Environment and Planning A 2000, volume 32, pages 9 – 32

DOI:10.1068/a3162

Statistical tests for spatial nonstationarity based on the geographically weighted regression model

Yee Leung
Department of Geography, Centre for Environmental Studies, and Joint Laboratory for Geoinformation Science, The Chinese University of Hong Kong, Shatin, Hong Kong;
e-mail: yeeleung@cuhk.edu.hk

Chang-Lin Mei, Wen-Xiu Zhang
School of Science, Xi'an Jiaotong University, Xi'an Shaanxi, 710049, People's Republic of China

Received 23 November 1998; in revised form 12 April 1999

Abstract. Geographically weighted regression (GWR) is a way of exploring spatial nonstationarity by calibrating a multiple regression model which allows different relationships to exist at different points in space. Nevertheless, formal testing procedures for spatial nonstationarity have not been developed since the inception of the model. In this paper the authors focus mainly on the development of statistical testing methods relating to this model. Some appropriate statistics for testing the goodness of fit of the GWR model and for testing variation of the parameters in the model are proposed and their approximated distributions are investigated. The work makes it possible to test spatial nonstationarity in a conventional statistical manner. To substantiate the theoretical arguments, some simulations are run to examine the power of the statistics for exploring spatial nonstationarity and the results are encouraging. To streamline the model, a stepwise procedure for choosing important independent variables is also formulated. In the last section, a prediction problem based on the GWR model is studied, and a confidence interval for the true value of the dependent variable at a new location is also established. The study paves the path for formal analysis of spatial nonstationarity on the basis of the GWR model.

1 Introduction

In spatial analysis, the ordinary linear regression (OLR) model has been one of the most useful statistical methods to identify the nature of relationships among variables (for example, see Dobson, 1990). In this technique, a variable y, called the dependent variable, is modeled as a linear function of a set of independent variables $x_1, x_2, \ldots, x_p$. Based on n observations $(y_i; x_{i1}, x_{i2}, \ldots, x_{ip})$, $(i = 1, 2, \ldots, n)$, from a study region, the model can be expressed as

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \epsilon_i , \tag{1}$$

where β_0 $\beta_1, \ldots, \beta_p$ are parameters, and ϵ_1 $\epsilon_2, \ldots, \epsilon_n$ are error terms which are generally assumed to be independent normally distributed random variables with zero means and constant variance σ^2. In this model, each of the parameters can be thought of as the 'slopes' between one of the independent variables and the dependent variable. The least squares estimate of the parameter vector can be written as

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0\ \hat{\beta}_1 \ldots \hat{\beta}_p)^{\mathrm{T}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{Y}, \tag{2}$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \qquad \boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}. \tag{3}$$

Statistical properties of these estimates have been well studied and various hypothesis tests have also been established.

It is important to note that the slopes (that is, the parameters) in the model in equation (1) are assumed to be universal across the study area. However, in many real-life situations, there is ample evidence for the lack of uniformity in the effects of space. Variation or spatial nonstationarity in relationships over space commonly exist in spatial data sets and the assumption of stationarity or structural stability over space may be highly unrealistic (for example, see Anselin, 1988; Fotheringham et al, 1996; Fotheringham, 1997). It is shown that, as stated in Brunsdon et al (1996), (a) relationships can vary significantly over space and a 'global' estimate of the relationships may obscure interesting geographical phenomena; (b) variation over space can be sufficiently complex that it invalidates simple trend-fitting exercises. So, when analyzing spatial data, we should consider such modeling strategies that take into account this kind of spatial nonstationarity.

Over the years, some approaches have been proposed to incorporate spatial structural instability or spatial drift into the models. For example, Anselin (1988; 1990) has investigated regression models with spatial structural change. Casetti (1972; 1986), Jones and Casetti (1992), and Fotheringham and Pitts (1995) have studied spatial variations by the expansion method. Cleveland (1979), Cleveland and Devlin (1988), Casetti (1982), Foster and Gorr (1986), Gorr and Olligschlaeger (1994), Brunsdon et al (1996; 1997), and Fotheringham et al (1997a; 1997b) have examined the following varying-parameter regression model:

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} x_{ik} + \epsilon_i \, . \tag{4}$$

Unlike the OLR model in equation (1), this model allows the parameters to vary in space. However, this model in its unconstrained form is not implementable because the number of parameters increases with the number of observations. Hence, strategies for limiting the number of degrees of freedom used to represent variation of the parameters over space should be developed when the parameters are estimated.

There are several methods for estimating the parameters. For example, the method of the spatial adaptive filter (Foster and Gorr, 1986; Gorr and Olligschlaeger, 1994) uses generalized damped negative feedback to estimate spatially varying parameters of the model in equation (4). However, this approach incorporates spatial relationships in a rather ad hoc manner and produces parameter estimates that cannot be tested statistically. Locally weighted regression method and kernel regression method (for example, see Cleveland, 1979; Casetti, 1982; Cleveland and Devlin, 1988; Cleveland et al, 1988; Brunsdon, 1995; Wand and Jones, 1995) focus mainly on the fit of the dependent variable rather than on spatially varying parameters. Furthermore, the weighting system depends on the location in the 'attribute space' (Openshaw, 1993) of the independent variables. For a further review of the varying-parameter model, see, for instance, Gorr and Olligschlaeger (1994) and Brunsdon et al (1996).

Along this line of thinking, Brunsdon et al (1996; 1997) and Fotheringham et al (1997a; 1997b) suggested a so-called geographically weighted regression (GWR) technique in which the parameters in equation (4) are estimated by a weighted least squares procedure, making the weighting system dependent on the location in geographical space and, therefore, allowing local rather than global parameters to be estimated. The typical output from a GWR model is a set of parameters that can be mapped in the geographic space to represent nonstationarity or parameter 'drift'.

Compared with other methods, the GWR technique appears to be a relatively simple but useful geographically oriented method to explore spatial nonstationarity.

Based on the GWR model, not only can variation of the parameters be explored, but significance of the variation can also be tested. Unfortunately, at present, only Monte Carlo simulation has been used to perform tests on the validity of the model. In this technique, under the null hypothesis that the global linear regression model holds, any permutation of the observations $(y_i; x_{i1}, x_{i2}, \ldots, x_{ip})$ $(i = 1, 2, \ldots, n)$ among the geographical sampling points is equally likely to occur. The observed values of the statistics proposed can then be compared with these randomization distributions and the significance tests can be performed accordingly. The computational overhead of this method is, however, considerable, especially for a large data set. Also, because the validity of these randomization distributions is limited to the given data set, this in turn restricts the generality of the proposed statistics. The ideal way to test the model is to construct appropriate statistics and to perform the tests in a conventional statistical manner.

The purpose of this paper is to solve the problem in the context of classical hypothesis testing. Specifically, we want to entertain two basic questions associated with the GWR model. (1) Given a situation, how can we tell whether the GWR model describes the data set significantly better than an OLR model? (2) Given that the GWR model is valid, does each set of its parameters vary significantly across the study region?

