Environment and Planning A 2002, volume 34, pages 733 – 754

DOI:10.1068/a34110

A general framework for estimation and inference of geographically weighted regression models: 1. Location-specific kernel bandwidths and a test for locational heterogeneity

Antonio Páez
Center for Northeast Asian Studies, Tohoku University, Kawauchi, Aoba-ku, Sendai 980-8576, Japan; e-mail: antonio@rs.civil.tohoku.ac.jp

Takashi Uchida
Graduate School of Engineering, Osaka City University, 3-3-138 Sugimoto, Sumiyoshi-ku, Osaka 9558-8585, Japan; e-mail: uchida@civil.eng.osaka-cu.ac.jp

Kazuaki Miyamoto
Center for Northeast Asian Studies, Tohoku University, Kawauchi, Aoba-ku, Sendai 980-8576, Japan; e-mail: miyamoto@plan.civil.tohoku.ac.jp
Received 14 May 2001; in revised form 7 December 2001

Abstract. Geographically weighted regression (GWR) has been proposed as a technique to explore spatial parametric nonstationarity. The method has been developed mainly along the lines of local regression and smoothing techniques, a strategy that has led to a number of difficult questions about the regularity conditions of the likelihood function, the effective number of degrees of freedom, and in general the relevance of extending the method to derive inference and model specification tests. In this paper we argue that placing GWR within a different statistical context, as a spatial model of error variance heterogeneity, or what might be termed locational heterogeneity, solves these difficulties. A maximum-likelihood-based framework for estimation and inference of a general geographically weighted regression model is presented that leads to a method to estimate location-specific kernel bandwidths. Moreover, a test for locational heterogeneity is derived and its use exemplified with a case study.

> "Rather than propose a new theory or unearth a new fact, often the most important contribution a scientist can make is to discover a new way of seeing old theories or facts."
>
> Richard Dawkins *The Selfish Gene*

1 Introduction

Important advances in the analysis of spatial data have been made over the last few years, in general moving from an initial focus on testing for spatial pattern using spatial autocorrelation statistics (Cliff and Ord, 1981), to modeling spatial pattern by means of regression models with spatial components (Anselin, 1988a; Griffith, 1988; Haining, 1990; Cressie, 1991). These early breakthroughs in testing and modeling are characterized by being global and therefore related to the study area as a single, discrete entity. It has been recognized, however, that spatial data frequently exhibit complex patterns that are usually difficult to capture, represent, and explain by using only global statistics and models. In such situations, these yield at best an approximation or 'average' of system-wide variation and behavior.

To improve this situation, a recent trend in spatial analysis has been to stress the local as opposite to the global. Results of this effort, in an exploratory framework, are Getis and Ord's family of $G_i(d)$ statistics of local spatial association (Getis and Ord, 1992; Ord and Getis, 1995; 2001), and Anselin's (1995) local indicators of spatial association (LISA), which are local counterparts of the best-known

global autocorrelation statistics. These statistics are designed to be location specific, a characteristic useful to detect what might be called spatial outliers—hot spots or places where spatial association is particularly intense—and to produce maps of spatial variation. Their use can help the analyst to detect revealing spatial patterns that would otherwise go unnoticed if global statistics were used (for example, see Páez et al, 2001).

In a modeling framework, the expansion method (Casetti, 1972) has been used to obtain local parameter estimates by letting them drift as functions of locational variables, sometimes simultaneously with spatial association modeling (Can, 1992). The expansion method is a form of analysis of the covariance (ANCOVA), but it has been noted that it works essentially by fitting a surface to the data, with its shape depending directly upon the complexity of the expansion parameters (Fotheringham et al, 1998). Arguably, this trend-fitting exercise might miss more complex spatial variation by oversimplifying or 'flattening' it away. To address the issue of complex spatial parametric variation or *spatial nonstationarity*, Fotheringham, Brunsdon, and Charlton (Brunsdon et al, 1996; Fotheringham et al, 1997; 1998) have proposed a simple but powerful method called geographically weighted regression (GWR) that gives all elements and diagnostics of a regression model, such as parameter estimates, goodness-of-fit measures (R^2) and *t*-values on a local basis.

GWR has been developed mainly along lines that parallel developments in the literature on smoothing methods, in particular local likelihood estimation, kernel regression, and locally weighted regression (see, inter alia, Tibshirani and Hastie, 1987; Cleveland and Devlin, 1988). From this perspective, GWR is seen as a locally weighted regression method that operates by assigning a weight to each and every observation i depending on its distance from a specific geographical location o, also called a focal point. The weighting system is based on the well-established concept of distance decay, made operational by means of a kernel function that reduces (that is, down-weights) the influence of distant observations on estimation for location o (which could be any point including i), while implicitly emphasizing the influence of neighboring observations. Take, for instance, the specification of GWR according to Leung et al (2000):

$$y_o = \beta_{o1} + \sum_{k=2}^{K} \beta_{ok} x_{ok} + \varepsilon_o \,, \tag{1}$$

where y is the dependent variable, the xs are independent variables, the βs are parameters and the εs are error terms, with (assumption 1, page 14)

$$\boldsymbol{\varepsilon}_o \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \mathbf{I}) \,. \tag{2}$$

The estimators for this model are given by the geographically weighted expression (in matrix form):

$$\hat{\boldsymbol{\beta}}_o = (\mathbf{X}^{\mathrm{T}} \mathbf{W}_o \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{W}_o \boldsymbol{Y} \,. \tag{3}$$

The above expression is identical to the estimators obtained from weighted least squares, with $\mathbf{W}_o$ as a diagonal matrix of geographic weights w_{oi} that are a decreasing function of distance d_{oi} between the focal point o and location i. A common choice of kernel function is the exponential distance-decay form $w_{oi} = \exp(-\gamma d_{oi}^2)$, although it should be noted that a number of valid functions exist, as, for example, those used in variogram modeling (see Chilès and Delfinier, 1999, pages 80–93). Regardless of functional choice, kernel functions used in GWR depend on a generally unknown constant γ, called the kernel bandwidth, which has to be calibrated (that is, a value of γ has to be determined) by means of a cross-validation procedure in order to make the method operational.

The similarities between GWR and the smoothing methods mentioned above have been emphasized before, and Leung et al, for instance, state that GWR "belongs to the local-regression methodology" (2000, page 14). Elements such as the adoption of a kernel function, the terminology, and in particular the cross-validation procedure used to calibrate the kernel function are perhaps the most evident debts that GWR owes to the smoothing literature. Upon closer inspection, however, the similarities give rise to a number of difficult questions. A problem with the local regression framework above is that it is unclear what the nature of the weights is, and what is the rationale that justifies their introduction, thus giving the impression of an ad hoc weighting scheme to obtain better local fits—perhaps a useful exploratory tool but not a proper statistical model. The picture is further complicated by the practice of using global kernel bandwidths, obtained from a cross-validation procedure, to estimate local regressions. Consequently, any proposal along these lines to bring GWR formally into the domain of modeling (in order to derive significance and model specification tests) eventually has to face questions such as: What is the effective number of parameters estimated? What is the effective number of degrees of freedom? Under what conditions can a regular likelihood function be derived for this model? Although some issues pertaining inference of local models are starting to be addressed in the smoothing literature (see Fan et al, 1998), it is not clear that these developments will carry over placidly to the spatial case, thus making the attempt of developing GWR into a formal model a difficult endeavor at best. Or so it seems.

In this paper we argue that most of the apparent difficulties on the way to generalizing GWR as a spatial statistical model can be solved by looking at the method from a different statistical perspective. Previous efforts to accommodate GWR into the mould of local regression methods have found a similar situation: the fit is not perfect and the details are troublesome. It is therefore our objective to propose a change of perspective for GWR. We concentrate on drawing the parallels between GWR and thoroughly studied *error variance heterogeneity* models and the weighting methods used to solve them, such as weighted least squares (WLS). Ultimately, we show how GWR can be seen as a model of error variance heterogeneity, with heterogeneity having a precise geographical interpretation—namely, how it is seen from a particular geographic location, an effect that we term *locational heterogeneity*. In this way, we propose a change of emphasis from modeling parametric nonstationarity to modeling locational heterogeneity that may result in parametric variation across space. The conceptual shift reveals itself fruitful and it will be seen that it helps to make all pieces fall in place: as a result of it the apparent ambiguity regarding the number of parameters being estimated vanishes; the geographical weights, instead of being an ad hoc device, are found to be part of a model of the variance; and the model is then shown to be an investigation into a spatial form of error variance heterogeneity.

Two situations of interest are discussed. First, there is the case when the kernel bandwidth is a known constant (that is, the covariance matrix is known except for a constant), a situation equivalent to the typical GWR method as presented in the literature. And second, there is the more realistic case when the variance parameters (variance and kernel bandwidth) are unknown. The latter situation naturally leads to a method to estimate location-specific kernel bandwidths and to a test for locational heterogeneity, which is in fact a model specification test for heteroscedasticity. Application of the method to data from Sendai City in Japan shows that locational heterogeneity can be effectively modeled and tested to gain insights about the spatial characteristics of the problem.

