SOME NOTES ON PARAMETRIC SIGNIFICANCE TESTS FOR GEOGRAPHICALLY WEIGHTED REGRESSION

Chris Brunsdon

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

A. Stewart Fotheringham and Martin Charlton

Department of Geography, University of Newcastle Upon Tyne, Newcastle Upon Tyne, NE1 7RU, U.K. E-mail: stewart.fotheringham@ncl.ac.uk and martin.charlton@ncl.ac.uk

ABSTRACT. The technique of geographically weighted regression (GWR) is used to model spatial 'drift' in linear model coefficients. In this paper we extend the ideas of GWR in a number of ways. First, we introduce a set of analytically derived significance tests allowing a null hypothesis of no spatial parameter drift to be investigated. Second, we discuss 'mixed' GWR models where some parameters are fixed globally but others vary geographically. Again, models of this type may be assessed using significance tests. Finally, we consider a means of deciding the degree of parameter smoothing used in GWR based on the Mallows C_p statistic. To complete the paper, we analyze an example data set based on house prices in Kent in the U.K. using the techniques introduced.

1. INTRODUCTION AND BACKGROUND

It has long been recognized that there are inherent difficulties in analyzing spatial data (Cressie, 1991; Haining, 1990; Griffith, 1988; Upton and Fingleton, 1985; Cliff and Ord, 1981). One of these, which we address directly in this paper, is spatial nonstationarity: the variation in relationships and processes over space (Bailey and Gatrell, 1995). Another is that of spatial dependency, which has been examined, primarily within a regression framework (Cliff and Ord, 1981; Odland, 1988; Anselin, 1993). One of the major effects of spatial dependency that has been discussed in some detail is that on the estimates of the standard errors of regression parameter estimates when the error terms in the model exhibit spatial autocorrelation. Anselin (1993) has developed spatial regression models that can ameliorate this problem. The problems of spatial nonstationarity and spatial dependency are obviously related. If spatially varying relationships are modeled within a global framework such as standard regression then the error terms in the global regression model will exhibit spatial autocorrelation. To demonstrate this, imagine a region in which y is regressed on x and that the resulting global model is

---

Received August 1997; revised May 1998; accepted August 1998.

(1) $$y_i = 1.5 + 2x_i$$

However, suppose that in one part of the region this relationship is actually

(2) $$y_i = 1.5 + 2.5x_i$$

and in another part of the region the actual relationship is

(3) $$y_i = 1.5 + 1.5x_i$$

Then applying Equation (1) to that part of the region in which Equation (2) holds will lead to underestimates of the y_i values in that region and positive residuals, assuming x_i is positive. Conversely, applying Equation (1) to that part of the region in which Equation (3) holds, will lead to overestimates of the y_i values and negative residuals. Hence, there will be a strong positive autocorrelation of the error terms resulting from the inability of the global model to deal with the spatial nonstationarity of the relationships being measured. The established method of dealing with this problem is to calibrate a spatial regression model that allows for spatial autocorrelation of the error terms (Anselin, 1993; Griffith, 1988; Cliff and Ord, 1981; Cressie, 1991) although there are problems in the implementation of such a model because of the need to assume a global autocorrelation function for the error terms and the need to determine a suitable weights matrix for the autocorrelation function.

An alternative, and more direct approach to the problem, is *Geographically Weighted Regression* (GWR). In GWR any spatial nonstationarity in the relationships being measured is accounted for by allowing the calibrated model to vary spatially (Brunsdon, Fotheringham, and Charlton, 1996). In this sense, the difference between GWR and the spatial error approach is that in the former, spatial drift from 'average' global relationships is measured directly, whereas in the latter it is measured as a second-order effect through the spatial distribution of residuals. We see merit in the GWR approach which also produces maps of parameter variations over space and can be used to improve our understanding of the processes being modeled, and thus separate local spatial anomalies in terms of each explanatory variable.

There are several reasons why spatial variations in relationships may occur. For example, in a regression model used to predict the price of houses it is possible that the value of an extra bedroom may vary from place to place—in areas close to good schools there may be more demand for houses by families with several children, so that the utility of an extra bedroom is high. Similarly, in cities where the housing stock is very old there may be few houses with garage facilities, so the added value of a garage may be greater than usual. This phenomenon can occur in many other fields of study—for example in modeling the spatial distribution of certain types of illness as a function of social and economic conditions (Fotheringham, Brunsdon, and Charlton, 1996), or modeling crime rates, voting behavior, and out-migration rates over space in terms of a casewise vector of explanatory variables. Indeed, in any analysis of spatial

data GWR may be used as a diagnostic for a global modeling approach to examine for the presence of spatial nonstationarity in relationships.

At this stage it is perhaps helpful to consider the mathematical expression of the GWR model. If an ordinary least squares (OLS) regression model takes the form

$$(4) \qquad y_i = \sum_j x_{ij} a_j + \varepsilon_i$$

then the model for GWR is

$$(5) \qquad y_i = \sum_j x_{ij} a_j (u_i, v_i) + \varepsilon_i$$

where (u_i, v_i) is the location in geographical space of the ith observation. Note that in the case where the function $a_j(u_i, v_i)$ is a constant for all (u_i, v_j), then the OLS model of Equation (4) holds. The method proposed to calibrate Equation (5) in GWR is basically a moving kernel window approach. To estimate $a_j(.,.)$ at the location of the ith observation one carries out a weighted regression where each observation is given a weight w_{ki}, if the observations are indexed by the variable k. Note that w_{ki} should be a monotone decreasing function of d_{ki}, the Euclidean distance between the points (u_i, v_i) and (u_k, v_k). For examples of applications of the technique, see (Fotheringham, Charlton, and Brunsdon, 1997; Brunsdon, Fotheringham, and Charlton, 1996; Fotheringham, Charlton, and Brunsdon, 1998).

Note that this approach may be generalized. An estimate of $a_j(.,.)$ can be obtained at any point (u, v) in this way—not just at the locations of the sample observations. This is particularly useful in, say, depicting $a_j(u, v)$ as a surface or a set of contour lines superimposed over the study area. There are several linkages between this approach and the idea of kernel probability density estimation (Silverman, 1986; Brunsdon, 1995), and that of kernel regression (Wand and Jones, 1995). Furthermore, as the sample size tends to infinity GWR can provide a consistent and unbiased estimator of $a_j(u, v)$ provided a suitable rule for choosing bandwidth as n increases is given (Stone, 1977; Wand and Jones, 1995).

An important geographical note is struck here—this technique provides a means of computing *localized* regression estimates. If the weighting function is a monotone decreasing function of distance then observations further from the point at which the model is calibrated weight less than observations closer to that point. Thus, GWR provides a localized set of parameters at this position in space. This may be advantageous in a geographical context. For example, if at some point the housing markets in two places are behaving very differently it is not helpful to calibrate models in either of these places using data from the other.

It should be noted that techniques already exist that can provide local estimates of regression parameters. Perhaps the best known of these is the

expansion method (Jones and Casetti, 1992; Casetti, 1972). In this framework, parameters of a global model are expanded in terms of other attributes that can include location so that parameter drift across space can be measured (Eldridge and Jones, 1991). However, as Fotheringham, Charlton, and Brunsdon (1998) show in some detail, the expansion method essentially calibrates prespecified parametric forms for trends in relationships over space rather than directly calibrating relationships as GWR does. Consequently, the expansion method may miss important spatial variations in relationships and produces relatively simplistic descriptions of any spatial nonstationarity.

However, like all models, GWR should be capable of being tested. In cases where a simple OLS model holds, the extra complexity of varying geographical coefficients is an unwelcome intrusion. It is therefore important to develop a set of statistical tests to see whether the results of fitting a GWR model are due to a genuine spatial drift in parameters, or whether the variations are simply caused by random fluctuations in the ε term in the model. Clearly the GWR approach will always give a better fit in terms of sums of squared residuals (RSS): the greater flexibility of the regression coefficients over space allows this to happen. However, as with any regression model it must be tested whether this inevitable reduction is statistically significant. The development of such tests will be the core of this paper. The most basic test (discussed in the next section) attempts to answer the question: *Is the performance of the GWR significantly better than that of ordinary regression?* In later sections we will consider more sophisticated tests relating to individual variables and the form of the kernel weighting function.

2.   A BASIC TEST OF THE GWR MODEL

One of the main products of GWR is a set of local parameter estimates that can be mapped and which describe the degree of spatial nonstationarity in a relationship. An obvious question to ask is then: Does significant spatial nonstationarity exist? Obviously spatial variations in parameter estimates will exist due to sampling error; we need to evaluate the degree to which variations exist and ask what is the probability the observed variations can be ascribed purely to sampling.

In an earlier paper, the degree of spatial nonstationarity was assessed visually and also by constructing experimental distributions based on Monte Carlo procedures (Fotheringham, Charlton, and Brunsdon, 1998). In the latter, the variance of the local parameter estimates was calculated and the data reorganized randomly across the spatial units. Rerunning GWR on this randomized data produced a new set of local parameter estimates which again generated a variance. Repeating this procedure yielded 99 experimental variances against which to compare the variance from the observed data set. However, this technique is computationally demanding because of the need to rerun the GWR procedure for each randomized data set, and therefore it is the task of the

remainder of this paper to develop a theoretical significance testing procedure for GWR.

   We need a test of the null hypothesis that the functions $a_i(u, v)$ are constant for all points (u, v) in the study area. If there is no evidence to reject this hypothesis, it suggests that an ordinary, global regression model is an adequate descriptor of the data. In more formal terms, the hypothesis is set out as

$$H_0 \qquad : \qquad \frac{\partial a_j}{\partial u} \equiv \frac{\partial a_j}{\partial v} \equiv 0 \qquad \forall j$$

against

$$H_1 \qquad : \qquad \frac{\partial a_j}{\partial u} \not\equiv 0 \text{ and } \frac{\partial a_j}{\partial v} \not\equiv 0 \qquad \forall j$$

or more specifically, H_0 will be compared against a subset of H_1 corresponding to a GWR-type estimation of the $a_j(u_i, v_i)$s.