It is necessary to answer the first question when the GWR technique is employed to analyze a given data set. A GWR model will certainly fit a given data set better than an OLR model. However, from the practical point of view, the simpler a model, the easier it can be applied and interpreted. If a GWR model does not perform significantly better than an OLR model, we would rather use the OLR model in practice, and therefore we can also know that there is no significant drift in any of the model parameters. By answering the second question, we can specify which set of the parameters drifts significantly over the study region.

In this paper we propose several appropriate statistics and derive their approximated null distributions for the statistical test of the above hypotheses. We also extend the stepwise procedure for the selection of independent variables in the OLR model to the GWR model. Based on the statistics and their approximated distributions, we further consider the prediction aspect of the GWR model and derive the interval estimate for the true value of the dependent variable at a new location.

We first introduce the GWR model and briefly describe the related problems in section 2. The goodness-of-fit tests based on the residual sum of squares for the GWR model and a stepwise procedure for choosing important independent variables are proposed in section 3. In section 4, we propose a statistical test to examine whether each set of the parameters of the GWR model varies significantly across a study region. In section 5, simulations are performed to examine the power of the proposed statistical methods, and in section 6, we consider the prediction problem of the GWR model and construct a confidence interval for the true value of the dependent variable. The paper is then concluded with a brief summary and discussion.

2 Geographically weighted regression

2.1 The model and the estimation of the parameters

The mathematical presentation of the GWR model is the same as the varying-parameter regression model in equation (4). Here, the parameters are assumed to be functions of the locations on which the observations are obtained. That is,

$$y_i = \beta_{i0} + \sum_{k=0}^{p} \beta_{ik} x_{ik} + \epsilon_i, \qquad i \in C = \{1, 2, \ldots, n\}, \tag{5}$$

where C is the index set of locations of n observations and β_{ik} is the value of the kth parameter at location i.

The parameters in the GWR model are estimated by the weighted least squares approach. The weighting matrix is taken as a diagonal matrix where each element in its diagonal is assumed to be a function of the location of observation. Suppose that the weighting matrix at location i is $\mathbf{W}(i)$. Then the parameter vector at location i is estimated as

$$\hat{\boldsymbol{\beta}}(i) = [\mathbf{X}^{\mathrm{T}}\mathbf{W}(i)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(i)\boldsymbol{Y}, \tag{6}$$

where $\mathbf{W}(i) = \mathrm{diag}[\mathrm{w}_1(i), \mathrm{w}_2(i), \ldots, \mathrm{w}_n(i)]$ and $\mathbf{X}$ and $\boldsymbol{Y}$ are defined as in equation (3). Here we assume that the inverse of the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{W}(i)\mathbf{X}$ exists.

In fact, according to the principle of the weighted least squares method, the generated estimators at location i in equation (6) are obtained by solving the following optimization problem. That is, determine the parameters $\beta_0, \beta_1, \ldots, \beta_p$ at each location i so that

$$\sum_{j=1}^{n} \mathrm{w}_j(i)(y_j - \beta_0 - \beta_1 x_{j1} - \ldots - \beta_p x_{jp})^2 \tag{7}$$

is minimized. Given appropriate weights $\mathrm{w}_j(i)$ which are a function of the locations at which the observations are made, different emphases can be given to different observations for generating the estimated parameters at location i.

2.2 Possible choices of the weighting matrix

As pointed out above, the role of the weighting matrix is to place different emphases on different observations in generating the estimated parameters. In spatial analysis, observations close to a location i are generally assumed to exert more influence on the parameter estimates at location i than those farther away. When the parameters at location i are estimated, more emphasis should be placed on the observations which are close to location i. A simple but natural choice of the weighting matrix at location i is to exclude those observations that are farther than some distance d from location i. This is equivalent to setting a zero weight on observation j if the distance from i to j is greater than d. If the distance from i to j is expressed as d_{ij}, the elements of the weighting matrix at location i can be chosen as

$$\mathrm{w}_j(i) = \begin{cases} 1, & \text{if } d_{ij} \leqslant d, \qquad j = 1, 2, \ldots, n\,. \\ 0, & \text{if } d_{ij} > d, \end{cases} \tag{8}$$

The above weighting function suffers from the problem of discontinuity over the study area. One way to overcome this problem is to specify $\mathrm{w}_j(i)$ as a continuous and monotone decreasing function of d_{ij}. One obvious choice might be

$$\mathrm{w}_j(i) = \exp(-\theta d_{ij}^2), \qquad j = 1, 2, \ldots, n, \tag{9}$$

so that, if i is a point at which observation is made, the weight assigned to that observation will be unity and the weights of the others will decrease according to a Gaussian curve as d_{ij} increases. Here, θ is a nonnegative constant depicting the way the Gaussian weights vary with distance. Given d_{ij}, the larger θ, the less emphasis is placed on the observation at location j. The problem of equation (9) amounts to assigning weights to all locations of the study area.

A compromise between the above two weighting functions can be reached by setting the weights to zero outside a radius d and to decrease monotonically to zero inside the radius as d_{ij} increases. For example, we can take the elements of the weighting matrix as a bisquare function, that is,

$$\mathrm{w}_j(i) = \begin{cases} (1 - d_{ij}^2/d^2)^2, & \text{if } d_{ij} \leqslant d, \qquad j = 1, 2, \ldots, n, \\ 0, & \text{if } d_{ij} > d, \end{cases} \tag{10}$$

The weighting function in equation (9) is the most common choice in practice.

2.3 Estimation of the parameter in the weighting matrices

In the process of calibrating a GWR model, the weighting matrix should first be decided. If we use, for example, Gaussian weighting in equation (9), the parameter θ should be predetermined. This can be done by a cross-validation procedure (for example, see Cleveland, 1979; Bowman, 1984; Brunsdon, 1995). Let

$$\Delta(\theta) = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}(\theta)]^2 , \tag{11}$$

where $\hat{y}_{(i)}(\theta)$ is the fitted value of y_i with the observation at location i omitted from the calibration process. Choose θ_0 as a desirable parameter such that

$$\Delta(\theta_0) = \min\Delta(\theta) . \tag{12}$$

In practice, plotting $\Delta(\theta)$ against the parameter θ will provide guidance on selecting an appropriate value of the parameter or it can be selected automatically by an optimization technique. The parameter in other forms of the weighting matrix can be determined similarly.

2.4 Testing for spatial nonstationarity of the parameters in the model

The importance of exploring spatial nonstationarity has been fully addressed by, for example, Fotheringham et al (1997b) and Charlton et al (1997). The GWR technique has been demonstrated to be a useful means for detecting spatial nonstationarity. By knowing whether or not the parameters in a GWR model vary significantly, we can provide not only a greater insight into both the data and the framework in which the data are examined, but also some valuable guidance for decision in future actions.

From the statistical point of view, the following two questions are the most important and should be tested statistically:

(1) Does a GWR model describe the data significantly better than an OLR model? That is, on the whole, do the parameters in the GWR model vary significantly over the study region?

