing Company: SAGE Publications Ltd
Print ISBN: 9780761959489
Online ISBN: 9781849209755
DOI: http://dx.doi.org/10.4135/9781849209755
Print pages: 162-183

Spatial Regression and Geostatistical Models

Introduction

One of the core techniques of the so-called quantitative revolution in geography was that of *linear regression*. Indeed, its use pre-dates the revolution — in physical applications by some three decades (Rose, 1936), and in human geography by at least one decade (McCarty, 1956). However, this was often rightly criticized for being an essentially aspatial method and therefore inadequate for modelling geographical processes. Gould (1970) states:

"there may be occasions when conventional inferential statistics can serve as research aids. These cases will probably be simple, non-spatial situations in which some educated guesses are capable of being investigated in inferential terms. Such occasions are likely to be rare."

Why was conventional regression seen as an inadequate tool for geographers? Some further insight can be gained by considering Hepple (1974):

"A glance at any set of maps of geographical phenomena will give a strong disposition against a classical urn-model generation of the pattern (and on which classical statistical inference is predicated)."

Some of the earliest attempts to provide more genuinely geographical approaches to data analysis were responses to these and similar criticisms. Examples include Ord (1975), Cliff and Ord (1973) and Hordijk (1974). These approaches may generally be described as *spatial regression models*. Although these are now relatively old ideas, they may be regarded as a turning point towards the modern view of spatial data analysis, and are therefore worthy of discussion here. Techniques of this sort have been applied to a wide range of study areas — for example, the spatial distribution of radon gas (Vincent and Gatrell, 1991), hydrology (Bras and Rodriguez-Irurbe, 1985), epidemiology (Cook and Pocock, 1983), geographical patterns in unemployment (Molho, 1995) and international trade (Aten, 1997). The range of dates for the preceding list of studies suggests that this kind of model continues to influence a broad span of disciplines.

To understand the shortcomings of ordinary linear regression and the theoretical contribution made by spatial regression models, consider the following practical example. Figure 7.1 shows the percentage of households that are owner occupied in wards in the county of Tyne and Wear, taken from the 1991 UK Census of Population. From this, it is clear that owner occupation tends to have higher rates on the periphery of the county, with the lowest rates being observed in the central area.

From the same source, a map of ward-based male unemployment rates in the same area is given in Figure 7.2. Note a general reversal of the pattern here — the *highest* rates of unemployment are in the centre. Suppose we went to investigate the linkage between unemployment and owner occupation. In particular, we would like to investigate whether male unemployment is a good predictor of owner occupation.

After some investigation of variable transforms, a scatterplot of the square root of male unemployment against owner occupation (Figure 7.3) suggests that, at least within the ranges in which the two variables commonly occur, a linear relationship exists between the two quantities. That is, after a square root transformation of the unemployment variable, one could model the relationship between the two quantities using linear regression. Therefore, a model of the form

Figure 7.1 Owner occupation in Tyne and Wear

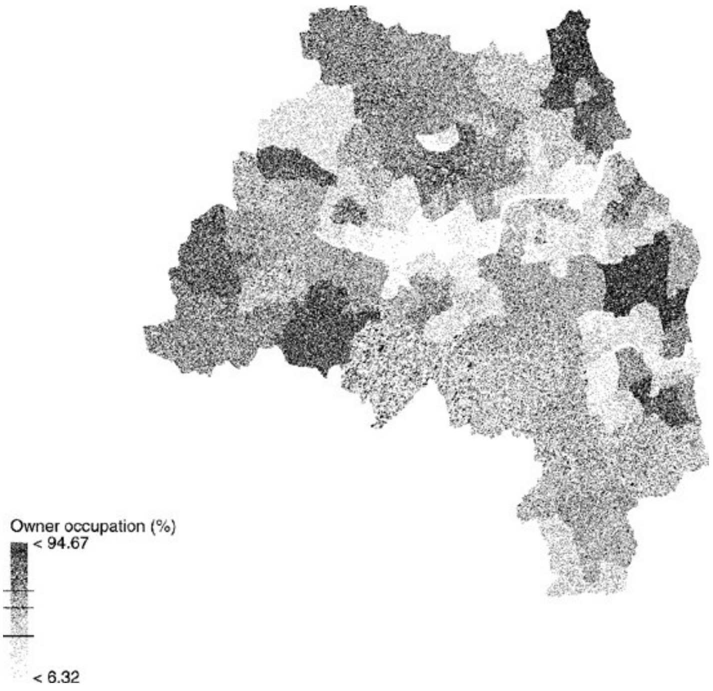
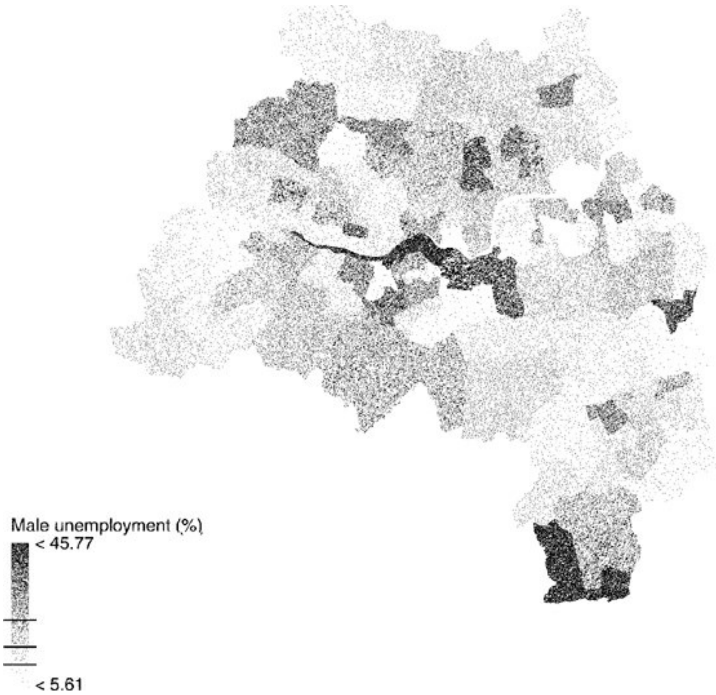


Figure 7.2 Male unemployment in Tyne and Wear



$$\text{Owner occupation} \sim \mathrm{N}(\beta_0 + \beta_1 \sqrt{\text{male unemployment}}, \sigma) \text{ (independently)} \tag{7.1}$$

is proposed. The values of β_0 and β_1 can be estimated using ordinary least squares regression. This gives the results shown in Table 7.1. Note that although it may seem strange that $\beta_0 > 100$ for the range of the explanatory variables encountered in the data set, the predicted levels of owner occupation lie within zero and 100% — the linear approximation seems reasonable *within the range of the data*. After fitting this model, the residuals are mapped. Residuals here are defined as the true value of owner occupation minus the fitted value. The resultant map is shown in Figure 7.4.