   Before proceeding, it is worth establishing some vector/matrix notatation. Let the matrix $\mathbf{X}$ have elements x_{ij} representing observation i on variable j and $\mathbf{y}$ be a vector with elements y_i representing observation i. These are the standard array form of the data used in many regression models. In the GWR model, the coefficients $a_j(u, v)$ vary across geographic space. Borrowing from the usual OLS notation, these can be gathered into a vector $\mathbf{a}(u, v)$, so that $\mathbf{a}(.,.)$ is a vector function mapping $\mathbb{R}^2$, a two-dimensional Euclidean plane, onto $\mathbb{R}^m$, an m-dimensional Euclidean hyperplane. Under OLS, the maximum likelihood estimate of $\mathbf{a}(u, v)$ is given by

$$\hat{\mathbf{a}}(u, v) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

where the hat denotes that $\hat{\mathbf{a}}(u, v)$ is an estimate. Note that although here the estimate is shown as a function of (u, v), in OLS it is in fact assumed constant with respect to them. To obtain an estimate of $\mathbf{y}$ one has to premultiply by $\mathbf{X}$

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

or, more simply

$$\hat{\mathbf{y}} = \mathbf{S}_o\mathbf{y}$$

where

$$\mathbf{S}_o = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

$\mathbf{S}_o$ is a *smoothing* operator that takes the observed $\mathbf{y}$ and 'smoothes' it to $\hat{\mathbf{y}}$, the fitted value after the random residuals have been removed. It is also sometimes referred to as the *hat matrix* or *hat operator* because it transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$. Note that $\mathbf{S}_o$ does not depend on $\mathbf{y}$.

   Now consider GWR. In this case, a weighted OLS is used to estimate $\mathbf{a}(u, v)$, but the weighting changes as (u, v) varies. Suppose $\mathbf{W}(u, v)$ represents a diagonal

matrix whose diagonal elements correspond to the weighting scheme for a particular (u, v). Then

$$\hat{\mathbf{a}}(u, v) = [\mathbf{X}^T\mathbf{W}(u, v)\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{W}(u, v)\mathbf{y}$$

We can see that a hat matrix exists for GWR by considering the expression for $\hat{\mathbf{y}}$ from the calibrated GWR model. At any given y_i this requires $\mathbf{X}_i$, the ith row of $\mathbf{X}$, and $\hat{\mathbf{a}}(u_i, v_i)$. Then we have

$$y_i = \mathbf{X}_i(\mathbf{X}_T\mathbf{W}(u, v)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(u, v)\mathbf{y}$$

Note that $\mathbf{X}_i(\mathbf{X}_T\mathbf{W}(u, v)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(u, v)$ is a row vector, say $\mathbf{r}_i$. Then if $\mathbf{S}_1$ is constructed so that its ith row is $\mathbf{r}_i$, we have

$$\hat{\mathbf{y}} = \mathbf{S}_1\mathbf{y}$$

So GWR also has a hat matrix, $\mathbf{S}_1$. As before, this does not depend on $\mathbf{y}$. In either of these models, the residuals may be expressed as

$$\hat{\varepsilon} = (\mathbf{I} - \mathbf{S}_z)\mathbf{y}$$

where $\mathbf{z}$ is either 0 or 1. Thus, the sum of squared residuals can be expressed as

$$
(6) \quad
\begin{aligned}
\hat{\varepsilon}^T\hat{\varepsilon} &= \mathbf{y}^T\left(\mathbf{I} - \mathbf{S_z}\right)^T\left(\mathbf{I} - \mathbf{S_z}\right)\mathbf{y} \\
&= \mathbf{y}^T\mathbf{R_z}\mathbf{y}
\end{aligned}
$$

where

$$\mathbf{R}_z = (\mathbf{I} - \mathbf{S}_z)^T(\mathbf{I} - \mathbf{S}_z)$$

Thus, for both OLS and GWR the sum of squared residuals are quadratic forms in $\mathbf{y}$. If it is assumed that the ε_is have independent and identical Normal distributions, then Equation (6) is a quadratic form of Normal variates.

A result from Kendall and Stuart (1977) concerning the test of a null model against an alternative becomes useful here. This basically states that if both models are expressed in hat matrix form, for Normally distributed $\mathbf{y}$, then the expression

$$
(7) \quad F = \left[\frac{\left(\mathbf{y}^T\mathbf{R}_0\mathbf{y}\right) - \left(\mathbf{y}^T\mathbf{R}_1\mathbf{y}\right)}{v}\right]\left[\frac{\left(\mathbf{y}^T\mathbf{R}_1\mathbf{y}\right)}{\delta}\right]^{-1}
$$

where $v = Tr(\mathbf{R}_0 - \mathbf{R}_1)$ and $\delta = Tr(\mathbf{R}_1)$ has an approximate F-distribution with degrees of freedom given by $(v^2/v', \delta^2/\delta')$, where $v' = Tr[(\mathbf{R}_o - \mathbf{R}_1)^2]$, $\delta' = Tr(\mathbf{R}_1{}^2)$, and v^2 and δ^2 are the respective squares of v and δ.

This approximation is based on the fact that both the numerator and denominator above are quadratic forms of Normal variates, and these are well approximated by χ^2 distributions with the degrees of freedom chosen so that

their first and second moments agree with those of the quadratic forms. Note that these degrees of freedom are not neccessarily integers, but that the F-distribution is well-defined for any positive degree of freedom parameters. In future analyses these two quantities will be referred to as the *effective degrees of freedom* in the model. Since GWR and OLS both have hat matrices, it is possible to compare these two models using the above statistic. In fact, this procedure is identical to the more usual procedure for comparing two OLS models, when one has included more explanatory variables than the other. It can be shown that in this case the degrees of freedom are always integers and the F-distribution assumption is exact. This suggests that GWR–OLS comparisons may be expressed in the form of an ANOVA table, with residual mean squares for both GWR and OLS being compared.

3.   AN EXAMPLE OF A 'WHOLE MODEL' TEST

As an example of the testing procedure outlined in the last section, consider the following data set. These data relate to house sales whose mortgages were arranged by the Nationwide Anglia Building Society in 1991. Geographically, these data are based on cases centered around Deal, in Kent, U.K. (see Figure 1). There are 82 cases in the study. For each mortgage arranged, details of the house itself, its agreed selling price, and its postcode are recorded. In this case the purpose of regression modeling is to attempt to predict selling price from other attributes of the house. This example is being used to demonstrate the use of the statistic introduced above, rather than to study house price in any detail so a relatively simple bivariate model will be used—attempting to predict house selling price (in thousands of pounds) as a function of floor area (in square meters). Typically, one would expect larger houses to sell at higher prices. Of course, far more detailed models could be used—there are several other features of houses recorded, such as number of bedrooms, number of bathrooms, and so on. However, because these factors correlate strongly with floor area it was felt that for this illustrative example it was sufficient to investigate this as a broad effect. A more detailed analysis of hedonic price models is currently the subject of work in progress by the authors.

If we refer to the dependent and independent variables as HOUSE PRICE and FLOOR AREA then a possible OLS regression model might be

(8) $(\text{HOUSE PRICE}) = a_0 + a_1 (\text{FLOOR AREA})$

Anybody who has attempted to buy housing in the UK (and quite probably anywhere else) will immediately identify one problem with Equation (8): the effect of geographical location is not allowed for. As well as size, one unique aspect of any house is its geographical location. Social conditions in an area, and its closeness to certain facilities have a strong influence on property values. The area of Kent being studied—around Deal—lies near to the U.K. coastline and has a castle, a golf course, two coastal launching sites, a railway station, and a major road passing through. Factors such as closeness to any of these features
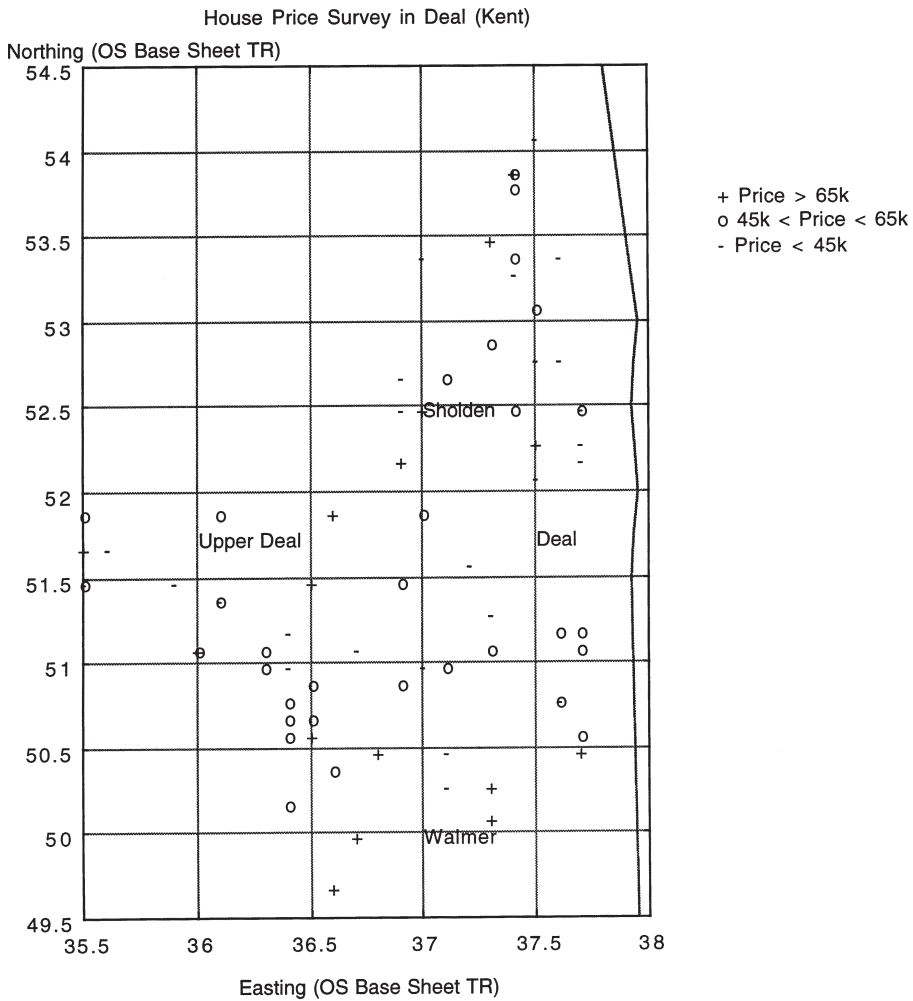
FIGURE 1: Locations of Samples.

may affect the demand for housing, and hence house prices. Although the study area itself spans only a few kilometers, it is likely that even over this space there will be geographical factors affecting house prices. Although price will vary with size in the same area, it is likely that the exact nature of the relationship will vary between areas. For this reason, a GWR model of house price is also proposed

(9)   HOUSE PRICE $= a_0(u, v) + a_1(u, v)$ FLOOR AREA

where, as before, (u, v) represents a spatial location.