2 The model

Consider the following linear model:

$$y_i = \beta_1 + \sum_{k=2}^{K} \beta_k x_{ik} + \varepsilon_i \,, \tag{4}$$

or expressed in compact matrix form:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \,. \tag{5}$$

In this model, $\boldsymbol{Y}$ is an $n \times 1$ vector of objective variable observations collected for a set of points or zones i, $i = 1, 2, ..., n$, and $\mathbf{X}$ is an $n \times K$ matrix of explanatory variables that include the usual constant term; $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters corresponding to K explanatory variables; and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of unobservable random error terms. A common assumption regarding the error terms is that they are independently normally distributed with constant variance (IID normal):

$$\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\mathbf{I}) \,. \tag{6}$$

There are many situations in diverse fields, such as econometrics, pharmacology, experimental design, and of course spatial econometrics and statistics, in which the assumption of constant variance or *variance homogeneity* in the errors is untenable, and corrective action might be required to ensure an efficient analysis. In other cases modeling the variance may be itself of interest (Box and Meyer, 1986). A central concern in spatial data analysis is that the unobservable underlying process might not be uniform across space. Analyzing land prices in a monocentric city, for instance, we have found that the determination process appears to operate differently in the neighborhood of the central business district than in the suburbs, and as a consequence, it is more efficiently described by region-specific parameters (Páez et al, 2001). What is needed is a model formulation that relaxes the above assumption.

2.1 Variance functions

A more general model that does not assume constant variance is obtained by defining the error terms in model (4) as follows:

$$\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega}) \,, \tag{7}$$

where $\boldsymbol{\Omega} = \sigma^2\mathbf{G}$ is a general covariance matrix with elements (see Davidian and Carroll, 1987):

$$\omega_{ii} = \sigma^2 \mathrm{g}_i(\boldsymbol{\gamma}, \boldsymbol{z}_i), \quad \text{and} \quad \omega_{ij} = 0, \qquad \forall i, j, \quad i \neq j \,. \tag{8}$$

This formulation allows us to model the variance corresponding to observation i as a function of a $p \times 1$ vector of known variables $\boldsymbol{z}_i$ associated with parameter vector $\boldsymbol{\gamma}$ ($p \times 1$), and a variance parameter σ^2. To ensure regularity conditions, it is assumed that $\omega_{ii} > 0$ for all i, and that ω_{ii} is a twice-differentiable function of σ^2 and $\boldsymbol{\gamma}$. The natural alternative is a model of variance homogeneity, defined by further assuming that there is a unique set of values $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ for which the function $\mathrm{g}_i(\boldsymbol{\gamma}^*, \boldsymbol{z}_i) = 1$ for all i. This of course means that $\boldsymbol{\Omega}$ reduces to the classic situation of constant variance ($\boldsymbol{\Omega} = \sigma^2\mathbf{I}$).

A number of functions have been proposed to model the variance. For instance, Davidian and Carroll (1987) consider, among other forms, a model of the variance that is quadratic in the explanatory variables ($p = 2K$):

$$\mathrm{g}_i(\boldsymbol{\gamma}, \boldsymbol{z}_i) = 1 + \sum_{p=1}^{k} \gamma_p x_{ip} + \gamma_{p+k} x_{ip}^2 \,. \tag{9}$$

This model suggests a spatial model of the variance that takes the form of a quadratic trend surface, if we let the known vector $\boldsymbol{z}_i$ (6×1) be the coordinates (c_x, c_y), squared coordinates, and $c_x - c_y$ interactions of location i:

$$g_i(\boldsymbol{\gamma}, \boldsymbol{z}_i) = 1 + \gamma_1 c_{xi} + \gamma_2 c_{yi} + \gamma_3 c_{xi}^2 + \gamma_4 c_{xi} c_{yi} + \gamma_5 c_{yi}^2 . \tag{10}$$

In this case $p = 5$ and the null hypothesis of variance homogeneity is given by $\boldsymbol{\gamma} = \boldsymbol{\gamma}^* = \boldsymbol{0}$. This variance function, however, can be subjected to the same criticism leveled at the expansion method, because the complexity of the trend surface depends on the choice of the variables, and moreover, the use of complex parameter and variable combinations can lead to interpretability problems.

Other authors have studied variance models of the exponential form (Box and Meyer, 1986; Verbyla, 1993; Lyon and Tsai, 1996):

$$g_i(\boldsymbol{\gamma}, \boldsymbol{z}_i) = \exp(\boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\gamma}) . \tag{11}$$

Again, the null hypothesis of homogeneity is given by $\boldsymbol{\gamma} = \boldsymbol{\gamma}^* = \boldsymbol{0}$. It is interesting to note that Cook and Weisberg, in a paper that discusses diagnostics for error variance heterogeneity of this form, comment that "when heteroscedasticity occurs, the variance may often depend on the values of one or more explanatory variables or on additional relevant quantities such as time and *spatial ordering*" (1983, page 1; our emphasis). So far, however, we are not aware of any spatial application that makes use of variance functions of this kind, although in a spatial context exponential functions have been used to model the autocovariance in geostatistics (Chilès and Delfinier, 1999). As previously noted, the exponential form is not the only valid possibility, and in fact any function that is twice differentiable in the parameters and produces a positive definite covariance matrix constitutes a valid candidate. In what follows, we adopt equation (11) for the model of GWR.

2.2 Geographically weighted regression

Let us define GWR as a model linear in the mean [see equation (1)]:

$$y_o = \beta_{o1} + \sum_{k=2}^{K} \beta_{ok} x_{ok} + \varepsilon_o ,$$

but now assume that the error terms are distributed as in equation (7) and define the covariance as a diagonal matrix $(n \times n)$:

$$\boldsymbol{\Omega}_o = \mathrm{E}(\boldsymbol{\varepsilon}_o \boldsymbol{\varepsilon}_o^{\mathrm{T}}) = \sigma_o^2 \mathbf{G}_o , \tag{12}$$

with the diagonal elements of $\boldsymbol{\Omega}_o$ as

$$\omega_{oi} = \sigma_o^2 g_{oi}(\boldsymbol{\gamma}, \boldsymbol{z}_i) = \sigma_o^2 \exp(\gamma_o d_{oi}^2) . \tag{13}$$

In this case, the covariance matrix is a particular form of equation (8) with $p = 1$. The variance is thus defined as a function of two parameters, namely σ_o^2 and γ_o. Clearly, the ith diagonal element of the covariance matrix $\omega_{oi} > 0$ as long as $\sigma_o^2 > 0$ (the usual nonnegativity condition of the variance), and ω_{oi} is a twice-differentiable function of σ_o^2 and γ_o.

In the above function, known variable $\boldsymbol{z}_i$ is the square of d_{oi}, the distance between a focal point o and location i for which an observation has been recorded, and subindex o is adopted to indicate that this is a model of error variance heterogeneity as seen from a specific geographical perspective. This local perspective of the variance is what we call locational heterogeneity. By convention, all parameters and statistics obtained from estimation at o are assigned to this geographical location, thus making GWR a locally linear model. It should be noted that each local model is completely defined by

the $(2 + K)$-parameter vector $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o^{\mathrm{T}}, \sigma_o^2, \gamma_o]^{\mathrm{T}}$ in exactly the same way that a model with heterogeneous variance is complete, the only difference being that the origin of the explanatory variable (that is, distance) is relocated to conduct local estimation. The underlying model of homogeneity is given by $\gamma_o = \gamma_o^* = 0$, in which case the model reduces to the usual constant variance situation and what might be termed the *global model* [equation (4)] because the variance does not depend on location, and consequently neither do other parameters. By allowing the possibility that variance varies in a given geographical direction we define a model of heteroscedasticity that is at the same time simple and that has a clear spatial interpretation [nongeographical models allow for the possibility that the variance increases in a given direction in *variable space* (see Cook and Weisberg, 1983)].

The power of adopting for estimation a specific geographical location GWR-style is that it allows us to move from a global, 'whole picture' perspective [as would be the case, for example, when using a variance function such as equation (10)], to a local perspective that gives us a 'field-level' view of the problem. In this sense, an important characteristic of GWR is that it is not limited to estimation at the location of recorded observations. Location o need not correspond to any given location i $(i = 1, ..., n)$, an observation at o is not indispensable for estimation, and in general any number of models can be estimated $(o = 1, ..., m)$, with $m \geqslant 1$ and no implied relationship between m and the number of observations n. The method can then be seen as the experiment of observing the same dataset from a number of different perspectives—meaning that we enjoy the advantage of being able to study the parametric 'landscape' from a number of different vantage points. In practice, models can be estimated and tested on a location-specific basis, for as few as a single location of interest, and parametric nonstationarity can be studied when locational heterogeneity exists and $m > 1$.

3 Estimation

In this section we discuss estimation of the proposed model. According to Carroll and Rupert (1988, page 11), when modeling the variance it is useful to distinguish two general cases depending on whether parameter vector $\boldsymbol{\gamma}$ is known or not (two other cases result when the variance is also a function of the mean response, but these are not of direct interest in the present context). For the model at hand, the two relevant cases, discussed below, regard our knowledge of parameter γ_o.