So what is wrong with proceeding in this way? The main problem is made apparent by the residual map. Note that the values of the residuals do not appear to vary randomly over space — high-valued residuals (darker areas on the map) seem to occur near to each other, as do low-valued residuals, and there are some quite sudden changes from the highest to the lowest values. This is at variance with the assumption in (7.1) that each observation is independent of the others, and in concordance with the comments of Gould (1970) and Hepple (1974) cited earlier in this section. If the independence assumption was true the error terms in the regression model could be modelled as uncorrelated random Gaussian variables, but here it seems that nearby terms tend to take the same sign. If one cannot assume independence in the error terms, then the least-squares calibration of β_0 and β_1 is no longer justified. The remainder of this chapter is concerned with the methods of addressing this. These fall into three broad categories, as shown in Table 7.2. The term *usual* data type is used here as it is possible to apply point-based techniques to area data, for example by replacing each zone by its centroid, and also to apply area-based techniques to point data, for example by computing Voronoi tessellations to a set of points. Note also that non-independent error terms may be modelled either by considering the correlation between error terms, or by modelling a spatial trend in the error terms, rather than assuming the expected value of the error is everywhere zero.

Figure 7.3 The square root of male unemployment vs. owner occupation in Tyne and Wear

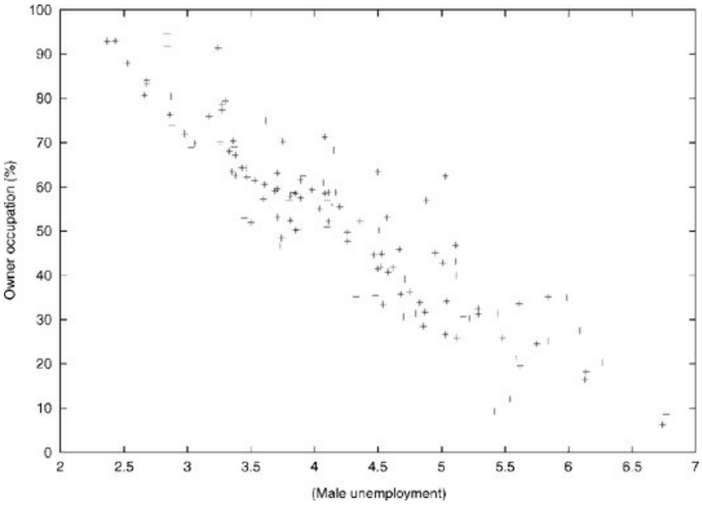
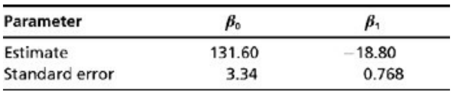


Table 7.1 Results of standard regression model for owner occupation data



Parameter	β_0	β_1
Estimate	131.60	−18.80
Standard error	3.34	0.768

Figure 7.4 Map of residuals for regression model (7.1) Table 7.2 Techniques considered in this chapter

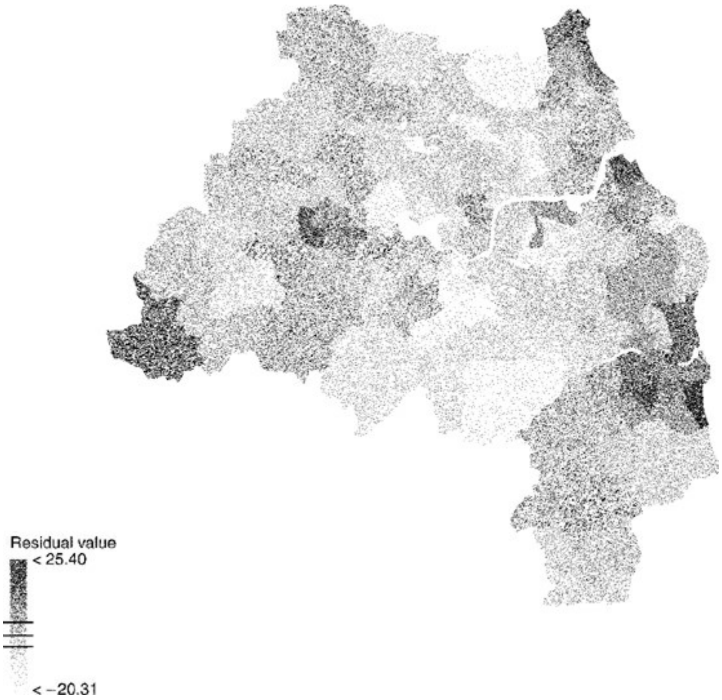
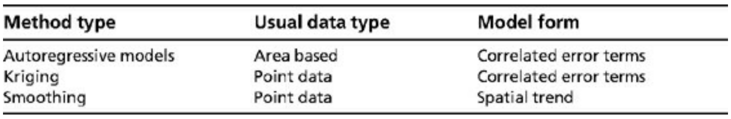


Table 7.2 Techniques considered in this chapter



Method type	Usual data type	Model form
Autoregressive models	Area based	Correlated error terms
Kriging	Point data	Correlated error terms
Smoothing	Point data	Spatial trend

Autoregressive models

In more general terms, the model (7.1) can be stated in terms of *n* observations of a dependent variable, contained in the vector y, and *n* observations of *m* dependent variables contained in the *n* by *m* matrix X. In this case, the ordinary regression model can be expressed in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad (7.2)$$

where *β* is a vector of regression coefficients, and ε is a vector of random errors. Here, the elements of ε are independently distributed about a mean of zero. Thus, for the ordinary model one may state

$$\boldsymbol{\epsilon} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2\mathbf{I}) \qquad (7.3)$$

where I is the identity matrix and 0 is the zero vector, and MVN(M, S) denotes a multivariate normal distribution with mean vector M and variance-covariance matrix S. The problem with this model is the assumption that the variance of ε is $\sigma^2\mathbf{I}$. A more general approach would allow for a broader class of variance-covariance matrices, which would allow for non-independence in the residuals. Most general in this sense would be to assume a model of the form

$$\epsilon \sim \mathrm{MVN}(\mathbf{0}, \mathbf{C}) \qquad (7.4)$$

where C is any valid variance-covariance matrix. This would resolve the general problem of the assumption of independence, but creates two new problems. First, Figure 7.4 implies that there is a spatial pattern in the dependence of residuals. The general covariance matrix C does not specifically reflect this. Secondly, although it is easy to calibrate β if **C** is known, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{C}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\mathbf{y}$ (Mardia et al., 1979), normally this is not the case and so both **C** and β must be calibrated from the sample data.