First, consider the geographic spread of the sample. In Figure 1 a map of the cases is provided, showing the coastline, major places, and the values of houses themselves. The castle is located towards Walmer, the railway station further north, near Sholden. The general shape of the residential region is indicated by the distribution of the sample points.

A scatter plot of the variables FLOOR AREA (in square meters) and HOUSE PRICE (in thousands of pounds) is shown in Figure 2. From this, it can be seen that there is a roughly linear relation between the two. Fitting an OLS model gives the results in Table 1.

Figure 3 shows a scatter plot of absolute residual values. At first glance this might be thought to indicate heteroskedasdicity in the residuals—however, one must be wary of such interpretations—see, for example, Cleveland (1979). Due to changes in the density of the x variable, the expected y range in any vertical section of the plot will vary—the denser the x variate the greater the range. Cleveland (1979) suggests compensating for this by adding a Loess fit of absolute values of residuals against FLOOR AREA to a plot of absolute values of residuals. This will measure mean absolute residuals for fixed size subsamples because a Loess plot is a moving window regression technique based on quantiles of the data set. This line has been added to Figure 3 and suggests a lack of strong trends in the size of residuals as a function of FLOOR AREA.

The above suggests that the pattern of the residual plot is not a problem, but it may be that the size of the residuals is a problem. The mean absolute error is about £11,000 whereas the average price is around £55,000; in practical terms this is an uncomfortably large margin of error. A more enlightening examination of the residuals is achieved using linked plots of the form advocated by Haslett,
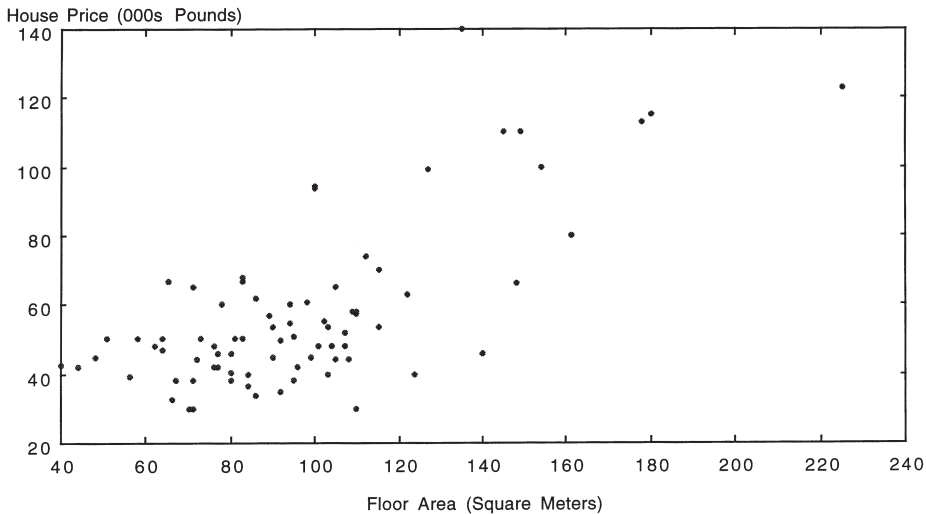

FIGURE 2: House Price Versus Floor Area.

TABLE 1: Results of Ordinary Least Squares Fit of HOUSE PRICE
on FLOOR AREA

Variable	$\hat{a}_i$	Standard Error ($\hat{a}_i$)
FLOOR AREA	0.496	0.059
Intercept	8.655	5.096

Wills, and Unwin (1990) and Brunsdon (1995). Here two interactive plots, one of geographical location and the other of residuals, can be 'brushed' with the cursor in a window-based computer environment to see which geographical location corresponds to which residual. Using this approach one effect that became clear was that positive and negative residuals exhibited strong spatial clustering. There were distinct clusters of over- and under-prediction.

This pattern is quite different from the purely random pattern one might expect under the assumptions of residual behavior made in an OLS model. The evidence suggests that adopting the GWR approach could improve the predictive performance of the model. Such a model was fitted to the data with a Gaussian kernel of the form

$$(10) \qquad\qquad K(d) = \exp(-d^2/k^2)$$

Here, the kernel bandwidth k is equal to 400 meters.

Inspecting the output of GWR in the form of shaded pixel maps (Figures 4 and 5), one can see higher values in the intercept and slope values in the south, towards Walmer. Interestingly, Walmer is the site of a historic monument and proximity to this appears to have a positive effect on house price. Note also that the intercept and slope values increase near to the two railway stations. This suggests that closeness to the railway station adds a premium to house price and also that the increase occurs in proportion to the size of the house. This could also be interpreted as demonstrating the exploratory nature of the technique. By combining geographical information to the GWR result the linkages to railway stations becomes apparent. This was achieved without explicitly specifying a variable such as distance to railway station, or a speculative parametric form for a distance-decay effect.

In Figure 6, a Loess fit is used to explore possible heteroskedasticity in the error term in the same manner as in Figure 3. Because there does seem to be some trend in the Loess curve, this is further investigated by producing 1 percent pointwise confidence intervals around the curve using Monte-Carlo simulation (Hastie and Tibshirani, 1990, p. 60–64). Here, the confidence bands are based on a null hypothesis of a constant level of error variance with floor area, and are simulated using random permutations of the absolute residuals against the FLOOR AREA values. It can be seen in Figure 6 that the Loess curve remains within the confidence bands and therefore it seems reasonable to work with the assumption of a fixed variance error term.
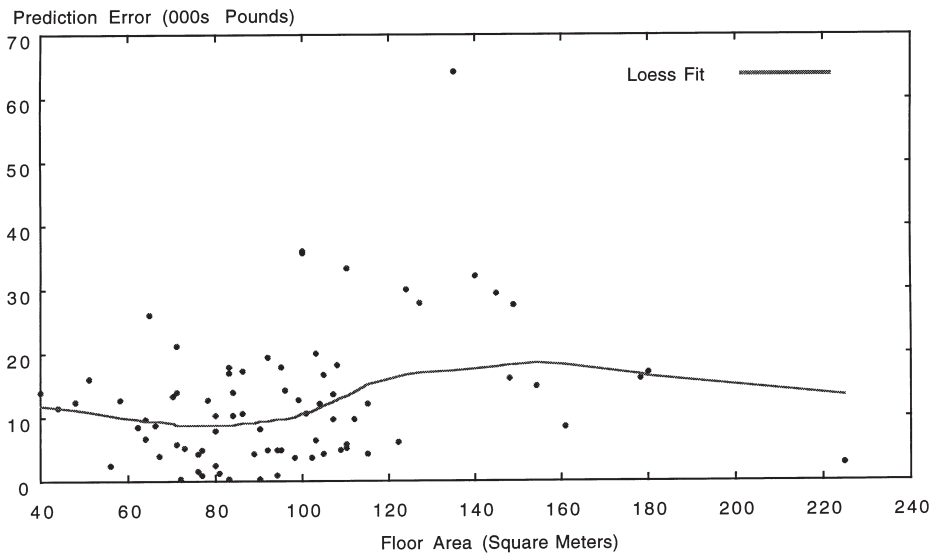
FIGURE 3: Residual Plot for OLS Model.

        This analysis may be formalized by carrying out an ANOVA test as proposed
in the previous section. This will compare the fit of the OLS model to that of
GWR. The results of this test are shown in Table 2. In this table, the first column
shows the residual sum of squares of the OLS residuals (first row), the RSS for
the GWR residuals (bottom row) and the difference between these two (middle
row). The second column gives the respective degrees of freedom for each of these.
The middle and bottom rows give v and δ in Equation (7) respectively, while the
top row shows two degrees of freedom for the linear model. The third column
MS, Mean Square, gives the results of dividing the sums of squares by their
respective degrees of freedom—the numerator and denominator in Equation (7).
Finally, dividing the former by the latter gives the psuedo-F statistic, and the
p-value may then be computed as set out earlier.
        Note the reduction in residual sum of squares when the GWR approach is
used. The results of this ANOVA suggest that the null hypothesis OLS model
should be rejected in favor of GWR. Note that the degrees of freedom are
noninteger, and also that the degrees of freedom for the improvement in residual
sum of squares is not equal to the difference of the GWR residuals and the OLS
residuals. These figures are obtained from Equation 7. Clearly, this suggests that
there is a significant improvement in the model fit when GWR is adopted.