3.1 Case 1: parameter γ known

The simplest estimation case results when parameter γ_o is known. In current GWR practice this does not mean actual a priori knowledge of its value, but rather, for lack of other methods, the application of a kernel bandwidth obtained from cross-validation (described below), assumed to be constant across the study area (that is, γ after dropping the location-specific subindex o). This is equivalent to assuming GWR as having heterogeneous error terms and a structure of the covariance known except for a constant, which means that the model can be estimated by means of weighted least squares (see, for example, Carroll and Rupert, 1988). WLS leads to the following expression for the parameters:

$$\hat{\boldsymbol{\beta}}_o = (\mathbf{X}^{\mathrm{T}}\mathbf{G}_o^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{G}_o^{-1}\boldsymbol{Y}. \tag{14}$$

As before, $\hat{\boldsymbol{\beta}}_o$ is $(K \times 1)$, $\mathbf{X}$ $(n \times K)$, $\boldsymbol{Y}$ $(n \times 1)$, and $\mathbf{G}_o$ $(n \times n)$. Equation (14) makes clear the inverse relationship between the kernel function behind the matrix of geographical weights in equation (3) and the exponential component of our variance function. The problem of selecting a kernel bandwidth to calculate the weights is solved by means of

a least squares criterion that aims at minimizing a cross-validation score of the form (Brunsdon et al, 1996):

$$C(\gamma) \;=\; \sum_{i=1}^{n} [\, y_i - \hat{y}_{(i)}(\gamma)]^2 \,. \tag{15}$$

This score is simply a measure of how well the local models fit the data. In equation (15) y_i is the value of the independent variable at $i = 1, ..., n$, whereas $y_{(i)}$ is the value predicted by a model estimated for the same location by using a given γ, after removing observation y_i from the sample to avoid perfect 'fits' that disregard all but one observation.

The above procedure has been amply discussed in the cited GWR references, and is the method commonly used to obtain parameter γ. In our view, however, it suffers from some important shortcomings. First, the method is extremely computer intensive, and thus expensive when dealing with large samples. Moreover, we have found instances where it fails to produce a definite minimum. Second, as mentioned above, obtaining and using a global kernel bandwidth implies the restrictive and unrealistic assumption that the rate of decay of the kernel function is constant across the study area. Third, and perhaps more importantly, it is not clear, having no references to compare, whether kernel bandwidths obtained in this fashion are accurate enough as to warrant their use as 'known' parameters of the variance. A more flexible approach that overcomes the above shortcomings is discussed next.

3.2 Case 2: parameter γ unknown

In practice, the structure of the covariance matrix is seldom known. Clearly, the assumption that parameter γ is known is a simplification needed to make the method operational under the perspective of locally weighted regression. In more realistic situations, from the viewpoint of a model of the variance, it can be as untenable as the assumption of variance homogeneity. The obvious solution to this problem is to estimate γ_o on a local basis as a parameter of function (13), in which case maximum likelihood appears as a more appropriate method of estimation. The log-likelihood of the model is given by the following expression:

$$L_o \;=\; -\frac{n}{2}\ln \pi - \frac{n}{2}\ln \sigma_o^2 - \frac{1}{2}\ln |\mathbf{G}_o| - \frac{1}{2\sigma_o^2}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_o)^{\mathrm{T}}\mathbf{G}_o^{-1}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_o)\,. \tag{16}$$

The parameter vector is $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o^{\mathrm{T}}, \sigma_o^2, \gamma_o]^{\mathrm{T}}$, and the number of parameters is $2 + K$. The above formulation is similar to a spatially autoregressive model in the sense that a more general form of the covariance matrix is used to introduce spatial effects (Haining, 1990). The main difference with said autoregressive models is that in the present case the covariance matrix is diagonal, and thus the error terms are assumed to be independent. The distinction is important because in a spatially autoregressive model the direct effect is on the means, which become a weighted average of the global and local means, whereas in the case of equation (16) the effect being modeled is directly on the variance.

3.2.1 *Estimators and concentrated likelihood*

To conduct estimation by maximum likelihood, the first-order conditions needed to calculate parameter estimates are first obtained by taking partial derivatives of the log-likelihood (16) with respect to parameter vector $\boldsymbol{\theta}_o = [\boldsymbol{\beta}_o^{\mathrm{T}}, \sigma_o^2, \gamma_o]^{\mathrm{T}}$. The resulting vector of partial derivatives, or score vector $\boldsymbol{s}$, is set to zero as follows:

$$\boldsymbol{s} \;=\; \frac{\partial L_o}{\partial \boldsymbol{\theta}_o} \;=\; \boldsymbol{0}\,. \tag{17}$$

The elements of the score vector result from taking partial derivatives with respect to each element of the parameter vector:

$$\boldsymbol{s}_{\beta} = \frac{\partial L_o}{\partial \boldsymbol{\beta}_o} = \frac{1}{\sigma_o^2}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_o)^{\mathrm{T}}\mathbf{G}_o^{-1}\mathbf{X}, \tag{18}$$

$$s_{\sigma^2} = \frac{\partial L_o}{\partial \sigma_o^2} = -\frac{n}{2\sigma_o^2} + \frac{1}{2\sigma_o^4}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_o)^{\mathrm{T}}\mathbf{G}_o^{-1}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_o), \tag{19}$$

$$s_{\gamma} = \frac{\partial L_o}{\partial \gamma_o} = -\frac{1}{2}\mathrm{tr}\mathbf{D}_o + \frac{1}{2\sigma_o^2}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_o)^{\mathrm{T}}\mathbf{G}_o^{-1}\mathbf{D}_o(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_o), \tag{20}$$

where tr is the trace operator and $\mathbf{D}_o$ is a diagonal $n \times n$ matrix:

$$\mathbf{D}_o = \begin{bmatrix} d_{o1}^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & d_{on}^2 \end{bmatrix}. \tag{21}$$

Based on the first-order conditions (18) and (19), it is simple to show that the estimators for this model are given by the following expressions:

$$\hat{\boldsymbol{\beta}}_o = (\mathbf{X}^{\mathrm{T}}\mathbf{G}_o^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{G}_o^{-1}\boldsymbol{Y}, \tag{22}$$

and

$$\hat{\sigma}_o^2 = \frac{1}{n}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_o)^{\mathrm{T}}\mathbf{G}_o^{-1}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_o). \tag{23}$$

When $\gamma_o = \gamma_o^* = 0$ then $\mathbf{G}_o = \mathbf{I}$ and the expressions reduce to the ordinary least squares (OLS) estimators of the global model. These estimators, when substituted and introduced into log-likelihood function (16), result in a concentrated likelihood that is a function of only one variable, namely parameter γ_o:

$$L_o^{\mathrm{C}} = -\frac{n}{2}\ln\left[\frac{1}{n}(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_o)^{\mathrm{T}}\mathbf{G}_o^{-1}(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right] - \frac{1}{2}\sum_{i=1}^{n}\gamma_o d_{oi}^2. \tag{24}$$

The above function can be maximized in γ_o to obtain a location-specific parameter that is then substituted in equations (22) and (23) to obtain maximum likelihood estimates for $\hat{\boldsymbol{\beta}}_o$ and $\hat{\sigma}_o^2$.

4 Inference

Inference and hypothesis testing are based on the asymptotic variance matrix (AVM) of the model. The elements of the information matrix $\mathbf{F}$ are obtained as the negative expected values of the second partial derivatives of the various parameter combinations, after applying the following expressions that result from the structure of the error terms:

$$\mathrm{E}(\boldsymbol{Y}) = \mathbf{X}\boldsymbol{\beta}_o, \tag{25}$$

and

$$\mathrm{E}(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}) = (\mathbf{X}\boldsymbol{\beta}_o)(\mathbf{X}\boldsymbol{\beta}_o)^{\mathrm{T}} + \sigma_o^2\mathbf{G}_o. \tag{26}$$

The element $\mathbf{F}_{\beta\beta}$ of the information matrix, for instance, is obtained as follows from the second derivative of the likelihood function with respect to $\boldsymbol{\beta}_o$:

$$\frac{\partial^2 L_o}{\partial \boldsymbol{\beta}_o \partial \boldsymbol{\beta}_o^{\mathrm{T}}} = -\frac{1}{\sigma_o^2}\mathbf{X}^{\mathrm{T}}\mathbf{G}_o^{-1}\mathbf{X}. \tag{27}$$

Because the individual elements of the information matrix are the negative, expected values of the second derivative, in this case equation (27), it follows that

$$\mathbf{F}_{\beta\beta} = -\mathrm{E}\left(\frac{\partial^2 L_o}{\partial \boldsymbol{\beta}_o \partial \boldsymbol{\beta}_o^{\mathrm{T}}}\right) = -\frac{1}{\sigma_o^2}\mathbf{X}^{\mathrm{T}}\mathbf{G}_o^{-1}\mathbf{X}\,. \tag{28}$$

The rest of the elements of the information matrix can be shown to be:

$$\boldsymbol{F}_{\sigma^2\beta} = \boldsymbol{0}\,, \tag{29}$$

$$F_{\sigma^2\sigma^2} = \frac{n}{2\sigma_o^4}, \tag{30}$$

$$\boldsymbol{F}_{\gamma\beta} = \boldsymbol{0}\,, \tag{31}$$

$$F_{\gamma\sigma^2} = \frac{1}{2\sigma_o^2}\mathrm{tr}\mathbf{D}_o\,, \tag{32}$$

$$F_{\gamma\gamma} = \frac{1}{2}\mathrm{tr}\mathbf{D}_o^2\,. \tag{33}$$

Then the information matrix can be shown to be block diagonal:

$$\mathbf{F}_{\theta\theta} = \begin{bmatrix} \mathbf{F}_{\beta\beta} & & \\ \boldsymbol{0} & F_{\sigma^2\sigma^2} & \\ \boldsymbol{0} & F_{\gamma\sigma^2} & F_{\gamma\gamma} \end{bmatrix}. \tag{34}$$

The asymptotic variance matrix $\mathbf{V}$ is obtained as the inverse of the information matrix ($\mathbf{V} = \mathbf{F}_{\theta\theta}^{-1}$), substituting the estimates that result from maximizing the concentrated likelihood (24). The diagonal elements of $\mathbf{V}$ become the variance of the corresponding parameters, and from these, the standard errors and the *t*-values needed to test the significance of individual parameters can be obtained as usual. In addition, the information matrix can be used to test for spatial nonstationarity (locational heterogeneity) as discussed in the following section.