Spatially autoregressive models

The problem of reflecting *spatial* dependency in the model will be considered first. One way to think of this is to postulate that for zone-based data, the random components of the model in adjacent zones are likely to be linked. For example, if considering house owner occupation rates, it is quite likely that the rate in a zone will mimic to some extent the rates of surrounding zones — especially if the zones are fairly geographically compact — say, one or two neighbourhood blocks. One way to account for this in a regression model would be to use, as an extra explanatory variable, the mean of the dependent variable for adjacent zones. Thus, in the house ownership example, for each zone we would compute the mean of the rates for adjacent zones and use this as an explanatory variable. Returning to the algebraic notation, if y is the dependent variable, we need to transform this into a vector of adjacent means. This is a linear operation, and can be written as **Wy**, where w_{iJ} = 0 if zones *i* and *j* are not adjacent, and w_{ij} = 1/a$_i$, if *i* and *j* are adjacent, and a, is the number of zones adjacent to zone *i*. In fact **W** is just the adjacency matrix for the zones, with rows rescaled to sum to one. Zones are not assumed to be adjacent to themselves, so w$_{ij}$ = 0 for all *i*. If the adjacent-mean variable is added to the regression model, the modified model becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{y} + \epsilon \qquad (7.5)$$

where ρ is the regression coefficient for the adjacent-mean variable. For obvious reasons, this is often called an *autoregressive* model. Subtracting ρ**Wy** from both sides of (7.5) gives

$$\mathbf{y} - \rho\mathbf{W}\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \qquad (7.6)$$

Factoring the left hand side gives

$$(\mathbf{I} - \rho\mathbf{W})\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \qquad (7.7)$$

and, assuming (**I** — ρ**W**) is invertible, pre-multiplying both sides by this expression gives

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\epsilon \qquad (7.8)$$

Thus, after transforming the **X** matrix, the model is of a similar form to (7.2), but with an error term that is a linear transformation of the original, independent vector £. In fact the new error term has variance-covariance matrix

$$\mathbf{C} = \sigma^2[(\mathbf{I} - \rho\mathbf{W})^{-1}]'(\mathbf{I} - \rho\mathbf{W})^{-1} \qquad (7.9)$$

If C is defined as a function of ρ as above, then it will reflect a spatial structure in error dependencies. Thus, once ρ is known, **C** can be computed and *β* can be estimated. The problem is now one of estimating ρ from the sample data.

This problem can be divided into two parts — first finding a valid range for ρ and secondly, finding an *optimum ρ* in terms of a maximum likelihood criterion. The first part is important because the variance-covariance matrix **C** must be valid — not all *n* by *n* matrices of real numbers meet this condition. The validity condition may be stated as below:

$$\sum_i \sum_j \alpha_i \alpha_j c_{ij} \geqslant 0 \qquad (7.10)$$

where the α_i, and α_j, are arbitrary real numbers. This is basically a condition that all linear combinations of the elements of y have a positive (or zero) variance. It can be shown (see for example Griffith, 1988) that for the autoregressive model this condition is satisfied if and only if| ρ| ≤ 1.

The second part of the problem can be approached by considering the expression for the likelihood of **y** given values of β and *ρ*. This is monotonically equivalent to

$$\sum_i \left(1 - \rho \sum_j w_{ij} y_j - \sum_j \beta_j x_{ij}\right)^2 \qquad (7.11)$$

Finding β and *ρ* to minimize this expression is equivalent to finding maximum likelihood estimates. The optimization process can be broken down into two parts:

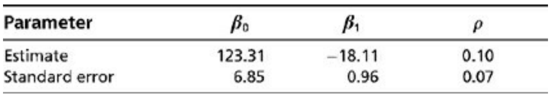
1
Find the optimum p.
2
Estimate β by substituting the *estimate* of *ρ* into the standard equation.

A number of calculation ‘tricks’ may be applied to make the first stage more computationally efficient; see Ord (1975) for a discussion of this. An alternative, but sometimes much less reliable, approach is to estimate β and *ρ* using least squares, although this is not the true maximum likelihood in this case. The consequence of this is that the least-squares estimates are often biased.

Applying the maximum likelihood methods to the owner occupation data gives the results of

Table 7.3. Note that there are a few changes from the results for the standard regression. First, the value for β_0 is a little lower and the value for β_1 is a little higher, suggesting a smaller negative slope on the linear regression. The estimated value for ρ suggests that some small degree of autocorrelation occurs, implying that neighbouring rates of owner occupation have some effect on the observed rate for a given zone. Finally, note that the standard errors for the estimates are a little larger than in the simple model. This is quite typical behaviour when extra parameters are introduced into models. In this case, the single parameter ρ has been introduced.

Table 7.3 Results of autoregressive model for owner occupation data



Parameter	β_0	β_1	ρ
Estimate	123.31	−18.11	0.10
Standard error	6.85	0.96	0.07

Spatial moving average models

A variation on the autoregressive model is the *moving average* model. In this model, it is not the dependent variable that is considered as autoregressive, but the *error term*. The regression model is thus

$$\begin{aligned} &\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \\ \text{where} \quad &\mathbf{u} = \rho\mathbf{W}\mathbf{u} + \epsilon \end{aligned} \qquad (7.12)$$

The inspiration for this model comes from time series analysis, where both autoregressive and moving average models may be used to model the behaviour of a variable observed at regular time intervals — see for example Kendall and Ord (1973). In many practical cases it is difficult to see whether this or an autoregressive approach gives a more accurate reflection of the spatial process under examination. One theoretical difference between the two models is that whereas in the autoregressive model it is typical that *all* error terms are correlated (albeit in a manner which reduces as zones become further apart), in the moving average model error terms are only correlated to their immediate neighbours, as specified in **W**. In more formal terms, for moving average processes, **C** is a sparse matrix if **W** is based on contiguities between zones.