4.   A MORE FLEXIBLE APPROACH TO GWR
        The result of the previous section suggests that, for the example data set, a
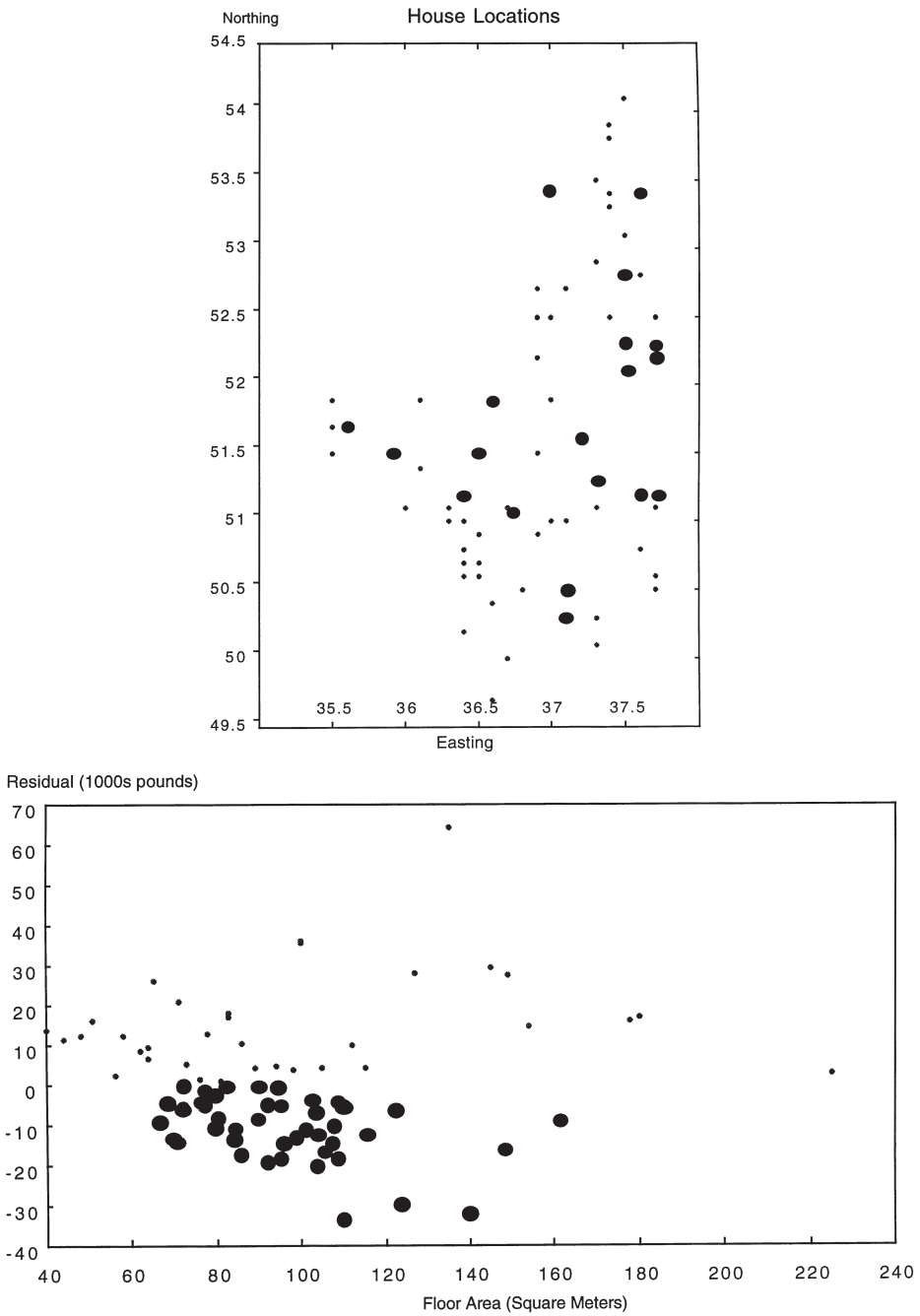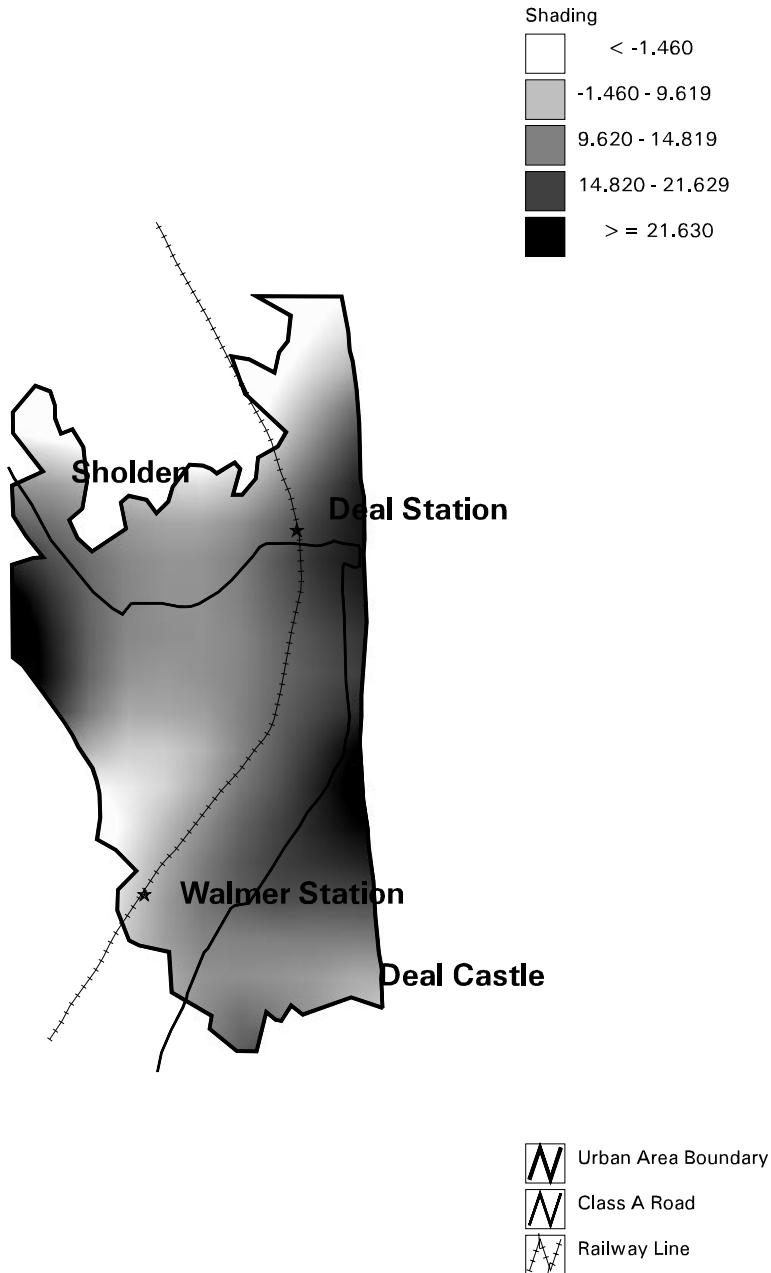model specification in which all of the terms are allowed to vary globally

FIGURE 4: Linked Plots Showing Relationship
Between Residuals and Geographical Location.
Large points are those selected by "Brushing."

Shading

	< -1.460
	-1.460 - 9.619
	9.620 - 14.819
	14.820 - 21.629
	> = 21.630

Sholden

Deal Station

Walmer Station

Deal Castle

Urban Area Boundary

Class A Road

Railway Line

Full GWR Model: Intercept

FIGURE 5: Full GWR Floor Area Coefficient Surface.
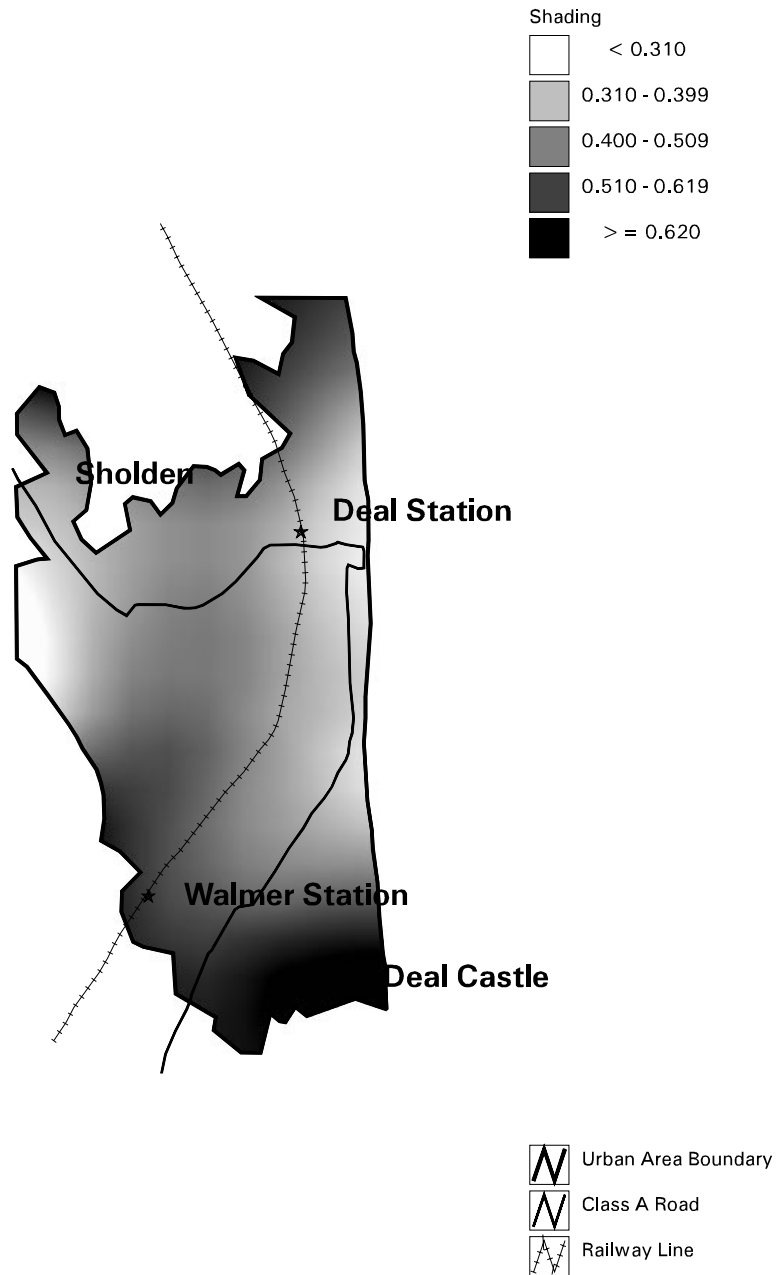
Full GWR Model: Gradient

FIGURE 6: House Price Residuals Loess Fit.
Note: See text for explanation of upper and lower bands.