5 A test for locational heterogeneity

Hypothesis testing, and in particular testing for spatial parametric nonstationarity, has been an important concern in the application of GWR since the inception of the method. The relevant question underlying this concern is (Brunsdon et al, 1996):

> Does the GWR model describe the data significantly better than a global regression model?

The question is important because it goes to the heart of the validity of the method. What it asks in effect is: are the benefits of estimating a more complex set of local models significantly larger than the costs? Why should we prefer GWR rather than a simpler, global model? To address this fundamental matter, Brunsdon et al (1996) proposed a Monte Carlo simulation method that attempts to assess whether the (global) kernel bandwidth γ obtained from a cross-validation procedure is significantly different from zero. It was reasoned that, as a bandwidth of zero implies a global model, departures from this value would constitute an indicator of spatial nonstationarity. The corresponding hypothesis would be:

$$\mathrm{H}_0\colon \gamma = 0\,,$$

which from the viewpoint of the model introduced in section 2.2. is equivalent to testing for locational heterogeneity. Rejection of the above hypothesis would mean that GWR is a better way to model the data.

Leung et al (2000) commenting on the above approach noted that, in addition to being computationally cumbersome, a Monte Carlo simulation produces results that

are limited to a given sample and therefore have low generalization value beyond a specific application. To improve this state of affairs concerning a question central to the validity of the method, they proposed a test based on the model given by equations (1)–(3), and aimed at solving the problem in the context of classical hypothesis testing. An important conceptual difficulty with their proposal, however, is that it parts from the premise that GWR is a locally weighted regression method.

Seen from a different perspective, once it has been shown that GWR can be seen as a model of locational heterogeneity, the specification used to derive the test seems rather suspect (for example, if the variance is constant what is the nature of the geographical weights?), and consequently it is not clear what is being tested. A second problematic issue is that the tests derived by Leung et al (2000) are built, as is the Monte Carlo procedure proposed by Brunsdon et al (1996), upon the restrictive assumption that the kernel bandwidth is constant across the study area. From the perspective given by a model of the variance, it seems clear that, in principle, a global parameter γ obtained from a cross-validation procedure is not testable, because it is applied under the assumption that it is a known constant. A more natural and direct approach to test for spatial nonstationarity, discussed in the following section, is to test parameter γ_o in a location-specific basis as a component of variance function (13).

5.1 A Lagrange multiplier approach to test for locational heterogeneity

A formal way of testing a model parameter, based on the maximum likelihood approach presented here for the case of unknown kernel bandwidths, makes use of the first-order conditions and information matrix of the GWR model with error covariance matrix defined as in equations (12)–(13). The method, known as the Lagrange multiplier (LM) test, was studied by Breusch and Pagan (1979; 1980) and Cook and Weisberg (1983), and was later extended by Anselin (1988b) to be used in spatial econometric modeling. This framework is attractive because the test can be evaluated by using results from the basic (restricted) model, and it does not require estimation of the more complex model.

To define the test, first consider a partitioning of the parameter vector, to isolate the parameter of interest as $[\gamma_o \,|\, \boldsymbol{\beta}_o^{\mathrm{T}}, \sigma_o^2]$. The relevant hypothesis is

$$\mathrm{H}_0\colon \gamma_o = \gamma_o^* = 0\,, \tag{35}$$

which implies homogeneity and consequently spatial parametric stationarity. The LM test, based on a constrained maximization of the log-likelihood, evaluates the cost of incrementing the number of parameters in the model. In the present case, the approach starts with a restricted model of variance homogeneity (the global model with $\boldsymbol{\Omega} = \sigma^2\mathbf{I}$), and evaluates the cost of moving towards a full model of locational heterogeneity that includes parameter γ_o. If the test does not reject hypothesis (35) within a given critical range α (usually 0.05) this indicates that the cost of introducing parameter γ_o is too high to be consistent with the data, thus favoring the hypothesis of spatial homogeneity. The test, distributed as χ^2 with 1 degree of freedom is (Engle, 1984)

$$\xi = s_{\gamma^*}^2 F_{\gamma^*\gamma^*}^{-1}\,. \tag{36}$$

In this case, the appropriately partitioned score vector is given by the first-order condition corresponding to parameter γ_o evaluated under the null hypothesis ($\gamma_o^* = 0$):

$$s_{\gamma^*} = \frac{1}{2}\left[\frac{1}{\sigma_o^2}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_o)^{\mathrm{T}}\mathbf{D}_o(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_o) - \mathrm{tr}\mathbf{D}_o\right]. \tag{37}$$

Information matrix (34) is block diagonal, and the incumbent element is given by the corresponding element of the partitioned inverse, in this case:

$$F^{-1}_{\gamma^*\gamma^*} = 2\left[\mathrm{tr}\mathbf{D}_o^2 - \frac{1}{n}(\mathrm{tr}\mathbf{D}_o)^2\right]^{-1}, \tag{38}$$

also evaluated under the null hypothesis of $\gamma_o^* = 0$. Parameters $\boldsymbol{\beta}_o$ and σ_o^2 are set to their maximum likelihood estimates—which in this case are identical to the global ordinary least squares estimates. Analytically, the LM test for spatial nonstationarity becomes:

$$\xi = \frac{1}{2}\left[\frac{1}{\hat{\sigma}^2_{\mathrm{OLS}}}(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})^{\mathrm{T}}\mathbf{D}_o(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}) - \mathrm{tr}\mathbf{D}_o\right]^2\left[\mathrm{tr}\mathbf{D}_o^2 - \frac{1}{n}(\mathrm{tr}\mathbf{D}_o)^2\right]^{-1}. \tag{39}$$

It is easy to verify that the above is a special form of the score test for heteroscedasticity derived by Cook and Weisberg (1983) with $p = 1$. The small sample behavior of the test was investigated by Cook and Weisberg (1983) who found that using the χ^2 approximation leads to a conservative test, which they consider to be adequate for diagnostic purposes. More recently, Lyon and Tsai (1996) compared a number of tests for heteroscedasticity and found that the LM test performed well under a number of circumstances. In addition, they found that the test retains its power in small samples and that in general it [and a modified version based on the residual likelihood (Verbyla, 1993)] is robust against high leverage points. Although it is clear from their paper that the LM test is not the only way to test the specification of the model, it is without doubt the most convenient in terms of ease of application and performance. Application of the test derived here is exemplified in section 7.

6 Simultaneous significance tests

The preceding discussion shows that the conceptual shift from local regression to locational heterogeneity modeling helps to dispel the apparent ambiguity regarding the number of parameters being estimated for a given model, and consequently the number of degrees of freedom for inference and hypothesis testing. The LM test of section 5, for instance, is χ^2 with 1 degree of freedom for any given location, regardless of whether models are estimated for one, two, or more locations of interest. A different issue arises, however, whenever the number of models estimated exceeds one ($m > 1$), because in that case we face the problem of *simultaneously* testing m hypotheses, which could be of locational heterogeneity, parameter significance, etc. Succinctly stated, the problem consists in giving protection to the group of statements through group error rates when a family of dependent or independent hypotheses is simultaneously tested (Miller, 1981).

A number of procedures are available that provide the kind of protection needed when conducting simultaneous inference. Many of them, however, require knowledge of the joint distribution and correlation structure of the tests (Miller, 1981), which clearly is not the present case. Fortunately, a simple procedure based on the Bonferroni probability inequality can be applied to our problem. The Bonferroni procedure (in so-called single-step tests) consists in dividing the overall significance level α (group error tolerance) by the number of observations n to obtain an adjusted significance level α/n (individual hypothesis error tolerance). This expeditious procedure is a valuable tool in simultaneous inference because it does not require distributional assumptions and is simple to use (Simes, 1986). These advantages, however, come at the price of lack of power and overconservative results when the tests are highly correlated (that is, not independent), as discussed by Simes (1986). The problem in the case of spatial data is that the tests are likely to be highly dependent, as discussed by Ord and Getis (1995; 2001) and Anselin (1995) in the context of local statistics of spatial association.

In the case of the proposed model, the main issue is heterogeneity. Suppose that the model is perfectly homogeneous (that is, $\gamma_o = 0$ for all o). In this case, irrespective of the presence of spatial autocorrelation, every local model becomes identical to the global model, and thus we would find ourselves testing the same model repeatedly. Clearly, the use of the crude α/n adjustment would be overly stringent in this situation (when dependence is perfect, the denominator in the adjustment should be 1), thus underlining the need for a method to improve the power of the procedure for highly correlated tests. Fortunately, methods exist that achieve this objective by introducing multistep rejective schemes, as opposed to the typical single-step procedure. Liu (1996) provides a recent review of relevant methods, and it is one of these that we adopt for simultaneous inference.