Calibrating moving average models is somewhat harder than calibrating autoregressive models. The simplest approach is to note that Equation (7.12) may be rearranged (in a similar manner to Equations (7.5)-(7.8)) to give

$$(\mathbf{I} - \rho\mathbf{W})\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})\mathbf{X}\boldsymbol{\beta} + \epsilon \qquad (7.13)$$

Thus, if ρ were known, β could be estimated by an ordinary least-squares regression of a linear transform of the **y** variables against the same transform of the **X** variables. Of course, ρ is unknown. However, an estimate of ρ could be obtained by calibrating an autoregressive model on the residuals of the ordinary least-squares model, that is finding the value of *ρ* minimizing ε ′ε in Equation (7.12). Once this is done, and a new β estimate is obtained, one can re-estimate the residuals, and then iterate the entire process. Repeating this can be shown to lead to a converging sequence of β and *ρ* estimates. Thus, the entire procedure may be set out as below:

1
Obtain an initial estimate of β using ordinary least squares:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

2
Use the result from step 1 to estimate a set of residuals. Obtain an initial estimate of *ρ* from these, using the autoregressive model calibration technique. Call this $\hat{\rho}$.
3
Use the current estimate of *ρ* to update the β estimate, given $(\mathbf{I} - \hat{\rho}\mathbf{W})\mathbf{y} = (\mathbf{I} - \hat{\rho}\mathbf{W})\mathbf{X} + \epsilon$..
4
Use the updated β estimate to obtain an updated set of residual estimates, and then an updated *ρ* estimate. Return to step 3 and iterate until $\hat{\rho}$ and $\hat{\boldsymbol{\beta}}$ converge.

Applying this procedure to the owner occupation data gives the results set out in Table 7.4. Here, the value of *ρ* is much larger than before, suggesting a strong degree of autocorrelation in the error term. Also, although the standard errors are still larger than those in the standard model (Table 7.1) they are considerably smaller than those seen in Table 7.3. It seems there is a stronger case for a spatial effect in the error term than there is for an autoregressive effect. Typically, this suggests that deviations from a global model take place at a very local scale — a suggestion which is also supported by the residual map.

Kriging

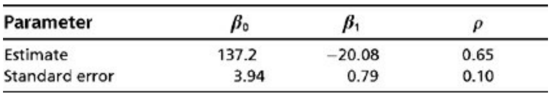
The statistical technique

The above section deals with approaches to spatial modelling for zone-based data, where *proximity* is represented by **W**, the adjacency matrix. However, with point-based data, the concept of adjacency is not so implicitly defined. A more natural approach here might be to consider **D**, a matrix of distances between the points, so that d_{Ij} is the distance between points *i* and j. This methodology plays a core role in the field of *geostatistics*. Suppose there is a set of *n* points with a distance matrix as defined above, also with a vector of attached dependent variables **y** and a set of attached independent variables, **X**. As before, it is intended to calibrate a regression model **y = βX + ε**, where the vector ε has distribution MVN(**0, C**), and it is required that the elements of **C**, the variance-covariance matrix of the error terms, reflect the spatial structure of the data.

A possible way of achieving this is to let c_{ij} be some function of d_{ij}, say $c_{ij} = f(d_{ij})$. This would mean that the covariance of a pair of error terms would depend on the distance between them. Typically, one would expect f to be a decreasing function, so that the linkage between error terms decreased as the distance between them increased. For this, the positive definiteness condition on the matrix C discussed earlier becomes important once again. Recall that this required

$$\sum_i \sum_j \alpha_i \alpha_j c_{ij} \geqslant 0$$

Table 7.4 Results of moving average model for owner occupation data



Parameter	β_0	β_1	ρ
Estimate	137.2	−20.08	0.65
Standard error	3.94	0.79	0.10

for all α_i, and α_j. In the current situation, this requires that

$$\sum_i \sum_j \alpha_i \alpha_j f(d_{ij}) \geqslant 0$$

Unfortunately, this means that one cannot choose an arbitrary function *f* as most do not satisfy the condition. There are, however, some functional forms which do meet this requirement. A number of possible candidates include the exponential form

$$c_{ij} = \sigma^2 \exp(-d_{ij}/h) \qquad (7.14)$$

or the power form

$$c_{ij} = \sigma^2 \exp(-d_{ij}^2/h)^2 \qquad (7.15)$$

or the spherical form

$$c_{ij} = \begin{cases} \sigma^2\left(1 - \dfrac{3d_{ij}}{h} + \dfrac{d_{ij}^3}{h^3}\right) & \text{if } d_{ij} < h \\ 0 & \text{if } d_{ij} \geqslant h \end{cases} \qquad (7.16)$$

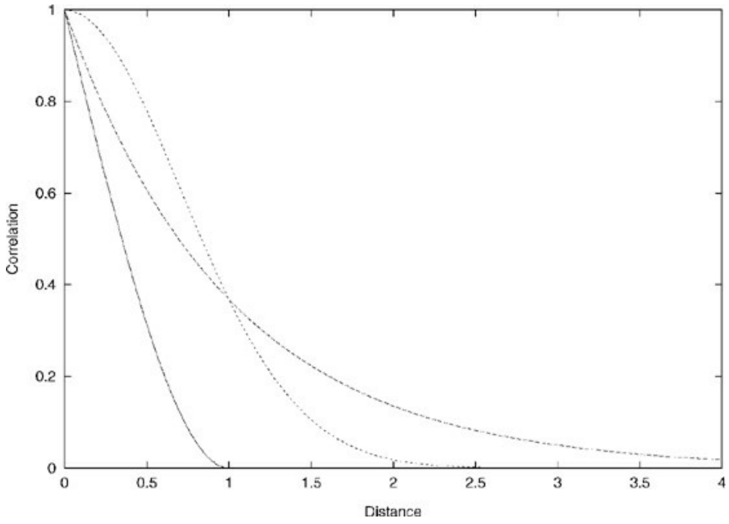
In each case, *h* acts in a similar manner to the bandwidth in kernel density estimation (see Chapters 4 and 6), and determines the distance around a given observation over which other observations are likely to be dependent: the parameter *h* is usually expressed in the same distance units as the distances between the points, d_{tj}. In each of the above equations there are two unknown parameters, σ and *h*. In simple terms *a* describes *how variable* the error term is, and *h* describes *how spatially influential* it is. Note also that by dividing any of (7.14)-(7.16) by σ^2 an expression for the *correlation* between ε_i, and ε_j is obtained. For example, for