(2) Does each set of parameters β_{ik} $(i = 1, 2, \ldots, n)$ exhibit significant variation over the study region?

For the first question, it is, in fact, a goodness-of-fit test for a GWR model. It is equivalent to testing whether or not $\theta = 0$ if we use equation (9) as the weighting function. In the second case, for any fixed k, the deviation of β_{ik} $(i = 1, 2, \ldots, n)$ can be used to evaluate the variation of the slope of the kth independent variable. Because it is very difficult to find the null distribution of the estimated parameter, say θ in equation (9), in the weighting matrix, a Monte-Carlo technique has been employed to perform the tests (Brunsdon et al, 1996; Fotheringham et al, 1997a). However, as pointed out above, the computational overhead of the method is considerable. Furthermore, the validity of the reference distributions obtained by the randomized permutation is limited to the given data set, and it in turn may restrict the generality of the corresponding statistics.

Some other approaches which account for spatial variation tend to add extra parameters to the model. For example, the structural change models examine spatial variation by dividing the study area into several regions and calibrating different regression models in different regions (for example, see Anselin, 1988; 1990); the expansion method embeds spatial variation by assuming the regression coefficients to be the other function of the locations (for example, see Jones and Casetti, 1992; Fotheringham and Pitts, 1995). In these cases, the number of parameters in the model is fixed and the classical likelihood ratio tests can therefore be used to test spatial nonstationarity. For the varying-parameter regression model in equation (1), the number of parameters in the model, as has been pointed out in section 1, increases with the number of observations. The GWR technique in fact exerts some constraints on the parameters by the weighted least squares method. Given the way parameters in the GWR model are estimated, it is unlikely to base the tests on the likelihood function to test spatial nonstationarity. However, it is still correct to perform the tests by using the residuals.

In the following two sections we use the notion of residual sum of squares to formulate some methods to examine the above questions within the conventional hypothesis-testing framework. Based on the proposed statistics and their approximated distributions, we also propose a stepwise procedure for choosing important independent variables of the GWR model.

3 Goodness-of-fit test and stepwise procedure for selection of independent variables

Based on the notion of residual sum of squares, some statistics are constructed in this section, and their approximated distributions are investigated to test whether a GWR model describes a given data set significantly better than an OLR model does. Our derivation is an extension of the method described in Cleveland and Devlin (1988) and Cleveland et al (1988). Furthermore, we propose a stepwise procedure to choose important independent variables within the framework of the GWR model.

In the sections to follow, we assume that the weighting matrix for calibrating the GWR model is given and the following two assumptions hold:

Assumption 1. The error terms $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are independently identically distributed as a normal distribution with zero mean and constant variance σ^2.

Assumption 2. Let $\hat{y}_i$ be the fitted value of y_i at location i. For all $i = 1, 2, \dots, n$, $\hat{y}_i$ is an unbiased estimate of $\mathrm{E}(y_i)$. That is, $\mathrm{E}(\hat{y}_i) = \mathrm{E}(y_i)$ for all i.

Assumption 1 is in fact the conventional assumption in theoretical analysis of regression. Assumption 2 is in general not exactly true for local linear fitting except that the exact global linear relationship between the dependent variable and the independent variables exists (see Wand and Jones, 1995, page 120–121 for the univariate case). However, the local-regression methodology is mainly oriented towards the search for low-bias estimates (Cleveland et al, 1988). In this sense, the bias of the fitted value could be negligible So, assumption 2 is a realistic one in the GWR model because this technique still belongs to the local-regression methodology.

3.1 The residual sum of squares and its approximated distribution

Let $\boldsymbol{x}_i^{\mathrm{T}} = (1\ x_{i1} \dots x_{ip})$ be the ith row of $\mathbf{X}$ $(i = 1, 2, \dots, n)$ and $\hat{\boldsymbol{\beta}}(i)$ the estimated parameter vector at location i. Then the fitted value of y_i is given by

$$\hat{y}_i = \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}(i) = \boldsymbol{x}_i^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(i)\mathbf{W}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(i)\boldsymbol{Y}. \tag{13}$$

Let $\hat{\boldsymbol{Y}} = (\hat{y}_1\ \hat{y}_2 \dots \hat{y}_n)^{\mathrm{T}}$ be the vector of the fitted values and let $\hat{\boldsymbol{\epsilon}} = (\hat{\epsilon}_1\ \hat{\epsilon}_2 \dots \epsilon_n)^{\mathrm{T}}$ be the vector of the residuals. Then

$$\hat{\boldsymbol{Y}} = \mathbf{L}\boldsymbol{Y}, \tag{14}$$

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\mathbf{I} - \mathbf{L})\boldsymbol{Y}, \tag{15}$$

where

$$\mathbf{L} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(1)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(1) \\ \boldsymbol{x}_2^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(2)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(2) \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(n)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(n) \end{pmatrix} \tag{16}$$

is an $n \times n$ matrix and $\mathbf{I}$ is an identity matrix of order n.

It should be noted that the role of the matrix $\mathbf{L}$ is to turn the observed values of the dependent variable into the fitted ones, which is the same as that of the hat matrix in the OLR literature (for example, see Neter et al, 1989). However, the hat matrix in the OLR literature is symmetric and idempotent, but, because of the weighting matrix $\mathbf{W}(i)$ $(i = 1, 2, ..., n)$, $\mathbf{L}$ is generally not.

We denote by $\mathrm{RSS_g}$ the residual sum of squares. Then

$$\mathrm{RSS_g} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \hat{\boldsymbol{\epsilon}}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y}^{\mathrm{T}}(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\boldsymbol{Y}. \tag{17}$$

This quantity measures the goodness of fit of a GWR model for the given data and can be used to estimate σ^2, the common variance of the error terms ϵ_i $(i = 1, 2, ..., n)$. Indeed, according to assumptions 1 and 2, we have

$$\mathrm{E}(\hat{\boldsymbol{\epsilon}}) = \mathrm{E}(\boldsymbol{Y}) - \mathrm{E}(\hat{\boldsymbol{Y}}) = \boldsymbol{0}, \qquad \mathrm{E}(\boldsymbol{\epsilon\epsilon}^{\mathrm{T}}) = \sigma^2\mathbf{I}, \tag{18}$$

where $\boldsymbol{\epsilon} = (\epsilon_1\ \epsilon_2 \dots \epsilon_n)^{\mathrm{T}}$ is the vector of the error terms. Therefore, $\mathrm{RSS_g}$ can also be expressed as

$$\begin{aligned} \mathrm{RSS_g} &= [\hat{\boldsymbol{\epsilon}} - \mathrm{E}(\hat{\boldsymbol{\epsilon}})]^{\mathrm{T}}[\hat{\boldsymbol{\epsilon}} - \mathrm{E}(\hat{\boldsymbol{\epsilon}})] \\ &= [\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})]^{\mathrm{T}}(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})[\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})] \\ &= \boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\boldsymbol{\epsilon}\,. \end{aligned} \tag{19}$$