TABLE 2: ANOVA comparing GWR to OLS

Source of Variation	SS	DF	MS	F	p-value
OLS Residuals	22152.	2.00			
GWR Improvement	18257.	47.42	385.0		
GWR Residuals	3895.	32.58	119.5	3.22	0.00

performs better than one in which they are fixed globally. That is, the effect of allowing two parameters to vary spatially was compared to the effect of allowing none to vary. However, a reasonable strategy may be to allow one of the parameters to vary over space while the other is held fixed. This may be an appropriate model in situations where certain factors influencing the dependent variable are global in nature, while others are local. For example, suppose there was a tax incentive for purchasing new housing stock, and this incentive was fixed on a national basis. In this case, provided the study area was entirely contained in the country where this took place, the effect on house price of a 'new house' dummy variable might be constant throughout the study area. However, the influence of other variables could vary geographically. In certain situations, it may be obvious which effects are geographically constant and which will vary, but this is not always so. For example, in the GWR analysis in the previous section it is not clear whether the noted improvement in model fit was due to the variability of the slope term, or the intercept term, or both. Each possible outcome has a different interpretation. If the FLOOR AREA term is fixed but the intercept varies then this implies that the marginal cost per unit area of property is global, but there is a size-independent premium for living in certain areas. However, if the FLOOR AREA term also varies, this suggests that difference in price between large and small houses also varies geographically. In its initial form, as documented in Brunsdon, Fotheringham, and Charlton (1996), GWR cannot help to answer this question. The approach set out is an all-or-nothing procedure in which spatial variation must be estimated for every parameter. In this section a new development in GWR that addresses this problem is introduced.

If it is desired to allow some parameters to vary over space, but others to be fixed, then a new model, which is a hybrid of Equations (4) and (5) should be specified as

$$(11) \qquad y_i = \sum_{j=1,k} x_{ij} a_j + \sum_{j=k+1,m} x_{ij} a_j (u_i, v_i) + \varepsilon_i$$

Here, the first k coefficients are fixed and the remainder vary. Returning to the matrix notation, the independent variable matrix $\mathbf{X}$ may be partitioned vertically after column k into the two matrices $\mathbf{X_c}$ and $\mathbf{X_v}$, representing the variables associated with constant and varying coefficients, respectively. Usually one of these will have a column of ones, giving an intercept term. Which one has this property will depend on whether the intercept term is to be fixed over space.

To calibrate this model, the *back-fitting* approach of Hastie and Tibshirani (1986) may be used. In this method, suppose we are fitting a model of the form

$$\mathbf{y} = \mathbf{f_c}(\mathbf{X_c}) + \mathbf{f_v}(\mathbf{X_v}) + \varepsilon$$

and suppose we have an initial guess for $\mathbf{f_c}(\mathbf{X_c})$, say $\hat{\mathbf{f}}_\mathbf{c}^{(0)}(\mathbf{X_c})$. Then, if we have a method for calibrating a model of the form

(12) $$\mathbf{y} = \mathbf{f_v}(\mathbf{X_v}) + \varepsilon$$

we can use this to obtain an estimate of $\mathbf{f_v}(\mathbf{X_v})$, say $\hat{\mathbf{f}}_\mathbf{v}^{(1)}(\mathbf{X_v})$ by calibrating the model

$$\mathbf{y} - \hat{\mathbf{f}}_\mathbf{c}^{(0)}(\mathbf{X_c}) = \mathbf{f_v}(\mathbf{X_v}) + \varepsilon$$

In a similar manner, if we have a method for calibrating

(13) $$\mathbf{y} = \mathbf{f_c}(\mathbf{X_c}) + \varepsilon$$

an estimate of $\hat{\mathbf{f}}_\mathbf{c}^{(1)}(\mathbf{X_c})$ can be obtained. Each two-stage iteration of this sort gives new calibrations of the two nonrandom terms making up the model, say $\hat{\mathbf{f}}_\mathbf{v}^{(i)}(\mathbf{X_v})$ and $\hat{\mathbf{f}}_\mathbf{c}^{(i)}(\mathbf{X_c})$ for the *i*th iteration, and eventually these should converge.

Fortunately, previous sections have shown that if $\mathbf{f_c}(\mathbf{X_c})$ is an OLS regression model and $\mathbf{f_v}(\mathbf{X_v})$ is a GWR regression model, then we do have ways of calibrating both of these, and so we may carry out the back-fitting procedure. In fact, the methods for estimating the functional values in Equations (12) and (13) are just premultiplication of the **y** vectors by the appropriate hat matrices. If we call the hat matrix for the OLS part of the model $\mathbf{S_c}$ and that for the GWR part $\mathbf{S_v}$ and then define

(14) $$\hat{\mathbf{y}}_\mathbf{c} = \mathbf{S_c}\mathbf{y}$$

(15) $$\hat{\mathbf{y}}_\mathbf{v} = \mathbf{S_v}\mathbf{y}$$

so that we have

(16) $$\hat{\mathbf{y}} = \hat{\mathbf{y}}_\mathbf{c} + \hat{\mathbf{y}}_\mathbf{v}$$

then the procedure for calibrating Equation (11) is as follows

Step 1. Supply an initial guess for $\hat{\mathbf{y}}_\mathbf{c}$, say $\hat{\mathbf{y}}_\mathbf{c}^{(0)}$. In practice, good results have been achieved by regressing $\mathbf{X_c}$ on **y** using OLS. This is achieved by setting $\hat{\mathbf{y}}_\mathbf{c}^{(0)}$ to $\mathbf{S_c}\mathbf{y}$.

Step 2. Set *i* equal to 1.

Step 3. Set $\hat{\mathbf{y}}_\mathbf{v}^{(i)}$ to $\mathbf{S_v}[\mathbf{y} - \hat{\mathbf{y}}_\mathbf{c}^{(i-1)}]$.

Step 4. Set $\hat{\mathbf{y}}_{\mathbf{c}}^{(i)}$ to $\mathbf{S}_{\mathbf{c}}[\mathbf{y} - \hat{\mathbf{y}}_{\mathbf{v}}^{(i)}]$.

Step 5. Set i equal to $i + 1$.

Step 6. Loop back to Step 3 unless $\hat{\mathbf{y}}^{(i)} = \hat{\mathbf{y}}_{\mathbf{c}}^{(i)} + \hat{\mathbf{y}}_{\mathbf{v}}^{(i)}$ is sufficiently close to $\hat{\mathbf{y}}^{(i-1)}$ to suggest convergence.

Provided some regularity conditions apply to both hat matrices we can be sure of convergence (Hastie and Tibshirani, 1990 pp. 108–110). Whether the value to which the $\hat{\mathbf{y}}_{\mathbf{c}}^{(i)}$s converge is a consistent estimate of $\hat{y}_c$ is not currently known. However, for a nongeographical form of a model similar to this (a generalized additive model with some linear terms) consistency has been shown (Hastie and Tibshirani, 1990 p. 118). Clearly, some future theoretical work in this area would be valuable.

As an example, consider again the data set from the previous section. A model of the form

(17) $\qquad$ HOUSE PRICE $= a_0(u, v) + a_1$ FLOOR AREA

is now proposed. This can be thought of as a house price consisting of an 'area effect' together with a linear 'size effect.' It is hoped that the spatially varying intercept term will reflect characteristics of the location of the house, such as nearness to facilities and socioeconomic characteristics of the area. However, unlike model (9), this assumes no interaction between the location of the house and the cost per square meter in the size component of the price. That is, although this model assumes that house price does alter geographically, the incremental cost of an extra square meter of floor space is global. Another viewpoint is that in estimating $a_0(u, v)$ here, one is attempting to construct a house-price surface after controlling for variations in the size of the housing stock.

The algorithm outlined above was used to calibrate this model. Convergence, although not rapid in this example, occured after about 80 iterations. In practical terms, this takes about 3 minutes on a Power Macintosh 7200/75, and about 1.5 minutes on a Macintosh Powerbook 1400/133, using the XLISP-STAT package (Tierney, 1990).

The intercept surface is shown as a contour map in Figure 7, and a residual plot for the mixed model is shown in Figure 8. The patterns in Figure 7 are similar to those followed by both the intercept and the slope in the full GWR model, favoring house prices around Walmer.