The procedure selected is of the step-down type to conduct multiple tests of significance and operates in the following fashion. Consider the problem of simultaneously testing a family of m null hypotheses $H_1, ..., H_m$ and let $P_1, P_2, ..., P_m$ be the p-values (the probability of making a type-I error) corresponding to hypotheses $H_1, ..., H_m$ (Liu, 1995).

In our case, m is the number of models being estimated. One family of hypotheses of interest is:

$$\begin{aligned} &H_1: \gamma_1 = 0, \\ &\vdots \\ &H_m: \gamma_m = 0. \end{aligned} \tag{40}$$

Because the test statistic for the above family of hypotheses is χ^2 distributed with 1 degree of freedom, the probability of making a type-I error for any given statement is:

$$P_i = 1 - F(\xi_i | v). \tag{41}$$

In the above expression, F is simply the cumulative distribution function, in this case χ^2, evaluated at ξ_i with $v = 1$ degrees of freedom. The test based on m constants $0 < l_1 \leqslant ... \leqslant l_m < 1$ would proceed as follows. Let $P^{(1)} \leqslant ... \leqslant P^{(m)}$ be the ordered p-values (41) and $H^{(1)}, ..., H^{(m)}$ the corresponding hypotheses. Start with the smallest p-value. If $P^{(1)} > l_1$ then stop testing and fail to reject all hypotheses; otherwise reject $H^{(1)}$ and proceed to the next step. In general, if the test has continued to the ith step and if $P^{(i)} > l_i$, then stop testing and fail to reject all remaining hypotheses; otherwise reject $H^{(i)}$ and proceed to step $i + 1$.

A constant l_i that satisfies the Bonferroni inequality is:

$$l_i = \frac{\alpha}{n - i + 1}, \tag{42}$$

where α is the nominal significance level, usually 0.05. The test controls strongly the type-I family-wise error rate at α, which means that the probability of a false rejection of any hypothesis is at most α regardless of how many and which hypotheses are true (note that this procedure is to conduct multiple tests of significance and cannot be used to construct confidence intervals). In this way the acceptance region is at most the nominal error tolerance, but observations with small probabilities are more heavily penalized. The same procedure can be used to test the significance of individual parameters β_{ok} by defining the appropriate family of hypotheses as follows:

$$\begin{aligned} &H_1: \beta_{1k} = 0, \\ &\vdots \\ &H_m: \beta_{mk} = 0. \end{aligned} \tag{43}$$

In this case, the test statistic is t distributed with $n-K$ degrees of freedom. If a two-tailed test is conducted, then the nominal level of significance becomes $\alpha/2$.

On a final note regarding the application of the above procedure, if at least some of the local models within the study area turn out to be homogeneous ($\gamma_o = 0$) and thus identical to the global model, the procedure is bound to be somewhat over conservative as a consequence of the increased number of perfectly dependent tests. The loss of power, however, is not likely to be nearly as serious as that resulting from the application of a crude α/n adjustment all throughout.

7 Application: urban heat island in Sendai, Japan

7.1 Data and global model

In what follows we apply the results obtained above regarding estimation and inference of GWR models, comparing the method under the two cases discussed in section 3, and applying the test for locational heterogeneity of section 5. To this end we use data that are part of a study of the effects of urbanization on the thermal anomaly known as an *urban heat island* in Sendai City, Japan (Páez et al, 1998). We first note that the purpose of this example is not to offer a comprehensive analysis of the problem, but rather to illustrate a number of points that we consider to be of interest in the application of the GWR method.

Before proceeding, we briefly describe the problem. The urban heat island is a meterological phenomenon that accompanies urbanization-related changes in surface cover, reduced humidity, and altered wind patterns, in combinations that result in a trend towards higher air temperatures. When large temperature differences appear, the city becomes separated from the general heat shield, thus resulting in an urban heat island episode. In a city such as Sendai, the temperature differential between city and countryside during a heat island episode can be as large as 6 °C. The controls of the urban heat island are both meteorological and physical, but from a planning viewpoint the interest is in a few controls that explain the presence of the city and its physical characteristics. Among these, city size, building density, and land-use distribution are prominent and frequently cited in the literature as factors to explain the heat island effect.

The variable that we aim to explain is the temperature (T in Kelvin) at different locations of the city during a heat island episode. The observed values for this variable (which range between 298 K and 303.5 K) are the result of a study conducted for Sendai City, and correspond to a summer afternoon. In addition to temperature, variables are introduced to represent the physical characteristics of the city. First, a physical measure of city size is considered, given by the natural logarithm of the distance from the center of the city to the location of the observation (LDIST). In addition, three land-use variables are introduced, in terms of the percentage of cover by zone, to represent three main types of use: commercial (C%), industrial (I%), and residential (R%) land uses. These variables are meant to indicate both land-use intensity by type, and the spatial distribution of uses, and are obtained from the Basic Planning Compendium for Sendai Metropolitan Area.

Using the above variables, a global model is estimated under the assumption of variance homogeneity, to produce the following results:

$$\hat{T} = 308.537 - 0.887\text{LDIST} + (0.0021\text{C}\%) + (0.0056\text{I}\%) + 0.0233\text{R}\% \,. \tag{44}$$

Parameters are significant at the 0.05 level, with the exception of those related to commercial and industrial land-use variables (in parentheses), and the performance of the model is relatively good considering the complexity of the problem, with an adjusted coefficient of determination $R^2 = 0.418$, and $\sigma^2 = 1.148$. Moreover, the

signs are intuitively correct, with temperature decreasing with distance from the center of the city, and more intense uses of land positively related to temperature. Good statistical practice requires us at this point to diagnose results for variance homogeneity and spatial autocorrelation among other factors (see, for example, Griffith and Layne, 1999). Presently, we focus on the study of *spatial* variance homogeneity (that is, locational heterogeneity) in order to apply the results of sections 3, 4, and 5, but note that autocorrelation concerns will be addressed in a forthcoming paper (Páez et al, 2002). The model in equation (44) serves as a benchmark against which to compare the results in the following section.

7.2 GWR: estimation and inference

In this section we consider estimation and inference of GWR models under the cases discussed in section 3. Although in general this need not be the case, we decide to conduct estimation at every location corresponding to a recorded observation, to give a total of $m = 493$ local models. In order to avoid numerical overflows when computing the exponential, distance d_{oi} among observations is given in kilometers. This has no effect on the results.

The first step to estimate a set of GWR models under case 1 (section 3.1, 'known' kernel bandwidth) is calibration of the kernel function by means of the cross-validation score (15). The procedure is simply to calculate the score by using different values of parameter γ (which is assumed to be spatially invariant) and to plot the results to find an estimate of the value that minimizes C. The results of this exercise appear in figure 1, where it can be seen that the cross-validation procedure suggests a value of around 0.16 for the kernel bandwidth. This is used as a starting value to minimize the problem automatically to give a more accurate value of 0.1668. The existence of a minimum could be taken by itself, in current practice, as an indication that GWR is a better representation of the data than the global model. Estimation of location-specific γ_o shows this to be only coincidentally true.

Under case 2 (section 3.2, unknown kernel bandwidth) local parameters γ_o are estimated at each location by minimizing the concentrated likelihood function (24). This is done for a total of 493 locations, for which the LM test is also calculated by using equation (39). The map in figure 2 shows the results of these two steps, and it can be seen there that parameter γ_o ranges between 0.0028 (the smallest *significant* kernel bandwidth) and 0.0543, and is significantly different from zero for most of the study