So,

$$\begin{aligned} \mathrm{E}(\mathrm{RSS_g}) &= \mathrm{E}[\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\boldsymbol{\epsilon}] \\ &= \mathrm{E}\{\mathrm{tr}[\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\boldsymbol{\epsilon}]\} \\ &= \mathrm{E}\{\mathrm{tr}[(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\boldsymbol{\epsilon\epsilon}^{\mathrm{T}}]\} \\ &= \mathrm{tr}[(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\mathrm{E}(\boldsymbol{\epsilon\epsilon}^{\mathrm{T}})] \\ &= \sigma^2\delta_1\,, \end{aligned} \tag{20}$$

where $\delta_1 = \mathrm{tr}[(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})]$. From equation (20) we know that

$$\hat{\sigma}^2 = \frac{\mathrm{RSS_g}}{\delta_1} \tag{21}$$

is an unbiased estimate of σ^2. From equation (19) $\mathrm{RSS_g}$ can be represented as a quadratic form of normal variables with a symmetric and positive semidefinite matrix $(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})$. It is well know that a quadratic form of standardized normal variables $\boldsymbol{\xi}^{\mathrm{T}}\mathbf{A}\,\boldsymbol{\xi}$ [that is, $\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{A}$ is symmetric] is distributed as χ^2 distribution if and only if $\mathbf{A}$ is idempotent (for example, see Johnson and Kotz, 1970, chapter 29; Stuart and Ord, 1994, chapter 15). For the random variable

$$\frac{\mathrm{RSS_g}}{\sigma^2} = \left(\frac{\boldsymbol{\epsilon}}{\sigma}\right)^{\mathrm{T}}(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\frac{\boldsymbol{\epsilon}}{\sigma}, \tag{22}$$

although $\epsilon/\sigma \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$, the matrix $(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})$ is generally not idempotent because of the complexity of the weighting matrix $\mathbf{W}(i)$ (it is different at each location i). Therefore $\mathrm{RSS}_{\mathrm{g}}/\sigma^2$ is generally not distributed as an exact χ^2 distribution. However, there are several ways to approximate the distribution of the quadratic form $\boldsymbol{\xi}^{\mathrm{T}}\mathbf{A}\,\boldsymbol{\xi}$ with normal variables $\boldsymbol{\xi}$ (for example, see Johnson and Kotz, 1970, chapter 29 and the related references; Solomon and Stephens, 1977; 1978; Stuart and Ord, 1994, chapter 15). One simple but accurate method is to approximate the distribution of this quadratic form by that of a constant c multiplied by a χ^2 variable χ_r^2 with r degrees of freedom if the matrix $\mathbf{A}$ is symmetric and positive semidefinite, where c and r are chosen in such a way that the first two moments of $c\chi_r^2$ are made to agree with those of the quadratic form, or equally, the mean and variance of $c\chi_r^2$ and those of the quadratic form are made to match each other. We know from equation (20) that the mean of $\mathrm{RSS}_{\mathrm{g}}/\sigma^2$ is δ_1. Using our knowledge of matrix algebra, we can obtain the variance of $\mathrm{RSS}_{\mathrm{g}}/\sigma^2$ as $2\delta_2$ (see appendix A), where

$$\delta_2 = \mathrm{tr}[(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})]^2\,. \tag{23}$$

It is a standard result that the mean and the variance of the random variable χ_r^2 are r and $2r$, respectively. So, the mean and variance of $c\chi_r^2$ are cr and $2c^2r$, respectively. According to the approximating rule stated above, let

$$\begin{cases} cr = \delta_1\,, \\ 2c^2r = 2\delta_2\,. \end{cases} \tag{24}$$

Solving the equations, we obtained $c = \delta_2/\delta_1$, and $r = \delta_1^2/\delta_2$. So, the distribution of $\mathrm{RSS}_{\mathrm{g}}/c\sigma^2 = \delta_1^2\hat{\sigma}^2/\delta_2\sigma^2$ can well be approximated by a χ^2 distribution with δ_1^2/δ^2 degrees of freedom.

3.2 Testing for goodness of fit

Using the residual sum of squares and its approximated distribution, we can test whether a GWR model describes a given data set significantly better than an OLR model does. If a GWR model is used to fit the data, under assumption 2, the residual sum of squares can be expressed as $\mathrm{RSS}_{\mathrm{g}} = \boldsymbol{Y}^{\mathrm{T}}(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\boldsymbol{Y}$, and the distribution of $\delta_1\mathrm{RSS}_{\mathrm{g}}/\delta_2\sigma^2$ can well be approximated by a χ^2 distribution with δ_1^2/δ_2 degrees of freedom. If an OLR model is used to fit the data, the residual sum of squares is $\mathrm{RSS}_{\mathrm{o}} = \boldsymbol{Y}^{\mathrm{T}}(\mathbf{I}-\mathbf{Q})\boldsymbol{Y}$, where $\mathbf{Q} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, and $\mathbf{I}-\mathbf{Q}$ is idempotent. So, $\mathrm{RSS}_{\mathrm{o}}/\sigma^2$ is exactly distributed as a χ^2 distribution with $n-p-1$ degrees of freedom (for example, see Neter et al, 1989; Hocking, 1996).

If the null hypothesis—H_0: there is no significant difference between OLR and GWR models for the given data—is true, the quantity $\mathrm{RSS}_{\mathrm{g}}/\mathrm{RSS}_{\mathrm{o}}$ is close to one. Otherwise, it tends to be small. Let

$$F_1 = \frac{\mathrm{RSS}_{\mathrm{g}}/\delta_1}{\mathrm{RSS}_{\mathrm{o}}/(n-p-1)}\,. \tag{25}$$

Then a small value of F_1 supports the alternative hypothesis that the GWR model has a better goodness of fit. On the other hand, the distribution of F_1 may reasonably be approximated by an F-distribution with δ_1^2/δ_2 degrees of freedom in the numerator and $n-p-1$ degrees of freedom in the denominator. Given a significance level α, we denote by $F_{1-\alpha}(\delta_1^2/\delta_2,\, n-p-1)$ the upper $100(1-\alpha)$ percentage point. If $F_1 < F_{1-\alpha}(\delta_1^2/\delta_2,\, n-p-1)$, we reject the null hypothesis and conclude that the GWR model describes the data significantly better than the OLR model does. Otherwise, we will say that the GWR model cannot improve the fit significantly compared with the OLR model.