5. SIGNIFICANCE TESTING FOR MIXED GWR MODELS

The example in the last section could be made more useful if some form of significance testing could be applied to mixed models of the general form of Equation (11). If this were the case, one could test whether allowing just the intercept term in the house-price model to vary was a significant improvement on OLS, and whether moving from the mixed model to full GWR was also a
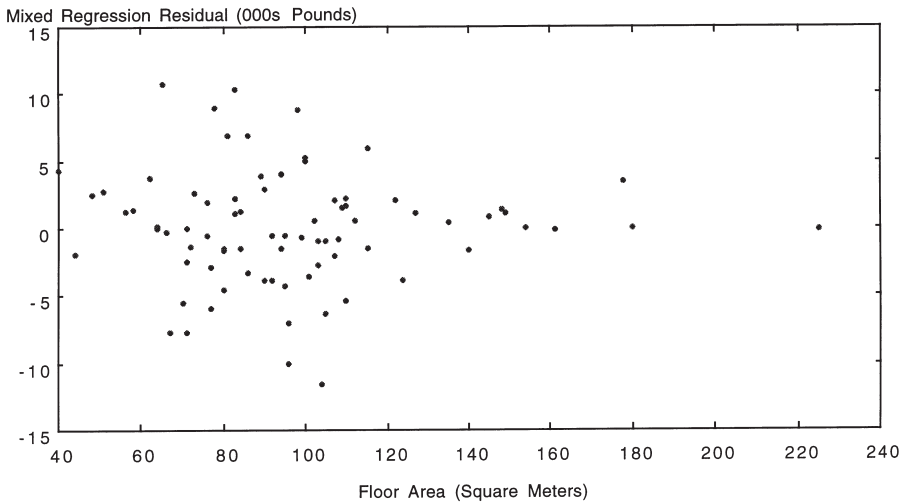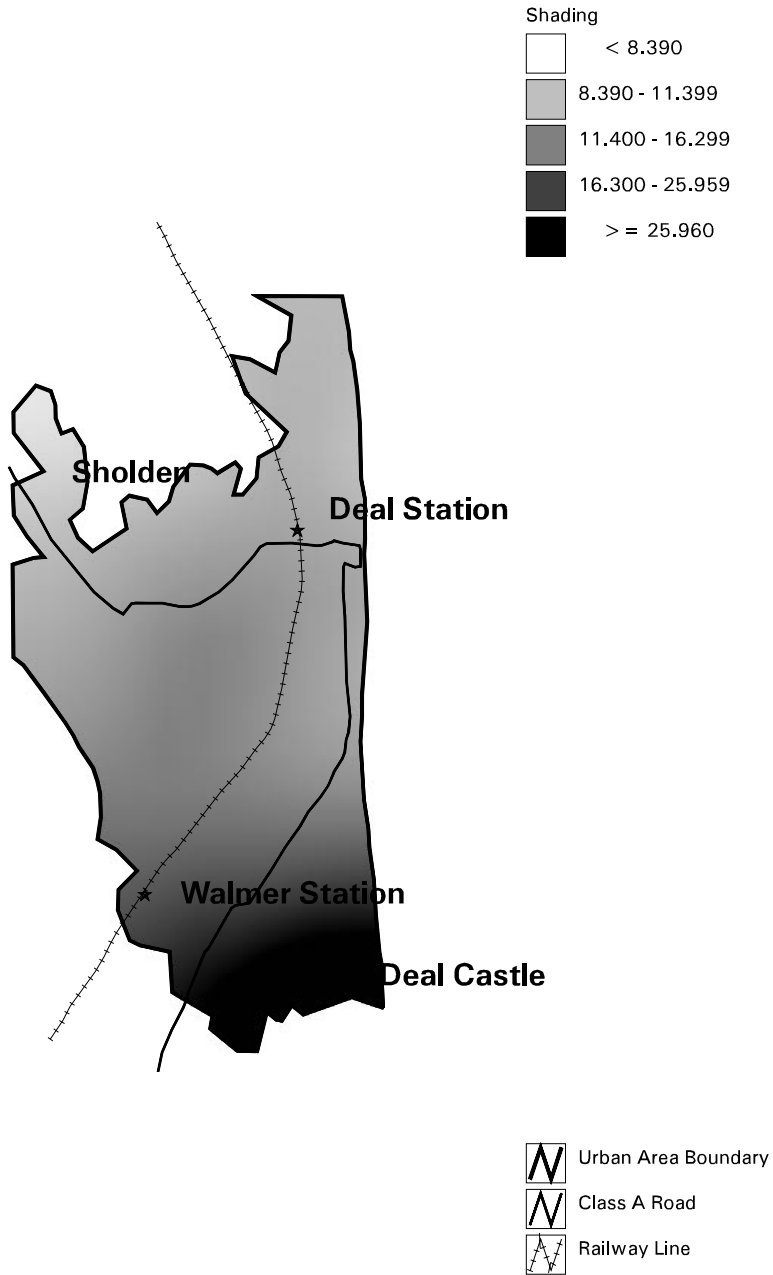
FIGURE 7: House Price Residuals (GWR Model).

significant model improvement. In order to derive a test of this sort, note that
in the mixed GWR algorithm, both $\mathbf{y}_{\mathbf{c}}^{(\mathbf{i})}$ and $\mathbf{y}_{\mathbf{v}}^{(\mathbf{i})}$ can be expressed as the result
of repeated matrix premultiplications applied to $\mathbf{y}$. Thus, we may write

$$\hat{\mathbf{y}}_{\mathbf{c}}^{(\mathbf{i})} = \mathbf{L}_{\mathbf{c}}^{(\mathbf{i})}\mathbf{y}$$
$$\hat{\mathbf{y}}_{\mathbf{v}}^{(\mathbf{i})} = \mathbf{L}_{\mathbf{v}}^{(\mathbf{i})}\mathbf{y}$$

noting that neither $\mathbf{L}_{\mathbf{c}}^{(\mathbf{i})}$ nor $\mathbf{L}_{\mathbf{v}}^{(\mathbf{i})}$ depend on $\mathbf{y}$. Eventually, when the algorithm
converges, we have

$$\hat{\mathbf{y}}_{\mathbf{c}} = \mathbf{L}_{\mathbf{c}}\mathbf{y}$$

$$\hat{\mathbf{y}}_{\mathbf{v}} = \mathbf{L}_{\mathbf{v}}\mathbf{y}$$

$$\hat{\mathbf{y}} = \mathbf{L}_{\mathbf{c}}\mathbf{y} + \mathbf{L}_{\mathbf{v}}\mathbf{y}$$

$$= (\mathbf{L}_{\mathbf{c}} + \mathbf{L}_{\mathbf{v}})\mathbf{y}$$

$$= \mathbf{L}\mathbf{y}$$

so that the mixed model estimator also has a hat matrix expression. Thus, once
again the Kendall and Stuart technique may be applied in ANOVA tests
comparing mixed models with full GWR models and OLS models.Computation-
ally, the appropriate hat matrix may be obtained by mirroring the operations
successively applied to $\hat{\mathbf{y}}_{\mathbf{c}}^{(\mathbf{0})}$ in the previous section, but with an *n*-by-*n* identity
matrix as the initial post-multiplicand.

Shading

☐	< 8.390
▨	8.390 - 11.399
▨	11.400 - 16.299
▨	16.300 - 25.959
■	> = 25.960

Sholden

Deal Station

Walmer Station

Deal Castle

Urban Area Boundary

Class A Road

Railway Line

Mixed GWR Model: Intercept

FIGURE 8: Residuals from GWR and Mixed GWR Models.

This method may be applied to the model used in the previous section, to test the full GWR model against a null hypothesis of Equation (17). The result, again in ANOVA format, is shown in Table 3. From this it may be seen that there is strong evidence that both the intercept and the slope vary. This may be interpreted by saying that the geographical effect does interact with the size of property, so that the cost per square meter of a house alters from place to place, as well as there being a premium added to the cost of a house which is independent of house size.

6. FURTHER TOPICS

Having introduced a significance testing method for GWR models, and considered ways in which the GWR model might be usefully extended, we use this closing section to explore some further issues and consider some new problems that may emerge.

Choosing Variables and Model Specification

The first problem here is almost as old as multiple regression itself: *How does one choose which variables belong in a regression model?* However, GWR adds a new twist to this—as well as deciding which variables should be added to a model, one must also decide whether these variables should be allowed to vary spatially. As with ordinary regression, these choices are rarely straightforward. Often, the order in which variables are added to regression models determines whether they are statistically significant or not. This problem can be particularly prominent when variables are strongly correlated. There is no guarantee that similar problems will not occur when sequentially replacing global regression terms of the form a_i with geographically specific $a_i(u, v)$ terms. The order in which replacements occur may affect the ANOVA tests outlined in this paper.

There is perhaps no clear-cut solution to this problem—unless there is some order of inserting regression terms (or allowing their spatial variability) in a sequence that is meaningful to some specific problem. For example, in a more comprehensive analysis of house prices it may be reasonable to allow the intercept term to vary first in an attempt to discover if there are any unobserved geographical variables influencing price. In this case the null hypothesis would take the form H_0: *all terms non-geographical* against H_1: *only the intercept varies geographically*. If the latter were true, this may suggest an external variable was having some effect.

TABLE 3: ANOVA comparing full GWR to mixed GWR

Source of Variation	SS	DF	MS	F	*p*-value
Mixed GWR Residuals	7848.5	36.8			
Full GWR Improvement	3953.8	12.64	312.8		
GWR Residuals	3894.6	32.58	119.5	2.61	0.01

A related, and important, issue is how reliant one should be on whether spatial variation is significant. The situation is best summarized by Chatfield (1994)

- A significant effect is not the same thing as an interesting effect

- A nonsignificant effect is not necessarily the same as no difference

The nightmare scenario might be a variant on stepwise regression, where as well as automatically inserting and deleting variables from a model, terms are allowed to vary spatially or globally constrained according to some ad hoc set of rules. Typically the final outcome of such a procedure depends on a set of parameters determining the behavior of the stepwise algorithm. In practice such parameters are frequently chosen arbitarily—perhaps as a set of default values on a computer software package—and the analyst is left with the task of constructing an interpretation from the output of this complex and ill-understood process. It is the view of the authors that good human knowledge of the subject matter underlying the statistical procedures is more likely to produce meaningful analyses than a black-box approach such as stepwise regression.

Choosing a Bandwidth

Another issue in GWR is the choice of bandwidth. The bandwidth in GWR determines the rate at which the regression weights decay around a given point (u, v). If the bandwidth is small, weights decay quickly with distance and the values of the regression coefficients change rapidly over space. Larger bandwidths produce smoother results. It is important to choose a suitable bandwidth to obtain reliable estimates of the spatial variation in the coefficients. To do this, one needs to discuss some properties of the estimator of y, say $\hat{y}$. At any geographical location, if we are given a set of predictors, $\mathbf{x}$, and a set of coefficient estimators, $\hat{\mathbf{a}}$, then $\hat{y} = \mathbf{x}^t\mathbf{a}$ is an estimate of y at that point. However, $\hat{\mathbf{a}}$ is an estimate of $\mathbf{a}$ based on a sample of spatially diffuse $\mathbf{x}$ and y observations. Due to the randomness of the y term, $\hat{\mathbf{a}}$ is random, and therefore so is $\hat{y}$. Two important properties of the distribution of $\hat{y}$ are its standard deviation and its expected value, SD[$\hat{y}$] and E[$\hat{y}$]. When, for all $\mathbf{x}$, E[y] = E[$\hat{y}$], the estimator is said to be unbiased. In this case, SD[$\hat{y}$] is, in itself, a useful measure of the quality of $\hat{y}$ as an estimator of y. However, zero bias does not in itself guarantee an optimal estimator.

Consider a pair of estimators for y. Call these $\hat{y}_1$ and $\hat{y}_2$ respectively. Suppose E[$\hat{y}_2$] $\neq y$ and E[$\hat{y}_1$] = y, but SD[$\hat{y}_2$] < SD[$\hat{y}_2$]. Although $\hat{y}_2$ is a biased estimator, its overall variability is less than that of $\hat{y}_1$. Thus, the likelihood of extreme values of potential errors in the prediction of y is less for $\hat{y}_2$—the only advantage $\hat{y}_1$ has to offer is that the error distribution is centered on zero. If one

were to consider the distributions of the squared error, that for the first estimator could have a much longer tail.