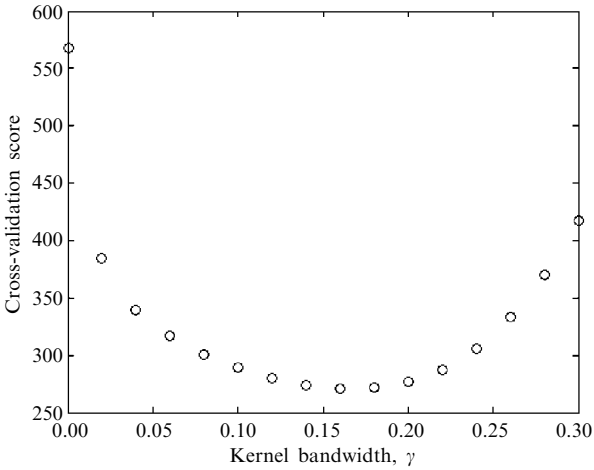


Figure 1. Cross-validation and selection of a global kernel bandwidth

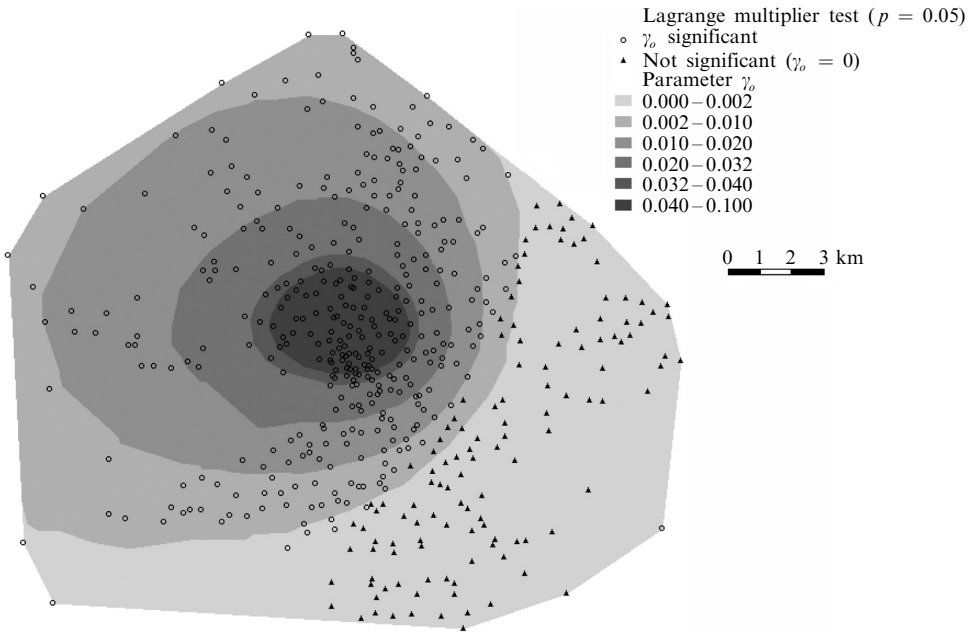


Figure 2. Location-specific kernel bandwidths and significance tests.

area, with the exception of 126 locations in the southeast region of the city. The significance of the parameter is tested by means of the LM approach discussed in section 5, with a nominal level of significance of 0.05 corrected by means of the step-down procedure introduced in section 6. From the figure it can be seen that spatial variation of parameter γ_o is rather regular, thus indicating that the underlying process is continuous but not uniform across space. It is also clear from the figure that the global kernel bandwidth obtained from the cross-validation procedure is widely off the mark all over the study area. This result provides some evidence that the cross-validation procedure commonly used in practice is not a reliable method to estimate the kernel bandwidth. Moreover, comparing the rates of decay given by the global and the maximum local kernel bandwidths (figure 3) shows that the global kernel bandwidth implies a much steeper rate of discount and considerably smaller estimation

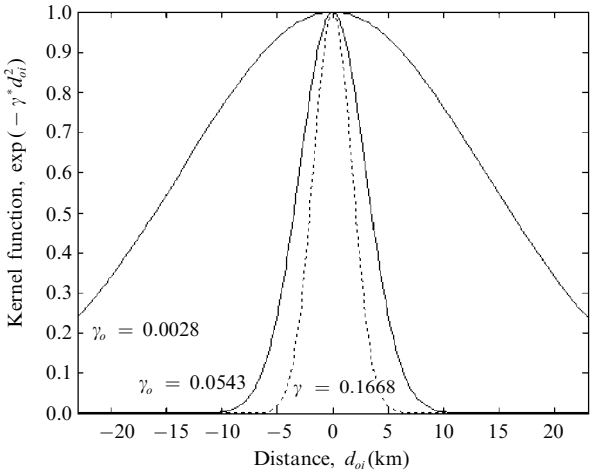


Figure 3. Rates of decay of the kernel function.

'neighborhoods' than the local counterparts. The solid lines in the figure denote the range of decay rates given by the smallest significant and the maximum γ_o, and the dotted line is the global kernel bandwidth obtained from the cross-validation procedure. Clearly, the use of the latter in the current case, will lead to considerably more parametric variation than the use of location-specific γ_o, as the parameters adapt to represent distinctive local conditions. In general, it seems very likely that the cross-validation procedure will tend to produce kernel bandwidths that overestimate the local values, as a consequence of trying to improve the 'fit' locally. This means that when genuine locational heterogeneity exists it will be exaggerated to varying degrees across the study area, resulting in spurious parametric drift in many cases. This seems to be confirmed by the following results regarding parameter estimation.

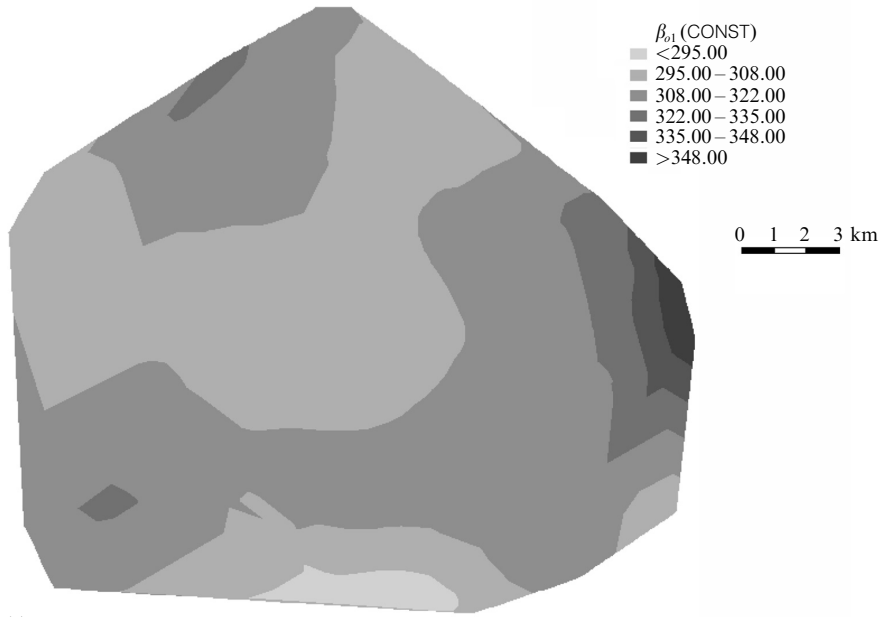
Applying the relevant kernel bandwidth value, local models are estimated for case 1 under the WLS approach and for case 2 with γ_o that result from minimizing the concentrated likelihood. The results are summarized in table 1. It should be appreciated there that the range of parametric variation is considerably wider for models that use the global kernel bandwidth, thus giving the impression that spatial parametric nonstationarity is more dramatic than it really is. A related consequence of this effect is that there are sign reversals for most parameters under case 1, but only one sign reversal for the case of the parameter corresponding to I% when location-specific kernel bandwidths are used (the latter variable is not significant for any location in the study area). In general, the results under case 2 tend to resemble more closely the global model given by equation (44).

Table 1. Parameter estimation with global and location-specific kernel bandwidths.

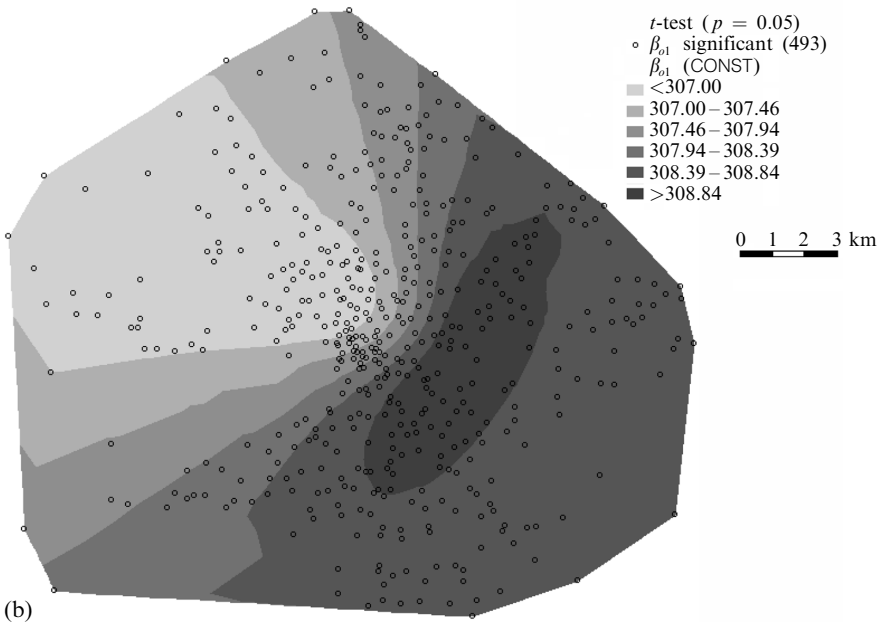
	γ	β_{o1} CONST	β_{o2} LDIST	β_{o3} C%	β_{o4} I%	β_{o5} R%
Case 1: 'known' kernel bandwidth						
Maximum	0.1668	361.3210	1.8002	0.3160	0.3269	0.1234
Minimum	0.1668	282.9090	−6.9200	−1.3107	−3.3168	−0.0214
Case 2: unknown kernel bandwidth						
Maximum	0.0543	309.3041	−0.5670	0.0031	0.0090	0.0238
Minimum	0.0028	306.5883	−0.9859	0.0012	−0.0018	0.0116

Results of estimating the parameters appear in figures 4–6, with the exception of β_{o3} which is significant in a very reduced (12) number of locations, and β_{o4} which is not significant at all. Obviously, some spatial nonstationarity results from the presence of locational heterogeneity, but not to the extent that the use of a global kernel bandwidth would lead us to believe. For the case of the constant, for instance, it can be seen that the range of variation of the local parameters is less than 3, which contrasts with a range of almost 80 in the case of the global kernel bandwidth. Variation ranges of this magnitude, on the other hand, are not uncommon in GWR models estimated by using global kernel bandwidths.