Another possibility to test the goodness of fit is to use the method of analysis of variance. Let

$$\mathrm{DSS} = \mathrm{RSS_o} - \mathrm{RSS_g} = \boldsymbol{Y}^{\mathrm{T}}[(\mathbf{I}-\mathbf{Q})-(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})]\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}\mathbf{A}\boldsymbol{Y}, \tag{26}$$

where $\mathbf{A}$, $\mathbf{A} = (\mathbf{I}-\mathbf{Q})-(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})$, is positive semidefinite because DSS $\geqslant 0$ for any $\boldsymbol{Y}$. This quantity measures the difference of the residual sums of squares. If the GWR model cannot significantly improve the goodness of fit compared with the OLR model, the quantity DSS must be sufficiently small. Under assumption 1 and the null hypothesis that the GWR model and the OLR model describe the data equally well, we can assume that the fitted values of y_i by the GWR model and the OLR model are all unbiased estimates of $\mathrm{E}(y_i)$ for all i. Therefore, by the aforementioned method, DSS can be expressed as $\mathrm{DSS} = \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{A}\boldsymbol{\epsilon}$ and the distribution of $v_1\mathrm{DSS}/v_2\sigma^2$ can be approximated by a χ^2 distribution with v_1^2/v_2 degrees of freedom, where $v_i = \mathrm{tr}(\mathbf{A}^i)$ $(i = 1, 2)$. Using this fact, we can construct a test statistic as follows:

$$F_2 = \frac{\mathrm{DSS}/v_1}{\mathrm{RSS_o}/(n-p-1)}. \tag{27}$$

A small value of F_2 will support the null hypothesis. Once again, the distribution of F_2 may be approximated by an F-distribution with v_1^2/v_2 degrees of freedom in the numerator and $n-p-1$ degrees of freedom in the denominator. Given a significance level α, we reject the null hypothesis if $F_2 > F_\alpha(v_1^2/v_2,\ n-p-1)$, where $F_\alpha(v_1^2/v_2,\ n-p-1)$ is the upper 100α percentage point of the F-distribution.

For simplicity of calculation, v_1 and v_2 can be expressed in a more explicit way. Indeed, it is easy to prove that (see the detailed proof in appendix B)

$$\begin{cases} \mathbf{LQ} = \mathbf{Q}, \\ \mathbf{A}^2 = (\mathbf{I}-\mathbf{Q})-2(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})+[(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})]^2. \end{cases} \tag{28}$$

Therefore, we have

$$\begin{cases} v_1 = \mathrm{tr}(\mathbf{A}) = n-p-1-\delta_1, \\ v_2 = \mathrm{tr}(\mathbf{A}^2) = n-p-1-2\delta_1+\delta_2. \end{cases} \tag{29}$$

3.3 A stepwise procedure for the selection of independent variables

In practice, we frequently suspect that a large number of independent variables might have some effect on the dependent variable. Nevertheless, too many independent variables included in the model make the explanation of the model superfluous and the computation unnecessarily complex. Furthermore, simulation results in Cleveland et al (1988) showed that the distributional approximations proposed above performs well when the number of independent variables in the model is small, but this approximation may become poor as the number of independent variables increases. So, it is essential to choose in the model calibration process independent variables that are important and exclude those that are less important.

An independent variable is said to be important if the residual sum of squares is significantly reduced when it is added to the model. Taking the ratio of the residual sums of squares before and after an independent variable is added to the model as a measurement of the change in the residual sum of squares due to the model, we propose a stepwise procedure for choosing important independent variables in calibrating a GWR model. The procedure is, to some extent, similar to that used in the building of an OLR model (for example, see Neter et al, 1989). It is a combined utilization of the forward selection procedure and the backward elimination procedure parallel to that of the OLR model. To facilitate our discussion, we first propose and discuss the forward and backward procedures.

The forward selection procedure can be structured as follows.

(1) Fit a GWR model with only the intercept terms as follows:

$$y_i = \beta_{i0} + \epsilon_i \,. \tag{30}$$

For a given weighting matrix, the parameters are estimated by equation (6) and the residual sum of squares is calculated by equation (17). In this special case, the estimated parameters are given by

$$\hat{\beta}_{i0} = \sum_{j=1}^{n} \mathrm{w}_j(i) y_j \Big/ \sum_{j=1}^{n} \mathrm{w}_j(i), \qquad i = 1, 2, ..., n, \tag{31}$$

and the residual sum of squares is given by

$$\mathrm{RSS}(0) = \boldsymbol{Y}^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(0)]^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(0)]\boldsymbol{Y}, \tag{32}$$

where

$$\mathbf{L}(0) = \begin{pmatrix} \dfrac{\mathrm{w}_1(1)}{\sum \mathrm{w}_j(1)} & \dfrac{\mathrm{w}_2(1)}{\sum \mathrm{w}_j(1)} & \cdots & \dfrac{\mathrm{w}_n(1)}{\sum \mathrm{w}_j(1)} \\ \dfrac{\mathrm{w}_1(2)}{\sum \mathrm{w}_j(2)} & \dfrac{\mathrm{w}_2(2)}{\sum \mathrm{w}_j(2)} & \cdots & \dfrac{\mathrm{w}_n(2)}{\sum \mathrm{w}_j(2)} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\mathrm{w}_1(n)}{\sum \mathrm{w}_j(n)} & \dfrac{\mathrm{w}_2(n)}{\sum \mathrm{w}_j(n)} & \cdots & \dfrac{\mathrm{w}_n(n)}{\sum \mathrm{w}_j(n)} \end{pmatrix}, \qquad j = 1, 2, ..., n \,. \tag{33}$$

(2) Let $x_1, x_2, ..., x_p$ be all the candidate independent variables. For each x_k, fit the following model:

$$y_i = \beta_{i0} + \beta_{ik} x_{ik} + \epsilon_i \,. \tag{34}$$

The parameters are still estimated by equation (6), but now $\mathbf{X}$ is an $n \times 2$ matrix with all of the elements in the first column being 1 and the second column being the n observations of x_k. Calculate the residual sum of squares $\mathrm{RSS}(x_k)$ from

$$\mathrm{RSS}(\boldsymbol{x}_k) = \boldsymbol{Y}^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(x_k)]^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(x_k)]\boldsymbol{Y}, \tag{35}$$

where $\mathbf{L}(\boldsymbol{x}_k)$ is similar to equation (16) in which the matrix $\mathbf{X}$ and the corresponding row vectors $\boldsymbol{x}_i^{\mathrm{T}}$ ($i = 1, 2, ..., n$) have to be changed to the corresponding matrix stated above.