Equation (7.14) we have

$$\mathrm{corr}(\epsilon_i, \epsilon_j) = \exp(-d_{ij}^2/h^2) \qquad (7.17)$$

The three different functional forms for the relationship between correlation and distance are shown in Figure 7.5. For each curve, $h = 1$. Note that although in general wider bandwidths imply increasing radii of non-trivial correlation, the same value of bandwidth for different functions does not imply exactly the same radius of influence for both functions. Note that in all of these functions, as d_{ij} approaches zero, the covariance approaches σ^2. However, in some cases this may be inappropriate — there will be some 'extra' variability at the point of observation perhaps due to measurement or sampling error. Here the limiting covariance as d_{ij} approaches zero is some quantity τ^2, say, which is not equal to σ^2· In this case, the covariance function undergoes a jump of size $\sigma^2 — \tau^2$ at the point $d_{ij} = 0$. This is sometimes referred to as the *nugget effect*. For example, Equation (7.14) can be extended to the form below:

Figure 7.5 Graphs of correlation functions: the light dotted line is for model (7.14) the heavier dotted line is for (7.15) and the solid line is for (7.16); in each case h = 1



$$c_{ij} = \begin{cases} \tau^2 \exp(-d_i j/h) & \text{if } d_{ij} > 0 \\ \sigma^2 & \text{if } d_{ij} = 0 \end{cases} \qquad (7.18)$$

Graphically, this relationship would appear much the same as the curve for Equation (7.14) shown in Figure 7.5, but with a single discontinuity at $d_{ij} = 0$.

The technique of *kriging* (Krige, 1966) involves finding estimates for the regression coefficients β as well as for *σ*, *τ* and *h*. This problem is not as simple as for ordinary least-squares regression, since one needs to know *σ*, *τ* and *h* in order to estimate β, but also one needs to know β in order to estimate *σ*, *τ* and h_i A compromise is to start with a *guess* at β, use this to estimate *σ*, *τ* and *h* and then to re-estimate β in the light of this. A plausible initial guess for β would be the ordinary least-squares estimate.

From this starting point, one can obtain a set of residuals from the observed $y_i$variables. Call these e_i,. These may be used as estimates for the true residuals ε From these, we need to estimate the parameters *σ*, *τ* and *h*, by considering the spatial relationship between the residual pairs e_i, and e_j, for points *i* and *j*, and the distance between the points d_{ij}. A useful way of measuring the difference between e_i and e_j is to consider the squared difference $(e_i — e_j)^2$. Consider the expected value of this expression, $E[(e_i — e_j)^2]$. Expanding, we have

$$E[(e_i - e_j)^2] = E(e_i^2) - 2E(e_ie_j) + E(e_j^2) \qquad (7.19)$$

but, given that e_i, and e_j, have expected values of zero, we can state that $E(e_i^2) = E(e_j^2) = \sigma^2$ and $E(e_ie_j) = \text{cov}(e_i, e_j) = c_{ij}$, so that

$$E[(e_i - e_j)^2] = 2\sigma^2 - 2c_{ij} \qquad (7.20)$$

Thus, the expected value of $(e_i — e_j)^2$ is very simply related to the covariance function, and can be expressed as a function of d_{ij}:

$$E[(e_i - e_j)^2] = 2\sigma^2 - 2f(d_{ij}) = g(d_{ij}) \qquad (7.21)$$

and $g(d_{ij})$ can therefore be expressed as a function with parameters *σ*, *τ* and say, *h*. For example, if the covariance model (7.14) is adopted, then $g(d_{ij})$ = σ^2[2–2 exp(-d_{ij}/*h*). Adding a nugget effect, we have

$$g(d_{ij}) = \begin{cases} (\sigma^2 - \tau^2)2 - [2\exp(-d_ij/h)] + \tau^2 & \text{if } d_{ij} > 0 \\ \sigma^2 & \text{if } d_{ij} = 0 \end{cases} \qquad (7.22)$$

Thus, by fitting a non-linear curve to $(e_i — e_j)^2$ as a function of d_{ij}, it is possible to estimate *h, σ* and *τ*. Sometimes the function *g* is divided by 2, since this results in the right hand side of the above equation losing an ‘untidy’ factor of 2. Noting that $E[(e_i — e_j)^2]$ = var($e_i — e_j$), this halved version of *g* is sometimes called the *semi-variance*, and a graph of *g* is called a *semi-variogram*. This is a particularly useful technique if the nugget effect is present, since *σ* can be more readily estimated as the asymptotic value of the semi-variogram as d_{ij} tends to infinity.

Calibrating the semi-variogram is itself a complex task. Even in a situation where the ε_i are independent, the e_i are not (Dobson, 1990), and so one cannot expect the $(e_i — e_j)^2$ to be. However, in general kriging practice this problem is ignored, and *h* and *σ* are usually estimated using non-linear least-squares approaches — that is, finding *h* and *σ* to minimize

$$\sum_{i,j}[(e_i - e_j)^2 - g(d_{ij}|h, \sigma, \tau)]^2 \qquad (7.23)$$

where $g(d_{ij}|h, \sigma, \tau)$ denotes the dependence of *g* on the three parameters. Having found reasonable approximations for *h* and *σ*, one can then obtain estimates for the c_{ij}, and proceed to recalibrate the model for β. If, as before, we denote the entire covariance matrix by **C**, we then have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'CX)^{-1}\mathbf{X}'\mathbf{C}\mathbf{y} \qquad (7.24)$$

as discussed in earlier sections when estimating β in situations where there is correlation between error terms. This rather lengthy procedure can be summarized as follows:

1

Estimate β using ordinary least squares.

2

Estimate residuals from this β estimate.

3

Calibrate the semi-variogram from the residuals.

4

Use the calibration of the semi-variogram to estimate **C**.

5

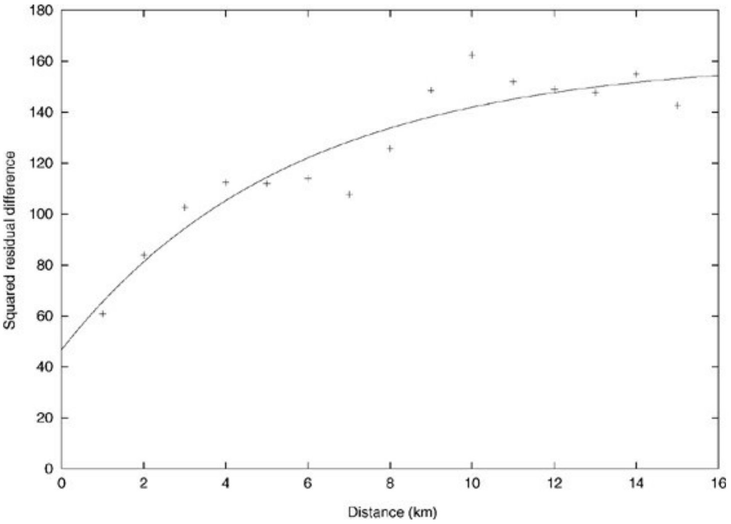
Re-estimate β using $(\mathbf{X}'\,CX)\text{-}^{1}\mathbf{X'Cy}$.