Similar results are obtained for the rest of the parameters. In particular, the lack of similarity between the spatial distribution of parameters estimated by using global and local kernels bandwidths (in a maximum likelihood framework) is notorious. These results suggest that care should be exercised when interpreting local parameters obtained by using a global kernel bandwidth. Regarding the maximum likelihood results, a possible extension of the exercise could be to combine the results of the estimation with a study of the spatial distribution of land uses and land-use intensities to give a more accurate picture of the process and its operation. In particular, testing for the significance of the parameters could help to give a clear idea of which variables



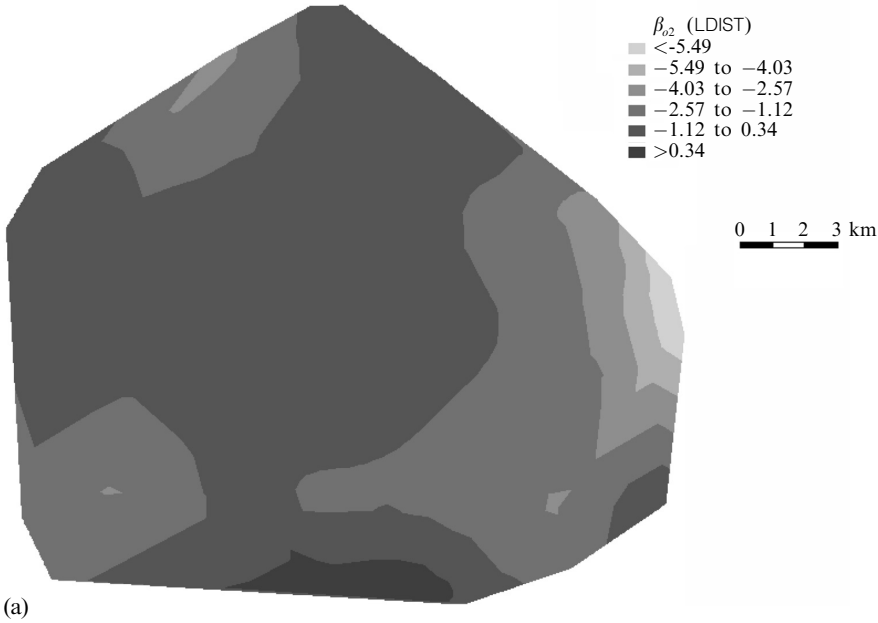
(a)



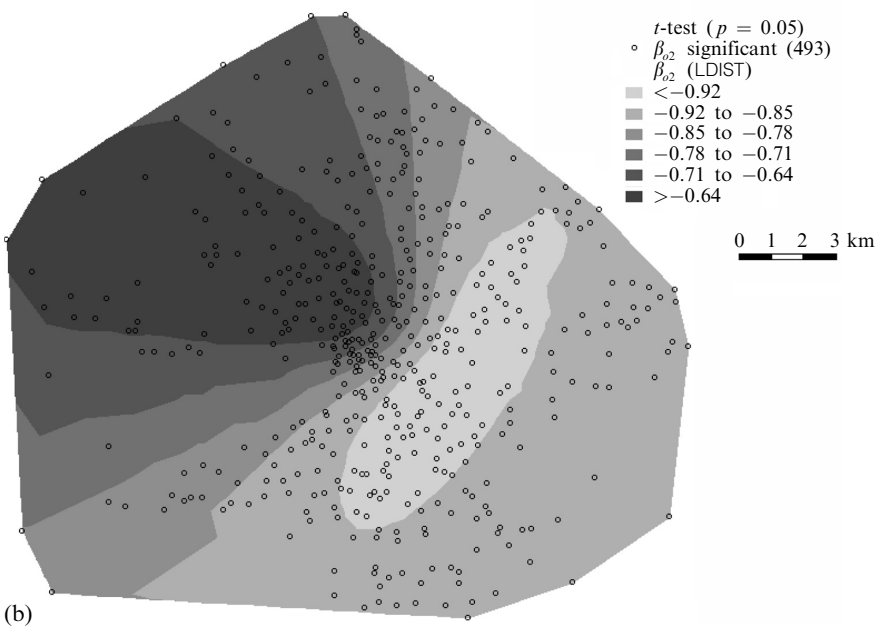
(b)

Figure 4. Constant parameter estimated (a) by using a global kernel bandwidth, (b) by using location-specific kernel bandwidths, and significance tests.

are operating more importantly in different parts of the city. The residential land-use variable, for instance, is tenuously but significantly contributing to determine higher temperatures in the suburbs of the city, in particular to the southeast, but not in more central parts of the study area. This kind of result cold provide an important contribution to bring the problem into focus. Parametric significance is tested using the step-down procedure discussed in section 6 with a nominal level of significance $\alpha = 0.05$.



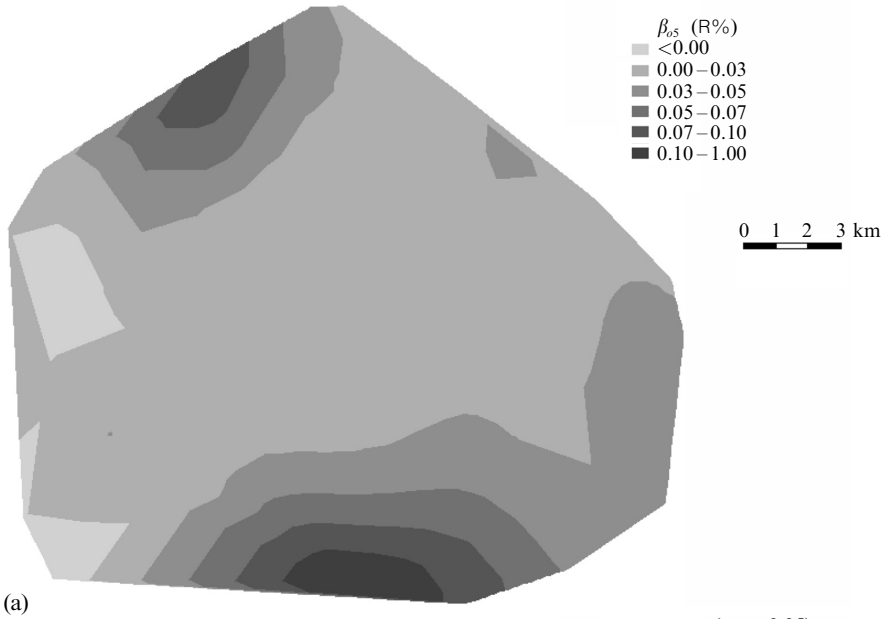
(a)



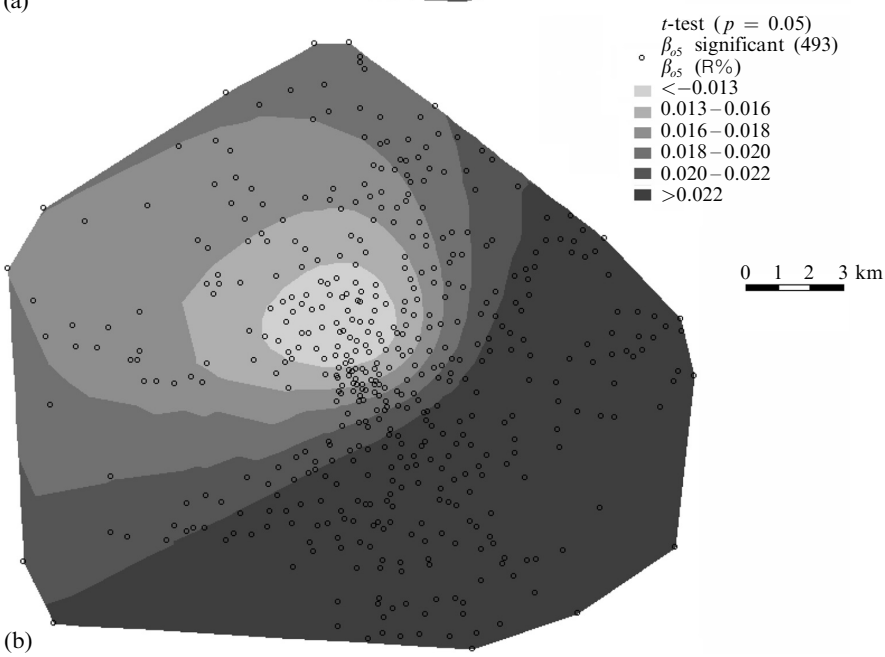
(b)

Figure 5. LDIST parameter estimated (a) by using a global kernel bandwidth, (b) by using location-specific kernel bandwidths, and significance tests.

Finally, figure 7 (see over) shows how local variance σ_o^2 estimated for our case study varies smoothly across space, from a minimum of 0.1042 to the west of the study area, to a maximum of 1.133 to the southeast. Clearly, heterogeneity does follow a pattern in the present case. Parametric nonstationarity, in particular with reference to parameter vector $\boldsymbol{\beta}_o$, is then seen as a consequence of locational heterogeneity. All in all, the above results argue in favor of shifting the focus from studying spatial parametric nonstationarity to a more direct study of the variance in a spatial context.