(3) For each $k = 1, 2, ..., p$, calculate the following quantity

$$\mathrm{R}(\boldsymbol{x}_k) = \frac{\mathrm{RSS}(x_k)/\delta_1(x_k)}{\mathrm{RSS}(0)/\delta_1(0)}, \qquad k = 1, 2, ..., p, \tag{36}$$

where

$$\begin{cases} \delta_i(x_k) = \mathrm{tr}\{[\mathbf{I} - \mathbf{L}(x_k)]^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(x_k)]\}^i, \\ \delta_i(0) = \mathrm{tr}\{[\mathbf{I} - \mathbf{L}(0)]^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(0)]\}^i, \end{cases} \qquad i = 1, 2 \,. \tag{37}$$

If adding x_k to the model in equation (30) does not reduce the residual sum of squares significantly, that is, the models in equations (30) and (34) fit the data equally well, we may assume that the mean of the fitted values of y_i in model (30) and in model (34) are all equal to $\mathrm{E}(y_i)$. Under this assumption, as aforementioned, the distribution of each $\mathrm{R}(x_k)$ can be approximated by the distribution $F[\delta_1^2(x_k)/\delta_2(x_k), \delta_1^2(0)/\delta_2(0)]$. Unlike the statistics for choosing independent variables in the OLR model, the degrees of freedom of the approximated distributions of $\mathrm{R}(x_k)$ are related to x_k ($k = 1, 2, \ldots, p$). So, the curves of the approximated distribution functions are different from one another. In this case, it is inappropriate to choose important x_k by comparing the quantity of $\mathrm{R}(x_k)$ because the smallest $\mathrm{R}(x_k)$ does not necessarily correspond to the most significant reduction in the residual sum of squares. But we may use the so called p-values of the statistics to choose important independent variables. In this case, the p-value of $\mathrm{R}(x_k)$ is the probability that a random variable distributed as $F[\delta_1^2(x_k)/\delta_2(x_k),\ \delta_1^2(0)/\delta_2(0)]$ takes on the value less than or equal to the observed value of $\mathrm{R}(x_k)$ (the general concept of p-value will be further explained in section 5).

Let ξ_k be a random variable with distribution $F[\delta_1^2(x_k)/\delta_2(x_k),\ \delta_1^2(0)/\delta_2(0)]$. For each $k = 1, 2, \ldots, p$, calculate the p-values:

$$p_k = \mathrm{P}[\xi_k \leqslant \mathrm{R}(x_k)], \qquad k = 1, 2, \ldots, p\,. \tag{38}$$

Suppose

$$p_{k_0} = \min_{1 \leqslant k \leqslant p} \{p_k\}\,. \tag{39}$$

For a given significance level α, if $p_{k_0} < \alpha$, we first choose x_{k_0} to enter the model. The reason is that the smallest p-value corresponds to the most significant reduction in the residual sum of squares. Otherwise, there are no independent variables which can be regarded as important for enhancing the fit of the GWR model.

(4) Under the condition of $p_{k_0} < \alpha$, add each of the x_k ($k \neq k_0$) to the model

$$y_i = \beta_{i0} + \beta_{ik_0} x_{ik_0} + \epsilon_i\,. \tag{40}$$

Calibrate the model and calculate the residual sum of squares

$$\mathrm{RSS}(x_{k_0}, x_k) = \boldsymbol{Y}^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(x_{k_0}, x_k)]^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(x_{k_0}, x_k)]\boldsymbol{Y}, \qquad k \neq k_0\,. \tag{41}$$

Let

$$\mathrm{R}(x_{k_0}, x_k) = \frac{\mathrm{RSS}(x_{k_0}, x_k)/\delta_1(x_{k_0}, x_k)}{\mathrm{RSS}(x_{k_0})/\delta_1(x_{k_0})}, \qquad k \neq k_0\,, \tag{42}$$

where the meanings of $\mathbf{L}(x_{k_0}, x_k)$, $\delta_i(x_{k_0}, x_k)$, etc are similar to those in the previous discussion. Compute the p-values

$$p_k = \mathrm{P}[\xi_k \leqslant \mathrm{R}(x_{k_0}, x_k)], \qquad k \neq k_0\,, \tag{43}$$

where $\xi_k (k \neq k_0)$ is distributed as

$$F\left[\frac{\delta_1^2(x_{k_0}, x_k)}{\delta_2(x_{k_0}, x_k)}, \frac{\delta_1^2(x_{k_0})}{\delta_2(x_{k_0})}\right]. \tag{44}$$

Let

$$p_{k_1} = \min_{k \neq k_0}\{p_k\}\,. \tag{45}$$

If $p_{k_1} < \alpha$, enter x_{k_1} into the model in equation (40). Otherwise, equation (40) is the final model.

Repeat step 4 until no independent variables among the candidate variables can be entered into the model, and the model at this point is the final model.

The main advantage of this procedure is its simplicity of calculation. The disadvantage of it is that, once a variable has been entered into the model, it stays in the model forever. In fact, when another variable enters the model afterwards, the variables previously entered may become insignificant in reducing the residual sum of squares. The method thus does not take into account the change in importance because of the switch in entrance order.

Unlike the forward selection procedure, the backward elimination procedure works as follows. A GWR model which includes all of the candidate independent variables is calibrated first. Then, the less important independent variables are eliminated one by one according to their significance in reducing the sum of squares with and without the variable entered into the model. That is, we first fit the model

$$y_i = \beta_{i0} + \sum_{m=1}^{p} \beta_{im} x_{im} + \epsilon_i \,. \tag{46}$$

The residual sum of squares is given by

$$\mathrm{RSS}(x_1, x_2, \ldots, x_p) = \boldsymbol{Y}^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(x_1, x_2, \ldots, x_p)]^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(x_1, x_2, \ldots, x_p)]\boldsymbol{Y}. \tag{47}$$

Then for each $k = 1, 2, \ldots, p$, fit the model

$$y_i = \beta_{i0} + \sum_{\substack{m=1 \\ m \neq k}}^{p} \beta_{im} x_{im} + \epsilon_i \,. \tag{48}$$

Its residual sum of squares is

$$\begin{aligned} &\mathrm{RSS}(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p) \\ &\quad = \boldsymbol{Y}^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)]^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)]\boldsymbol{Y}. \end{aligned} \tag{49}$$

Calculate

$$\mathrm{R}(\boldsymbol{x}_k) = \frac{\mathrm{RSS}(x_1, \ldots, x_p)/\delta_1(x_1, \ldots, x_p)}{\mathrm{RSS}(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)/\delta_1(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)}, \tag{50}$$

where

$$\left\{ \begin{aligned} &\delta_i(x_1, x_2, \ldots, x_p) = \mathrm{tr}\{[\mathbf{I} - \mathbf{L}(x_1, x_2, \ldots, x_p)]^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(x_1, x_2, \ldots, x_p)]\}^i, \\ &\delta_i(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p) \\ &\quad = \mathrm{tr}\{[\mathbf{I} - \mathbf{L}(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)]^{\mathrm{T}}[\mathbf{I} - \mathbf{L}(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)]\}^i, \end{aligned} \right. \tag{51}$$

and $i = 1, 2$.

Like the forward selection procedure, we still use the p-value of $\mathrm{R}(\boldsymbol{x}_k)$ as a standard to choose independent variables. Calculate the probability

$$p_k = \mathrm{P}[\xi_k \leqslant \mathrm{R}(x_k)]\,, \tag{52}$$

where ξ_k is a random variable distributed as

$$F\left[\frac{\delta_1^2(x_1, \ldots, x_p)}{\delta_2(x_1, \ldots, x_p)}, \frac{\delta_1^2(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)}{\delta_2(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)}\right]. \tag{53}$$

Suppose

$$p_{k_0} = \max_{1 \leqslant k \leqslant p} \{p_k\}\,. \tag{54}$$

For a given significance level α, if $p_{k_0} \geqslant \alpha$, then remove x_{k_0} from the model in equation (46). Otherwise, keep all the independent variables in the model. This process is repeated until no variables can be eliminated. The drawback of this method is that, once a variable has been deleted, it cannot reenter the model. Again, the order of deleting is not dealt with in this method.