This is an example of *bias-variance trade-off*—an important issue that occurs in many types of statistical modeling. It is certainly an issue in GWR. If regression coefficients vary continuously over space, then using weighted least-squares regression is unlikely to provide a completely unbiased estimate of $\hat{\mathbf{a}}(u, v)$ at a given point (u, v). For each observation, there will be a different value of $\hat{\mathbf{a}}$, but the regression requires that this value is the same for all observations. The best one can hope is that the values do not vary too much. This is best achieved by only considering observations close to the (u, v) at which we wish to estimate $\hat{\mathbf{a}}(u, v)$. However, because this reduces the effective sample size for the estimate the standard error of $\hat{\mathbf{a}}(u, v)$ will increase. Thus, the question arises as to how close to (u, v) should points be considered? Too close and the variance becomes large, too far and the variance drops but bias increases. At one extreme, if a global model is chosen, then $\hat{\mathbf{a}}(u, v)$ is assumed constant for all (u, v). If there is much variability in the true $\hat{\mathbf{a}}(u, v)$ then clearly bias will cause problems. It is consideration of this phenomena that should guide the choice of bandwidth.

One approach is based on an idea of Mallows (1973) who proposes a statistic for ordinary linear regression defined by

$$c_p = \frac{\text{SS}}{\hat{\sigma}^2} + 2p - n$$

where SS is the sum of squared residuals for a given model, p is the number of parameters in the model, n is the number of cases and $\hat{\sigma}^2$ is the mean squared error estimate from a second regression model using extra variables. The idea is that the second regression model may be over-parametrized but will provide an unbiased (although perhaps inefficient) estimator of σ^2. In this case, we have $E[\hat{\sigma}^2] = \sigma^2$, and if the first regression model is also unbiased then standard regression theory (see for example, Dobson (1990) gives us that $E[\text{SS}] = (n - p)\sigma^2$. In this case $E[c_p] = p$. On the other hand, if the first regression model is biased, then $E[\text{SS}] > (n - p)\sigma^2$ and so $E[c_p] > p$. If c_p greatly exceeds p for some regression model, this suggests that the omission of some variables has caused bias. In this way, the statistic can be used to select a subset of variables from a list of candidates that still provide an unbiased model.

At first, this may seem irrelevent to the choice of bandwidth in a GWR model. However, choice of bandwidth is strongly linked to reduction of bias in models. A GWR with a small bandwidth will tend to give unbiased estimates because it will 'wrap' itself around features in the data. However, it will also provide estimates of y having very large standard errors because each locally weighted regression will only make use of a small number of variables. Decreasing bandwidth has a similar effect on the performance of GWR models as adding variables does to OLS models. Something similar to c_p would be useful here to check whether increasing bandwidth would introduce bias into the model. This

suggestion is not entirely novel; Cleveland (1979) suggests a similar approach for choosing the level of smoothing in Loess regression.

In this case, $\hat{\sigma}^2$ will be estimated from a GWR with a very small bandwidth (as before, we require an estimate with little bias). Then, as a statistic we use

$$M_k = \frac{\text{SS}}{\hat{\sigma}^2} + 2\text{Tr}\left(\mathbf{S}_k^T \mathbf{S}_k\right) - n$$

where $\mathbf{S}_k$ is the hat matrix for the GWR model under examination, with bandwidth k. If the bandwidth is providing a virtually unbiased estimate of y then the expected value of M_k will be $\text{Tr}(\mathbf{S}_k^T \mathbf{S}_k)$. This value is very similar to p in the original statistic. In fact, in the limiting case where $k \rightarrow \infty$, we have $\text{Tr}(\mathbf{S}_k^T \mathbf{S}_k) \rightarrow p$.

In fact, because M_k may be expressed as a linear function of the F-statistic used earlier in the paper, confidence limits of M_k under the null hypothesis that GWR with a bandwidth of k provides an unbiased estimator of y can be computed, and used in diagnostic plots. To see this in practice, M_k is plotted against-k in Figure 9. The vertical lines show the upper and lower 95 percentage points of the distribution of M_k, with a central check to show the expected value. The black dots are the observed values. When the dot falls below the check we have that c_p falls below $\text{E}[c_p]$ under a null hypothesis of no bias in the GWR calibration. On the basis of these diagrams k was chosen to be 0.18 kilometers. This may seem small, but it perhaps illustrates that house price demand exhibits spatial fluctuations even between streets in the same neighborhood.

Even with this method bandwidth choice is still a matter of judgement, essentially one requires a bandwidth giving an estimate of negligible bias, but
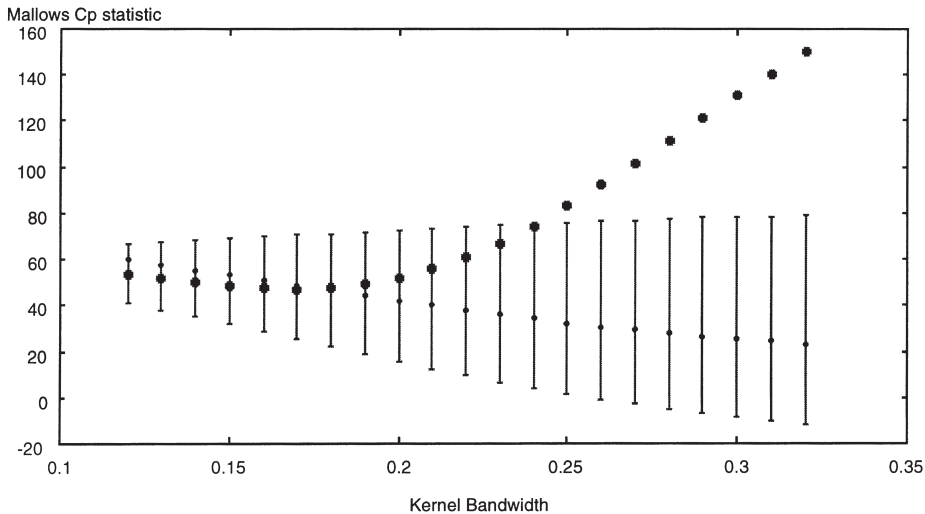

FIGURE 9: Mallows *Cp* Plot for Kernel Bandwidth.

beyond this the choice of whether one wants a very small bandwidth or a larger one with smaller confidence intervals around parameter estimates depends on the situation to which GWR is applied. In cases where low standard error estimates of coefficients in specific (u, v) locations are required, larger bandwidths may be best. However, in cases where the primary aim is cartographic exploration smaller bandwidths may be preferable. In this situation the human eye and brain are reasonable smoothers of 'noisy' map patterns, but are not able to reconstruct genuine features of the data that have been filtered out by GWR before the graphic has been produced.

Spatial Autocorrelation in the Error Term

Although we have considered the GWR model as an alternative to standard spatial autocorrelation models for modeling spatial effects in data, the two models are not mutually exclusive. For example, one could consider a generalization of models (5) or (11) where the distribution of the error terms ε_i are no longer independent, but exhibit spatial autocorrelation. This is an important consideration because nonindependence of the residual terms may cause problems in interpreting the Mallows' C_p statistic proposed earlier and may also lead to undersmoothing in the GWR estimates (the latter phenomenon is discussed in general terms in Hastie and Tibshirani (1990) and Wand and Jones (1995). Two issues should be addressed here: first, techniques for diagnosing the presence of spatial autocorrelation in the residuals of models such as (5) and (11), and second, the calibration of models in which there is spatial autocorrelation.

The first issue may be addressed using either exploratory diagnostics or formal tests. It is anticipated that formal tests may be developed along the lines of those for identifying residual spatial autocorrelation in global regression models (see, for example, Griffith, 1988). One exploratory approach is through the 'semivariogram plot' often used in Kriging (Mallows, 1973). In Kriging, error terms are assumed to be spatially autocorrelated with $\text{Cov}[\varepsilon_i, \varepsilon_j] = C(u_i - u_j, v_i - v_j)$, where (u_i, v_i) and (u_j, v_j) are the geographic coordinates locating observations i and j, and C is a symmetric, nonnegative definite function. Furthermore, $\text{E}(\varepsilon_i) = \text{E}(\varepsilon_i) = 0$. Thus, the autocorrelation is a stationary process depending only on the relative positions of observations i and j. This situation is often further simplified to the model $\text{Cov}[\varepsilon_i, \varepsilon_j] = C(d_{ij})$, with d_{ij} being the distance between observation i and j. This effectively adds an assumption of directional invariance to the correlation structure. Note that under this assumption

$$\text{E}\left[(\varepsilon_i - \varepsilon_j)^2\right] = \tfrac{1}{2}\text{Var}\left(\varepsilon_i - \varepsilon_j\right) = C(0) - C(d_{ij}) = \gamma(d_{ij})$$

Thus, we have a function $\gamma(.)$ linking the expected value of $\varepsilon_i - \varepsilon_j$ to d_{ij}. The task of the semivariogram plot is to identify $\gamma(.)$ by graphing d_{ij} against $\text{E}(\varepsilon_i - \varepsilon_j)$. A more detailed discussion is given in Bailey and Gatrell (1995). Typically, this is done by 'binning' the $n(n-1)/2$ observation pairs into a number of d_{ij} (possibly overlapping) range classes and plotting the central d_{ij} of each class against the

mean value of $(\varepsilon_i - \varepsilon_j)^2$ in that class, although a Loess regression of d_{ij} against $(\varepsilon_i - \varepsilon_j)^2$ for each of the observation pairs is another possibility. The case where the mean value of $(\varepsilon_i - \varepsilon_j)^2$ shows no trends suggests that there is no spatial autocorrelation in the errors, as this corresponds to the case where $E(\varepsilon_i - \varepsilon_j)^2$ does not depend on d_{ij}. In other words the diagnostic plot suggests that $\gamma(d_{ij})$ is constant. Again, Bailey and Gatrell (1995) provide a more detailed discussion of this.