A worked example

The technique of kriging is best illustrated with a worked example. Here the data described earlier will be used. Although this is zone rather than point data, one can attribute the data to ward centroids. In this way it is possible to compare the results of this analysis with those of previous sections in this chapter. As with previous analyses, the explanatory variable here will be the *square root* of unemployment, as this appears to have a linear relationship with owner occupation.

To recap, the ordinary least-squares estimates for the elements of β were β_0 = 31.6 and β_1 >= −18.8. From this, residuals were computed, as shown earlier in Figure 7.1. This supplies us with a set of residuals which may be used to estimate *h* and *a*. In this case, there are 120 wards, and therefore 120 centroids. Thus, there are 7140 d_{ij} values to consider, together with the same number of values of $(e_i — e_j)^2$. For a non-linear regression problem this may be computationally intense. Therefore, another common computational shortcut used in kriging is to 'bin' the d_y categories into a series of distance intervals, and consider the average $(e_i — e_j)^2$ values for each category as a function of the central distance values for that category. In some cases, the bin categories might overlap (Bailey and Gatrell, 1995). Here bins of width 2 km centred on the d_{ij} values 1 km, 2 km, …, 16 km were used, and a semi-variogram curve of the form (7.22) was fitted to the observed average $(e_i — e_j)^2$ values. The results of this exercise are shown in Figure 7.6. Here, the fitted parameter values are h = 5.6 km, $\sigma^{2\;=}$ 160.8 and τ^2 = 46.75. From

Figure 7.6 Results of fitting a theoretical semi-variogram of the form in model (7.22)

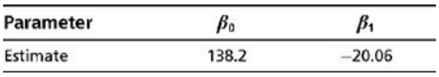


these, we can construct the **C** matrix, and re-estimate β_0 and β_1. The re-estimates are shown in Table 7.5. This gives a very similar value of β_0 and β_1, to the spatial moving average model.

Trend surfaces from kriging residuals

One very useful product of the kriging approach to regression is that it is possible to predict the value of residuals at points other than those at which the observations are taken. Suppose a new point is added to the data set, at a location *(u, v)*. Without any other information, the mean value of the error term expected at this point is zero. However, if there is spatial autocorrelation amongst the error terms, and the values of some errors *near* to the new point are known, then this situation changes. For example, if many nearby error terms are positive, one might reasonably expect the error at *(u, v)* to be positive also. Essentially, we are now considering the *conditional* probability distribution of the error at (u, *v) given* the errors at the

Table 7 5 Results of kriging model for owner occupation data



Parameter	β_0	β_1
Estimate	138.2	–20.06

other locations. Using this distribution, it is possible to predict the likely error term at *(u, v)* and therefore apply a correction to the ordinary least-squares predictor of *y* at *(u, v)* given the predictor variables.

In practice we do not know the error terms at each point, but we do know the residuals from fitting the regression model after kriging. The latter may be used as an approximation for the former. Next we must consider how one can predict the error at *(u, v)* given a vector of *n* errors {ε_i}, or residuals {ε_i}. One way of addressing this is to consider the *expected mean square error* in predicting the error term. If we denote the true value of the error term at *(u, v)* by *e(u, v)*, and the estimated value of this by $\hat{\epsilon}(u, v)$, then we are attempting to minimize

$$\mathrm{E}\{[\epsilon(u, v) - \hat{\epsilon}(u, v)]^2\} \qquad (7.25)$$

In particular, suppose we consider estimates that are linear combinations of the e_i so that

$$\hat{\epsilon}(u, v) = \sum_i \gamma_i e_i \qquad (7.26)$$

Then, it can be shown (Bailey and Gatrell, 1995), that if *γ* is the vector of the *γ*, in the above expression then (7.25) is minimized when

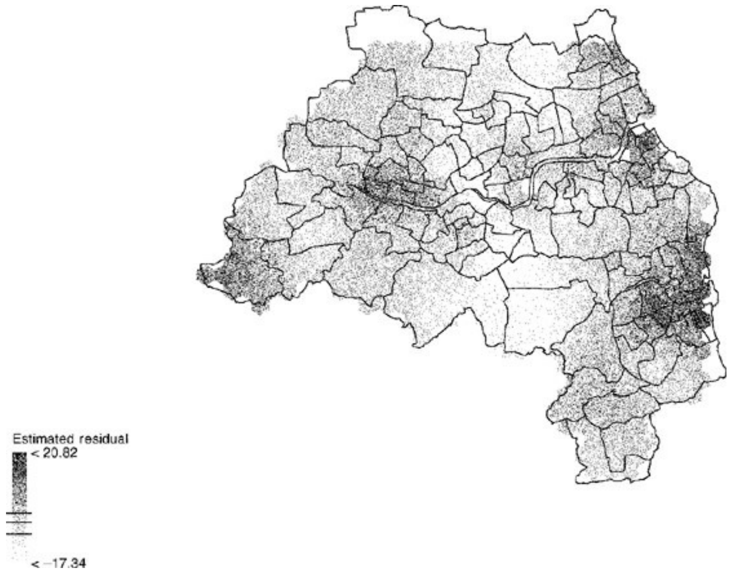
$$\gamma = \mathbf{C}^{-1}\mathbf{c}(u, v) \qquad (7.27)$$

where C is the *n* by *n* covariance matrix between the errors at the *n* observed points, and c(u, *v)* is the *n* by 1 vector of covariances between the error at *(u, v)* and the error at the *n* observed points. Recall that in the kriging procedure we calibrated a functional relationship between the covariance of a pair of error terms and the distance between the points at which the error terms occurred. This not only allows an estimate of C to be made as before, but also allows an estimate of c(u, *v)* to be made, based on the distances between *(u, v)* and the *n* observed point locations. Thus, if e is a vector of the residuals {e_i,}, then we have

$$\hat{\epsilon}(u, v) = (\mathbf{C}^{-1}\mathbf{c}(u, v))'\mathbf{e} \qquad (7.28)$$

Thus, it is possible to predict the error term (or at least compute its conditional mean value) at any point *(u, v)* in the study area. Thus, it is possible to consider spatial trends in the error term, and by mapping $\hat{\epsilon}(u, v)$ identify regions where the kriging regression model is likely to over- or underpredict. For the home ownership application discussed above, a map of $\hat{\epsilon}(u, v)$ is shown in Figure 7.7. One notable

Figure 7.7 Map of $\hat{\epsilon}(u, v)$ for owner occupation example in Section 7.3.2



feature here is a region of positive predicted residuals in the south-eastern part of the map. This region roughly corresponds to the city of Sunderland. A number of other prominent features can also be identified, for example around the coastal community of South Shields, just to the south of the mouth of the River Tyne.