The drawbacks of the forward selection and backward elimination procedures can be overcome by a stepwise procedure parallel to that in OLR. It is actually a combination of the two procedures just described. It starts with the model in equation (30) and selects variables one by one to enter the model in the manner of the forward selection procedure. But once a new variable has been entered into the model, the variables that have been entered in the previous steps are checked one by one again for their significance in the manner of the backward elimination procedure. When a previously entered variable becomes insignificant, it will be eliminated. Otherwise, it will be kept in the model. This process is repeated until no other variables can be added to the model and also no variables can be deleted from the model.

Remark. In practical applications, the parameter in the weighting matrix may be unknown. This parameter should be determined before the procedure for choosing important variables is performed. In this case, we may first fit a GWR model with all the candidate independent variables and estimate the parameter in the weighting matrix by the cross-validation approach stated in section 2.3. After the process of choosing important independent variables has been completed, the cross-validation method is used again to refine this parameter, but now the model used includes only those independent variables that have been chosen. After redetermining the parameter in the weighting matrix, we can fit such a GWR model that includes only the important independent variables as our final model.

Given that the main purpose of the GWR technique is to explore spatial nonstationarity among the model parameters, the proposed stepwise methods may be thought of as a by-product when we can find the approximated distribution of the ratio of residual sum of squares. These model selection methods are useful when the fitted model is used for explanation and prediction. On the other hand, from the exploration point of view, these methods may also be used for certain exploratory analysis because the process of selecting an independent variable is indeed the test of whether or not the coefficient of this variable is zero. If a variable is not included in the selected model, it means that the coefficient of this independent variable is not only constant but is also zero at some significance level.

4 Test for variation of each set of parameters

After a final model has been selected, we can further test whether or not each set of parameters in the model varies significantly across the study region. For example, if the set of parameters $\{\beta_{ik};\ i = 1, 2, ..., n\}$ of x_k (if $k = 0$, the parameters examined correspond to the intercept terms) is tested not to vary significantly over the region, we can treat the coefficient of x_k to be constant and conclude that the slope between $\boldsymbol{x}_k$ and the dependent variable is uniform over the area when the other variables are taken to be fixed. Statistically, it is equivalent to testing the hypotheses

$$\mathrm{H}_0\colon\ \beta_{1k} = \beta_{2k} = ... = \beta_{nk}\ \textit{for a given } k\,,$$

$$\mathrm{H}_1\colon\ \textit{not all}\ \beta_{ik}\ (i = 1, 2, ..., n)\ \textit{are equal}\,.$$

First, we must construct an appropriate statistic which can reflect the spatial variation of the given set of parameters. A practical and yet natural choice is the sample variance of the estimated values of β_{ik} $(i = 1, 2, ..., n)$ (Brunsdon et al, 1996; 1997). We denote by V_k^2 the sample variance of the n estimated values, $\hat{\beta}_{ik}$ $(i = 1, 2, ..., n)$ for the kth parameter. Then

$$V_k^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_{ik} - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_{ik}\right)^2, \tag{55}$$

where $\hat{\beta}_{ik}$ $(i = 1, 2, ..., n)$ are obtained from equation (6).

The next stage is to determine the sampling distribution of V_k^2 under the null hypotheses H_0. Let $\hat{\boldsymbol{\beta}}_k = (\hat{\beta}_{1k}\ \hat{\beta}_{2k}\ ...\ \hat{\beta}_{nk})^{\mathrm{T}}$ and let $\mathbf{J}$ be an $n \times n$ matrix with unity for each of its elements. Then V_k^2 can be expressed as

$$V_k^2 = \frac{1}{n}\hat{\boldsymbol{\beta}}_k^{\mathrm{T}}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\hat{\boldsymbol{\beta}}_k \,. \tag{56}$$

Under the null hypothesis that all the β_{ik} ($i = 1, 2, ..., n$) are equal, we may assume that the means of the corresponding estimated parameters are equal, that is,

$$\mathrm{E}(\hat{\beta}_{1k}) = \mathrm{E}(\hat{\beta}_{2k}) = ... = \mathrm{E}(\hat{\beta}_{nk}) = \mu_k \,. \tag{57}$$

So,

$$\mathrm{E}(\hat{\beta}_k) = \mu_k \boldsymbol{1}\,, \tag{58}$$

where $\boldsymbol{1}$ is a column vector with unity for each element. From equation (58) and the fact that $\boldsymbol{1}^{\mathrm{T}}(\mathbf{I} - \frac{1}{n}\mathbf{J}) = \boldsymbol{0}$ and $(\mathbf{I} - \frac{1}{n}\mathbf{J})\boldsymbol{1} = \boldsymbol{0}$, we can further express V_k^2 as

$$V_k^2 = \frac{1}{n}[\hat{\boldsymbol{\beta}}_k - \mathrm{E}(\hat{\boldsymbol{\beta}}_k)]^{\mathrm{T}}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)[\hat{\boldsymbol{\beta}}_k - \mathrm{E}(\hat{\boldsymbol{\beta}}_k)]\,. \tag{59}$$

Furthermore, let $\boldsymbol{e}_k$ be a column vector with unity for the $(k+1)$th element and zero for other elements. Then

$$\hat{\beta}_{ik} = \boldsymbol{e}_k^{\mathrm{T}}\hat{\boldsymbol{\beta}}(i) = \boldsymbol{e}_k^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(i)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(i)\boldsymbol{Y}, \tag{60}$$

and

$$\hat{\boldsymbol{\beta}}_k = (\hat{\beta}_{1k}\ \hat{\beta}_{2k}\ ...\ \hat{\beta}_{nk})^{\mathrm{T}} = \mathbf{B}\boldsymbol{Y}, \tag{61}$$

where

$$\mathbf{B} = \begin{pmatrix} \boldsymbol{e}_k^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(1)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(1) \\ \boldsymbol{e}_k^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(2)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(2) \\ \vdots \\ \boldsymbol{e}_k^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(n)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(n) \end{pmatrix}. \tag{62}$$

Substituting equation (61) into equation (59), we obtain

$$\begin{aligned} V_k^2 &= \frac{1}{n}[\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})]^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{B}[\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})] \\ &= \boldsymbol{\epsilon}^{\mathrm{T}}\left[\frac{1}{n}\mathbf{B}^{\mathrm{T}}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{B}\right]\boldsymbol{\epsilon}, \end{aligned} \tag{63}$$

where $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\mathbf{I})$, and $\frac{1}{n}\mathbf{B}^{\mathrm{T}}(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{B}$ is positive semidefinite.