(a)



(b)

Figure 6. R% parameter estimated (a) by using a global kernel bandwidth, (b) by using location-specific kernel bandwidths, and significance tests.

8 Summary and conclusions

Development of the method known as geographically weighted regression has followed the general lines of methods in smoothing, in particular local likelihood estimation and local regression. This approach, however, has led to some conceptual and technical difficulties. In this paper we have argued that GWR is more fruitfully seen as a model of spatial variance heterogeneity, because many of the apparent technical difficulties on the way to developing GWR as a spatial statistical model vanish with this change of

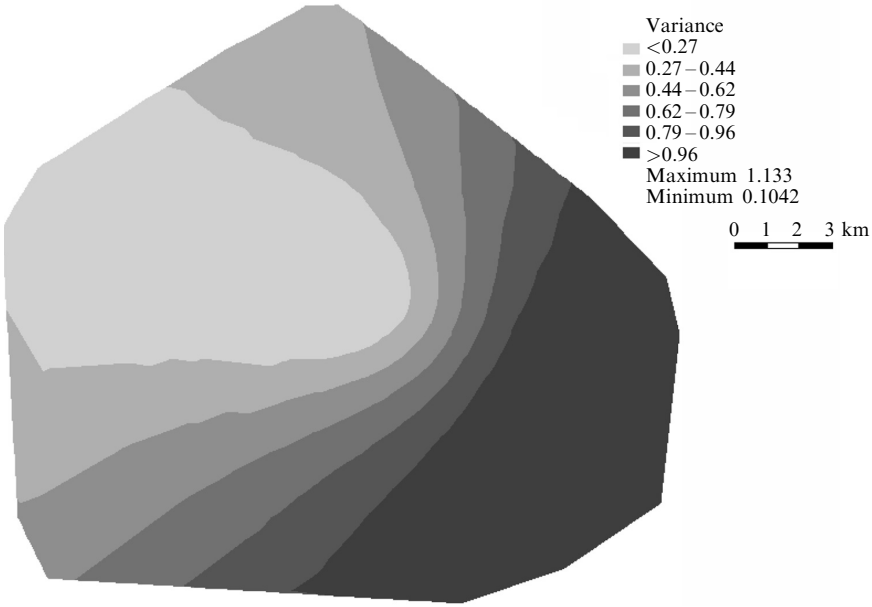


Figure 7. Location-specific variance parameter, σ_o^2.

perspective. We have therefore proposed a conceptual shift from smoothing methods to variance modeling methods. The techniques to diagnose variance heterogeneity and to model variance functions have been studied before (for example, Cook and Weisberg, 1983; Davidian and Carroll, 1987) but to the best of our knowledge not in a specifically spatial context. The power of estimation GWR style is that it allows us to move away from global models and to study local relationships resulting from situations of what we have termed locational heterogeneity. By bringing GWR into this statistical perspective we hope to have dispelled many difficult questions surrounding the application of the method, and to have contributed to the establishment of GWR as a proper statistical model.

Our results can be summarized as follows. We have demonstrated how the variance can be modeled as a smooth function of geographical distance. As the variance function is general enough to cover the case of spatial ordering as an explanatory factor, the standard regularity conditions apply and we can derive asymptotic results needed for inference. The maximum likelihood approach introduced here allows for more formal ways of estimating the parameter of the kernel function, do it on a local basis, and test the hypothesis of locational heterogeneity (which implies spatial nonstationarity) by using a Lagrange multiplier approach. In practical terms, these results mean that GWR can be safely applied whenever the hypothesis of variance homogeneity is rejected by means of a statistical test. Moreover, to address the issue of multiple comparisons, we adopted a step-down procedure of simultaneous inference that overcomes the problem of loss of power in the presence of highly dependent tests. In order to keep the presentation in focus, we have deliberately avoided an extensive discussion of the important topic of spatial autocorrelation. This is the object of a forthcoming research paper.

Acknowledgements. We are indebted to Professor Stewart Fotheringham and Dr Chris Brunsdon for a critical lecture of an earlier version of this paper, and for their comments and discussion of the concept of local likelihood estimation. We are especially grateful to three unnamed reviewers and to Professor Eric Sheppard for their criticism and comments, which importantly contributed to bring the paper into focus. We naturally assume full responsibility for any and all errors or shortcomings.

References

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

Anselin L, 1988b, "Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity"*Geographical Analysis* **20** 1 – 17

Anselin L, 1995, "Local indicators of spatial association—LISA" *Geographical Analysis* **27** 93 – 115

Box G E P, Meyer R D, 1986, "Dispersion effects from fractional design" *Technometrics* **28**(1) 19 – 27

Breusch T S, Pagan A R, 1979, "A simple test for heteroscedasticity and random coefficient variation" *Econometrica* **47** 1287 – 1294

Breusch T S, Pagan A R, 1980, "The Lagrange multiplier test and its applications to model specification in econometrics" *Review of Economic Studies* **48** 239 – 253

Brunsdon C, Fotheringham A S, Charlton M E, 1996, "Geographically weighted regression: a method for exploring spatial nonstationarity" *Geographical Analysis* **28** 281 – 298

Can A, 1992, "Specification and estimation of hedonic housing price models" *Regional Science and Urban Economics* **22** 453 – 474

Carroll R J, Rupert D, 1988 *Transformation and Weighting in Regression* (Chapman and Hall, New York)

Casetti E, 1972, "Generating models by the expansion method: applications to geographic research" *Geographical Analysis* **4** 81 – 91

Chilès J P, Delfinier P, 1999 *Geostatistics: Modeling Spatial Uncertainty* (John Wiley, New York)

Cleveland W S, Devlin S J, 1988, "Locally weighted regression: an approach to regression analysis by local fitting" *Journal of the American Statistical Association* **83** 596 – 610

Cliff A D, Ord J K, 1981 *Spatial Processes: Models and Applications* (Pion, London)

Cook R D, Weisberg S, 1983, "Diagnostics for heteroscedasticity in regression" *Biometrika* **70**(1) 1 – 10

Cressie N ,1991 *Statistics for Spatial Data* (John Wiley, New York)

Davidian M, Carroll R J, 1987, "Variance, function estimation" *Journal of the American Statistical Association* **82** 1079 – 1091

Engle R F, 1984, "Wald, likelihood ratio, and Lagrange multiplier tests in econometrics", in *Handbook of Econometrics* Eds Z Griliches, M D Intriligator (Elsevier Science, Amsterdam) pp 775 – 826

Fan J, Farmen M, Gijbels I, 1998, "Local maximum likelihood estimation and inference" *Journal of the Royal Statistical Society B* **60** 591 – 608

Fotheringham A S, Charlton M, Brunsdon C, 1997, "Measuring spatial variations in relationships with geographically weighted regression", in *Recent Developments in Spatial Analysis* Eds M M Fischer, A Getis (Springer, Berlin) pp 60 – 82

Fotheringham A S, Charlton M, Brunsdon C, 1998, "Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis" *Environment and Planning A* **30** 1905 – 1927

Getis A, Ord J K, 1992, "The analysis of spatial association by use of distance statistics" *Geographical Analysis* **24** 189 – 206

Griffith D, 1988 *Advanced Spatial Statistics* (Kluwer Academic, Dordrecht)

Griffith D A, Layne L J, 1999 *A Casebook for Spatial Statistical Data Analysis* (Oxford University Press, Oxford)

Haining R, 1990 *Spatial Data Analysis in the Social and Environmental Sciences* (Cambridge University Press, Cambridge)

Leung Y, Mei C-L, Zhang W-X, 2000, "Statistical tests for spatial nonstationarity based on the geographically weighted regression model" *Environment and Planning A* **32** 9 – 32

Liu W, 1996, "Multiple tests of a non-hierarchical finite family of hypotheses" *Journal of the Royal Statistical Society B* **58** 455 – 461

Lyon J D, Tsai C L, 1996, "A comparison of tests for heteroscedasticity" *The Statistician* **45** 337 – 349

Miller R G, 1981 Simultaneous Statistical Inference 2nd edition (Springer, New York)

Ord J K, Getis A, 1995, "Local spatial autocorrelation statistics: distributional issues and an application" *Geographic Analysis* **27** 286 – 306

Ord J K, Getis A, 2001, "Testing for local spatial autocorrelation in the presence of global autocorrelation" *Journal of Regional Science* **41** 411 – 432

Páez A, Uchida T, Miyamoto K, 1998, "Urbanization and the urban heat island effect from a spatial descriptive approach", in *Papers on City Planning 33* (The City Planning Institute of Japan, Tokyo) pp 67 – 72

Páez A, Uchida T, Miyamoto K, 2001, "Spatial association and heterogeneity issues in land price models" *Urban Studies* **38** 1493 – 1508

Páez A, Uchida T, Miyamoto K, 2002, "A general framework for estimation and inference of geographically weighted regression models: 2. Spatial association and model specification tests" *Environment and Planning A* **34** forthcoming

Simes R J, 1986, "An improved Bonferroni procedure for multiple tests of significance" *Biometrika* **73** 751 – 754

Tibshirani R, Hastie T, 1987, "Local likelihood estimation" *Journal of the American Statistical Association* **82** 559 – 567

Verbyla A P, 1993, "Modelling variance heterogeneity: residual maximum likelihood and diagnostics" *Journal of the Royal Statistical Society B* **55** 493 – 508