Semi-parametric smoothing approaches

A final method which models spatial deviations from a simple regression model is the *semi-parametric* model (Hastie and Tibshirani, 1990). This differs from the other models in that it does not assume that the error terms are correlated. Here the model takes the form

$$y_i = f(u_i, v_i) + \sum_j \beta_j x_{ij} + \epsilon_i \qquad (7.29)$$

The model is termed semi-parametric since one makes no assumptions about the functional form *f*. In fact *f* is used to model spatial deviations from the simple, non-spatial regression model. A plot of *f*(*u, v*) as a surface in the study area serves a similar purpose to that of Figure 7.6, in that it indicates spatial trends in over- or underprediction of the simple regression model.

To calibrate a model of this kind, a very different approach is required. First, suppose that the error terms $\{\varepsilon_i\}$ were known. One could estimate *f* by applying a spatial smoothing to these quantities. This is achieved at point *(u, v)* by computing a weighted average of the ε_i where the greatest weight is assigned to observations closest to *(u, v)*. For example, the Gaussian weighting scheme could be defined by

$$w_i(u, v) = k \exp[-d_i^2(u, v)/h^2] \qquad (7.30)$$

where W_i(u, *v)* is the weight applied to the ith residual to estimate *f(u, v)*, and *d,(u, v)* is the distance between the ith observation and the point *(u, v), h* is a bandwidth controlling the degree of smoothing as discussed in Chapters 2, 4, 5 and 6, and it is a normalizing constant to ensure that $\sum_i w_i(u, v) = 1$. Using these weights, we can estimate *f* at *(u, v)* using

$$\hat{f}(u, v) = \sum_i w_i(u, v)\epsilon_i \qquad (7.31)$$

It is worth noting the similarity between this approach and the geographically weighted regression (GWR) technique described in Chapter 5 (Brunsdon et al., 1996). As with kriging, we do not have observed values of ε_j However, if β were known we could use the residual terms $\{e_j\}$ as estimates, giving a revised estimate

$$\hat{f}(u, v) = \sum_i w_i(u, v)e_i \qquad (7.32)$$

The problem here is that we do not know β. However, rearranging (7.29) and replacing *f* by its estimate $\hat{f}$ we have

$$y_i - \hat{f}(u_i, v_i) = \sum_j \beta_j x_{ij} + \epsilon_i \qquad (7.33)$$

Thus, β could be estimated by regressing $y_i - \hat{f}(u_i, v_i)$ on the explanatory variables. Obviously, this estimate, say $\hat{\boldsymbol{\beta}}$ could then be used in Equation (7.32). In matrix form, the relationship between $\hat{f}$ and $\hat{\boldsymbol{\beta}}$ can be expressed by the simultaneous equations

$$\begin{aligned}\hat{\mathbf{f}} &= \mathbf{W}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \hat{\mathbf{f}})\end{aligned} \qquad (7.34)$$

where **W** is the matrix whose ijth element is $w_i(u_j,\ Vj)$ and $\hat{\mathbf{f}}$ is a column vector whose ith element is $\hat{f}(u_i, v_i)$. After a lengthy sequence of algebraic manipulation it is possible to solve explicitly for $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'(\mathbf{I} - \mathbf{W})\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{W})\mathbf{y} \qquad (7.35)$$

and

$$\hat{\mathbf{f}} = \mathbf{W}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad (7.36)$$

so that it is possible to calibrate the entire model using explicit formulae. It is also worth noting the similarity between Equation (7.35) for the semi-parametric approach and Equation (7.24) for kriging. The smoothing approach and the kriging approach can give very similar estimates for β. In fact if **I** — W = **C** then the two approaches give identical estimates.

One final issue to be addressed is the choice of *h*, the smoothing bandwidth. When *h* is very small, the spatial smoothing applied to the residuals will be extremely ‘spikey’ and will follow the random variations in the error terms too closely. However, when *h* is very large then oversmoothing may occur, so that genuine features in *f(u, v)* may be ‘smoothed out’ and not detected in $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$. Also, selection of *h* is not akin to estimating a model parameter. The parameters one is trying to estimate are the function *f* and the vector β. The most appropriate *h* for one particular data set will not be the same for another — factors such as the density of the observation points as well as the shape of the unobserved function *f* will all affect the choice of *h*. One approach to choosing *h* is to consider the expected mean square error for the prediction of *y*, as was done when predicting residuals in kriging. In the semi-parametric case, there is no simple algebraic solution giving *h* which minimizes this quantity. Instead, one has to rely on estimating this quantity from the data, and find *h* which minimizes this. There are a number of ways of estimating the expected mean square error from the data — one of these is to compute the *generalized cross-validation score* or GCV score (Hastie and Tibshirani, 1990). This is defined to be

$$\mathrm{GCV} = \frac{1}{n}\sum_i \left\{ \frac{y_i - \hat{y}_i}{1 - \mathrm{tr}(\mathbf{S})/n} \right\}^2 \qquad (7.37)$$

where $\hat{y}_i$ is the fitted value for y_i, tr(X) denotes the sum of the leading diagonal (or *trace)* of the matrix **X**, and **S** is a matrix such that $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, where $\hat{\mathbf{y}}$ is a column vector of predicted *y* values. For the semi-parametric model seen here, it can be shown that

$$\mathbf{S} = \mathbf{X}(\mathbf{X}'(\mathbf{I}-\mathbf{W})\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{W}) + \mathbf{I} - \mathbf{W} \qquad (7.38)$$

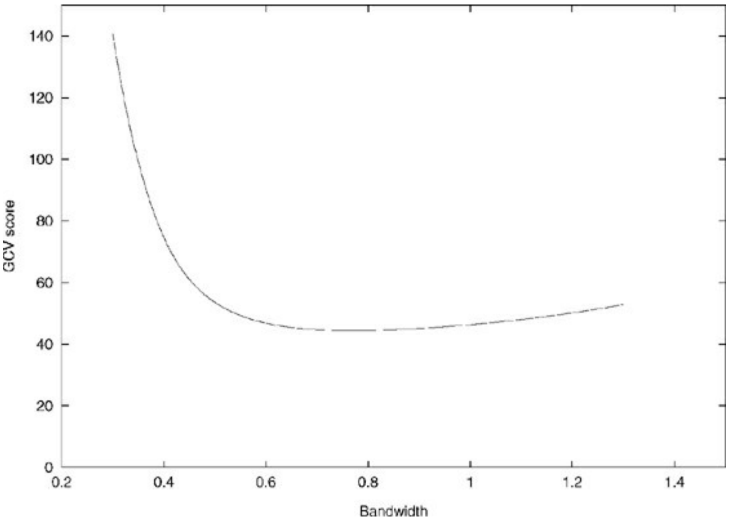
As *h* changes, then in turn S and GCV alter. To choose *h*, one has to plot its value against the associated GCV estimate, and find the minimum point on this graph. Note that with a large number of observations, computation of S becomes a problem, since it consists of n^2 elements. Typically, sparse matrix manipulation algorithms are called for here.