To illustrate, the 'binning' approach was applied to residuals from the GWR model fitted to the house-price data in Section 3. Distance classes of (0.15,0.25], (0.25,0.35], . . . (0.95,1.05] were used as bins. Figure 10 shows a plot as described above. A bar graph was used here because it gives a visual impression of the binning process. In this case, there is no clear trend in the heights of the bars as distance increases, suggesting that a constant function is a reasonable estimate for $\gamma(.)$, and therefore that the residuals from this house-price model show negligible spatial autocorrelation.

The second issue to be considered is that of calibrating GWR models in the case where autocorrelation is present. An initial problem arising here is that of *identifiability*. As discussed in the introduction to this paper, nonstationarity in regression coefficients, and correlation between nearby error terms both give rise to similar phenomena—namely that of spatial clusters of residuals of the same sign when ordinary least squares regression models are fitted. This makes it extremely difficult to determine which of the two effects have caused the observed spatial residual patterns. Indeed, there are situations where Kriging estimates based on stationary models of spatial autocorrelation and nonparametric surface models without autocorrelation using thin-plate splines lead to
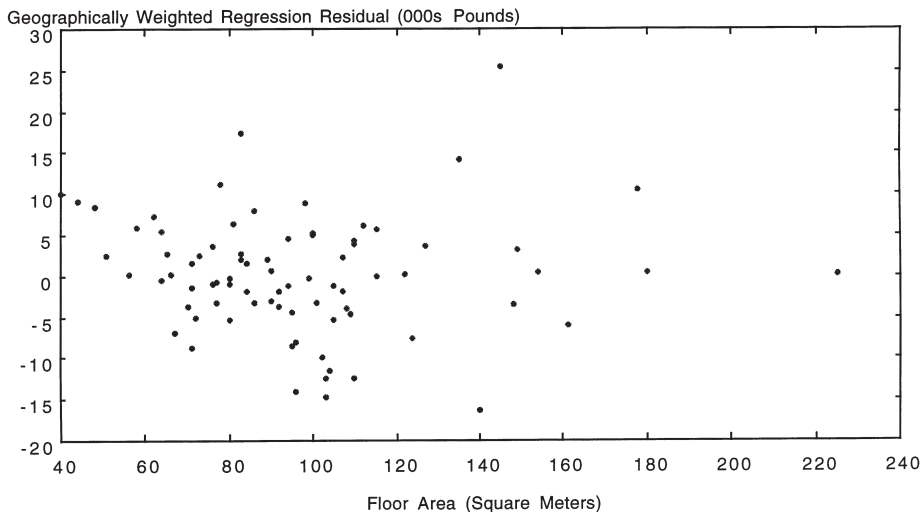

FIGURE 10: Mixed Model Residuals.

the same set of fitted values (Wahba, 1990). In this situation there is complete nonidentifiability between the two models. Despite these problems, Wang (1996) demonstrates how smoothing spline models with correlated random errors may be calibrated. This approach provides one future direction for research into calibrating spatially autocorrelated GWR models.

Another approach is to consider relaxing the stationarity constraint on the spatial autocorrelation and to use geographical weighting to calibrate not only the regression coefficients but also the degree of spatial autocorrelation in the error term. This extra flexibility would allow for situations in which the degree of correlation in the error term varied from place to place. This has been explored by Brunsdon, Fotheringham, and Charlton (1998).

A Vector of Bandwidths

In Section 4 we introduced the notion of mixed GWR and put forward ideas about calibrating this model. One could argue that such mixed models are a special case of a more general model in which the bandwidth for calibrating the coefficient for each variable may be different. One allowable bandwidth value could be infinity—this is equivalent to a globally constant coefficient. In terms of calibration, techniques similar to that of back-fitting may be used. Note that because each individual variable has an associated hat matrix, the back-fitting approach yields a hat matrix for the entire model calibration. Again, a Mallows-like statistic could be constructed, but in this case we replace a bandwidth of scalar k is replaced by a vector $\mathbf{k}$. This, of course, creates new problems. If the dimension of $\mathbf{k}$ is large then graphical techniques such as that used in Figure 9 become difficult to impliment. One compromise may be to use some form of sequential approach in which new variables are added one at a time, each time choosing an appropriate bandwidth. However, the criticisms leveled earlier against stepwise regression may also have some validity here. At some stage in any GWR modeling process it must be considered whether using multiple bandwidths would provide a helpful extension of the simpler approaches discussed earlier in the paper.

Adaptive Bandwidths

A final extension of the GWR method is to allow bandwidths to vary geographically. For example, when considering models in human geography, urban areas may well have more dense clusters of observed cases than rural areas. In this case a similar number of observations for calibrating a GWR may be obtained using a smaller bandwidth in the urban areas. An analogous problem in kernel-based probability density estimation has been considered by Silverman (1986) and Brunsdon (1995). In the GWR case, several methods of developing 'adaptive' kernels are currently being investigated (Fotheringham, Brunsdon, and Charlton, 1998). In many cases, associated hat matrices also exist and so many of the techniques outlined in this paper may again be useful.

7. CONCLUSIONS

In this paper we consider a number of theoretical issues in GWR. An important contribution to GWR is the introduction of analytical techniques to test for spatial nonstationarity as an alternative to the Monte Carlo methods put forward in Brunsdon, Fotheringham, and Charlton (1996). The advance here is twofold: the F-statistic approach increases theoretical understanding of GWR as well as finding a less computationally-intensive means of testing GWR-based hypotheses. Another important advance is the development of *mixed* models—allowing the investigation of *partial* nonstationarity in models. Uniting both of these advances is the use of the *hat matrix* in exploring theoretical issues and practical computations. Hat matrices play a key role in much ordinary linear regression theory, in generalized linear models, and in generalized additive models (see, for example, Hastie and Tibshirani, 1990). These matrices give much insight into all of these techniques, and the finding that hat matrices also play a key role in GWR gives promise of further theoretical advances in this area in the near future. With this increase in the theoretical understanding of GWR, it is hoped that it may be applied in a number of areas and become a valuable modeling tool in spatial data analysis.

REFERENCES

Anselin, Luke. 1993. "SpaceStat: A Program for the Statistical Analysis of Spatial Data," Technical Report 93106–4060, Department of Geography, University of California at Santa Barbara: NCGIA.

Bailey, Trevor C. and Anthony C. Gatrell. 1995. *Interactive Spatial Data Analysis*. Essex: Longman Scientific and Technical.

Brunsdon, Chris. 1995. "Estimating Probability Surfaces for Geographical Points Data: An Adaptive Kernel Algorithm," *Computers and Geosciences*, 21, 877–894.

Brunsdon, Chris, A. Stewart Fotheringham, and Martin Charlton, 1996. "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity," *Geographical Analysis*, 28, 281–289.

———. 1998. "Spatial Nonstationarity and Autoregressive Models," *Environment and Planning A*, 30, 1905–1928.

Casetti, Emilio. 1972. "Generating Models by the Expansion Method: Applications to Geographic Research," *Geographical Analysis*, 4, 81–91.

Chatfield, Chris. 1994. *Problem Solving, A Statistician's Guide.* London: Chapman and Hall.

Cleveland, William S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Cliff, Andrew D. and J. Keith Ord. 1981. *Spatial Processes: Methods and Applications.* London: Pion.

Cressie, Noel. 1991. *Statistics for Spatial Data*. New York: Wiley.

Dobson, Annette J. 1990. *An Introduction to Generalized Linear Models*. London: Chapman and Hall.

Eldridge, John and John-Paul Jones. 1991. "Warped Space: A Geography of Distance-Decay," *Professional Geographer,* 43, 500–511.

Fotheringham, A. Stewart, Chris Brunsdon, and Martin Charlton. 1996. "The Geography of Parameter Space: An Investigation into Spatial Non-Stationarity," *International Journal of Geographical Information Systems*, 10, 605–627.

———. 1998. "Scale Issues and Geographically Weighted Regression," in N. Tate (ed.), *Scale Issues and GIS*. Chichester: Wiley.

Fotheringham, A. Stewart, Martin Charlton, and Chris Brunsdon. 1997. "Two Techniques for Exploring Non-Stationarity in Geographical Data," *Geographical Systems,* 4, 59–82.

———. 1998. Geographically Weighted Regression: A Natural Extension of the Expansion Method for Spatial Data Analysis," *Environment and Planning A*, 30, 1905–1928.

Griffith, Daniel. 1988. *Advanced Spatial Statistics*. Dordrecht: Kluwer.

Haining, Robert. 1990. *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge: Cambridge University Press.

Haslet, John, Graham Wills, and Anthony Unwin. 1990. "SPIDER—An Interactive Statistical Tool for the Analysis of Spatially Distributed Data," *International Journal of Geographical Information Systems*, 4, 285–296.

Hastie, Trevor J. and Robert J. Tibshirani. 1986. "Generalized Additive Models (With Discussion," *Statistical Science*, 1, 297–318.

———. 1990. *Generalized Linear Models*. London: Chapman and Hall.

Jones, John-Paul and Emilio Casetti. 1992. *Applications of the Expansion Method*. London: Routledge.

Kendall, Sir Maurice and Alan Stuart. 1977. *The Advanced Theory of Statistics*. London: Griffin.

Mallows, Colin L. 1973. "Some Comments on C_p," *Technometrics,* 15, 661–667.

Odland, John. 1988. *Spatial Autocorrelation*. London: Sage Publications.

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

Stone, Charles J. 1977. "Consistent Nonparametric Regression," *Annals of Statistics,* 5, 549–645.

Tierney, Luke. 1990. *LISP-STAT: An Object Oriented Environment for Statistical Computing and Dynamic Graphics*. Chichester: Wiley.

Upton, Graham and Bernard Fingleton. 1985. *Spatial Data Analysis by Example, Volume 1: Point Pattern and Quantitative Data*. Chichester: Wiley.

Wahba, Grace. 1990. *Spline Models for Observational Data*. Philadelphia: SIAM.

Wand, Matthew and Chris Jones. 1995. *Kernel Smoothing*. London: Chapman and Hall.

Wang, Yu. 1996. "Smoothing Spline Models with Correlated Random Errors." Technical Report 996, Department of Statistics, University of Wisconsin.