As an example, consider the home ownership data once again. In this case, the model fitted takes the form

$$\text{Owner occupation} = \beta_1\sqrt{\text{unemployment}} + f(u, v) + \epsilon \qquad (7.39)$$

Note that there is no β_0 here, since *f(u, v)* takes an arbitrary form, and so could contain an additive constant such as β_0. For these data, *h* is plotted against GCV in Figure 7.8. This graph shows that the lowest prediction mean square error occurs when *h* is around 0.75 km, although the curve is quite flat around this region, and values in the range 0.75–1.0 km will give more or less similar results in terms of

Figure 7.8 Plot of GCV vs. h for the home ownership model

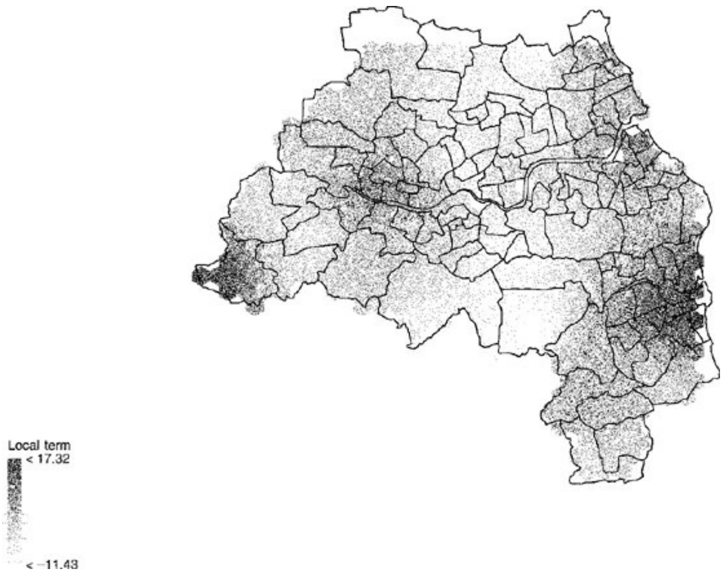


predicting the *y* values. When the optimal value of *h* is used we obtain an estimate for β_1 of −18.26. The corresponding estimate for β_1 is shown in Figure 7.9. It is worth noting the similarity between this map and that of Figure 7.6. Also of interest is the fact that this model could be further extended, by allowing *all* of the coefficients to vary over space. In doing this we would arrive at the GWR approach. Finally, it is possible to extend this model even further by allowing for autocorrelation effects in the model *as well as* spatial variation — and furthermore to allow the value of *p* to vary geographically (Brunsdon et al., 1998a)¡

Conclusions

In this chapter, a number of ways of approaching regression which allow for spatial dependency amongst the observed points are discussed. The first two, the auto-regressive and moving average approaches, are essentially zone based. The covariance between error terms is defined in terms of the adjacencies between zones, or the distances between zone centroids. Clearly, this depends very heavily on the choice of areal units. In instances where such information is available, it may be helpful to apply autoregressive or moving average models to the same data aggregated in more than one way. This may give some idea of how sensitive the models are to the areal units used. In fact it is possible to obtain a range of results

Figure 7.9 Map of $\hat{f}(u_i, v_i)$ for home ownership model

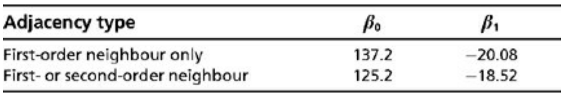


with a single data set, by defining adjacency between zones in different ways. This is illustrated in Table 7.6, where the owner occupation moving average model is calibrated first with the adjacency matrix used earlier in the chapter, and secondly with adjacency being defined as a first- or second-order neighbour. This could perhaps be regarded as a new manifestation of the modifiable areal unit problem (Openshaw, 1984), where not only the *scale* and *location* of zones but also the *topology* can influence the results of a statistical analysis. It is worth noting that although it is possible to compute confidence intervals for the model parameters in these models, these do not allow for any uncertainty in the definition of adjacency. In real terms, these confidence intervals are almost certain to be too small.

One way in which kriging differs from the latter techniques is that some attempt is made to calibrate the *extent* of the error term correlation from the observed data. Rather than assuming that dependencies are in terms of immediate zonal neighbours, a correlation function is calibrated based on the distance between points. Once this has been done the regression model is then calibrated using this model of error term correlation. However, as before, this calibration assumes the correlation function to be known exactly, rather than being an estimated curve from residual data. This results in a similar problem to that with autoregressive and moving average models: although confidence intervals can be computed for the regression coefficients, these will most likely be too optimistic since in actuality there is some uncertainty about the correlation function. This is addressed in Diggle et al. (1998), however, using a Bayesian framework (Besag and Green, 1993).

It is interesting to note the similarity between the semi-parametric model and the kriging model. Although these two approaches assume that different processes generated the observed data, they give quite similar predictions. Indeed, by an appropriate choice of correlation function for kriging, and smoothing function for the semi-parametric approach, one can obtain identical predictions. This demonstrates a common problem with spatial regression: it can sometimes be very difficult to infer which underlying process generated the data. However, one positive point is that even if one chooses the wrong process to model, the consequences in terms of prediction error will not be great. In such situations it is perhaps reasonable to adopt the computationally simplest approach, which in this case is perhaps the semi-parametric approach. However, despite the difficulties involved in any of these methods, to the geographer all of these approaches offer an improvement on the simple regression technique, since they all take into account the spatial nature of the process being analysed.

Table 7.6 Variability of a spatial moving average model as the definition of adjacency is altered



Adjacency type	β_0	β_1
First-order neighbour only	137.2	−20.08
First- or second-order neighbour	125.2	−18.52

http://dx.doi.org/10.4135/9781849209755.n7