Similar to the method employed in section 3, the distribution of $\gamma_1 V_k^2/\gamma_2\sigma^2$ can be approximated by a χ^2 distribution with γ_1^2/γ_2 degrees of freedom, where

$$\gamma_i = \mathrm{tr}\left[\frac{1}{n}\mathbf{B}^{\mathrm{T}}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{B}\right]^i, \qquad i = 1, 2\,. \tag{64}$$

Because σ^2 is unknown, we cannot use $\gamma_1 V_k^2/\gamma_2\sigma^2$ as a test statistic directly. But from section 3, we know that the distribution of $\delta_1^2\hat{\sigma}_2/\delta_2\sigma^2$ can be approximated by a χ^2 distribution with δ_1^2/δ_2 degrees of freedom, where $\hat{\sigma}^2$ is an unbiased estimator of σ^2 which is defined by equation (21), and $\delta_i = \mathrm{tr}[(\mathbf{I} - \mathbf{L})^{\mathrm{T}}(\mathbf{I} - \mathbf{L})]^i$ ($i = 1, 2$). So, for the statistic

$$F_3(k) = \frac{V_k^2/\gamma_1}{\hat{\sigma}^2}, \tag{65}$$

under the assumption in equation (57), its distribution can be approximated by an F-distribution with γ_1^2/γ_2 degrees of freedom in the numerator and δ_1^2/δ_2 degrees of freedom in the denominator. Therefore, we can take F_3 as a test statistic. The large value of F_3 supports the alternative hypothesis H_1. For a given significance level α, find the upper 100α percentage point $F_\alpha(\gamma_1^2/\gamma_2, \delta_1^2/\delta_2)$. If $F_3 \geqslant F_\alpha(\gamma_1^2/\gamma_2, \delta_1^2/\delta_2)$, reject H_0; accept H_0 otherwise.

5 Simulations

In this section, several simulated data sets with known parameters are used to examine the test power of the statistics proposed in sections 3 and 4.

The models for generating the data sets are taken from Foster and Gorr (1986) where the data were used to compare the percentage improvement in fit and bias of the spatial adaptive filter over the ordinary least squares regression estimates. The spatial region of parameter variation consists of coordinates (z_{i1}, z_{i2}) taken from a uniform, two-dimensional grid consisting of $m \times m$ lattice points with unit distance between any two of them along the horizontal and vertical axes (as illustrated in figure 1). That is, the points for taking the observations are at the following lattice points:

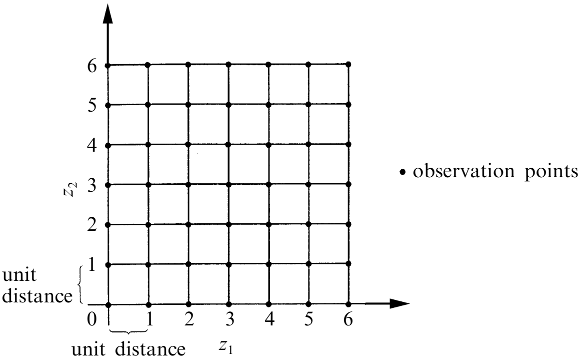

Figure 1. A grid for generating the simulation data set: (the case of $m = 7$).

$$\{(z_{i1}, z_{i2});\; z_{i1}, z_{i2} = 0, 1, 2, \ldots, m-1\}. \tag{66}$$

The general form of the model being simulated is as follows:

$$y_i = \beta_{i0} + \beta_{i1} x_i + \epsilon_i, \tag{67}$$

where $\epsilon_i \sim N(0, 4)$. The values of the independent variable x are drawn randomly from a uniform distribution on interval (0, 1) and therefore contain no information on spatial variation of parameters. Besides the five particular cases in Foster and Gorr (1986), two other cases (linear change in both the intercept and slope, and step change in both the intercept and slope) are also considered. We list these seven cases as follows:

Case 1. Stationary: $\beta_{i0} = 10$; $\beta_{i1} = 15$.

Case 2. Linear change in intercept: $\beta_{i0} = 8.333 + 1.667 z_{i1}$; $\beta_{i1} = 10$.

Case 3. Linear change in slope: $\beta_{i0} = 5$; $\beta_{i1} = 12.5 + 2.5 z_{i1}$.

Case 4. Linear change in both intercept and slope: $\beta_{i0} = 8.333 + 1.667 z_{i1}$; $\beta_{i1} = 12.5 + 2.5 z_{i1}$.

Case 5. Step change in intercept: $\beta_{i0} = 10$ if $z_{i1} < [\frac{m}{2}]$, and $\beta_{i0} = 15$ otherwise, where $[\frac{m}{2}]$ represents the integer part of $\frac{m}{2}$; $\beta_{i1} = 10$.

Case 6. Step change in slope: $\beta_{i0} = 5$; $\beta_{i1} = 20$ if $z_{i1} < [\frac{m}{2}]$, and $\beta_{i1} = 30$ otherwise.

Case 7. Step change in both intercept and slope: $\beta_{i0} = 10$ if $z_{i1} < [\frac{m}{2}]$, and $\beta_{i0} = 15$ otherwise; $\beta_{i1} = 20$ if $z_{i1} < [\frac{m}{2}]$, and $\beta_{i1} = 30$ otherwise.

In each case, we take $m = 6, 7, 8, 9, 10$ and perform the simulations. For a given m, the values x_i are drawn from the uniform distribution on interval (0, 1); the error terms ϵ_i $(i = 1, 2, \ldots, m^2)$ are drawn from N(0, 4) and the corresponding values of the dependent variable y_i are generated by the model in equation (67). The Gaussian weighting function in equation (9) is used and the parameter θ in the weighting function is first determined by the cross-validation approach stated in section 2.3. Then, we fit the GWR model and calculate the values of the test statistics F_1 in equation (25), F_2 in equation (27), $F_3(k)$ in equation (65) and the corresponding degrees of freedom. We use the p-values of the statistics to show the degree of significance. For a one-tailed test, the p-value of the test statistic is the probability that the statistic could have been more extreme than its observed value under the null hypothesis. A large p-value supports the null hypothesis whereas a small p-value supports the alternative hypothesis. A test can be carried out by comparing the p-value with a given significance level α. If the p-value is less than α, we reject the null hypothesis; we accept it otherwise. It should be noted that the p-value is now widely employed in applied statistics because of its ease in use. For our proposed statistics, the p-values for F_1, F_2, and $F_3(k)$ are respectively:

$$\begin{cases} p_1 &= \mathrm{P}(F_1 \leqslant f_1)\,, \\ p_2 &= \mathrm{P}(F_2 \geqslant f_2)\,, \\ p_3(k) &= \mathrm{P}[F_3(k) \geqslant f_1(k)]\,, \end{cases} \tag{68}$$

where f_1, f_2, and $f_3(k)$ are the observed values of the test statistics F_1, F_2 and $F_3(k)$, respectively.

5.1 Model validity

In the first case, that is, the stationary one, for each value of $m = 6, 7, 8, 9, 10$, we get $\theta = 0$ by the cross-validation method. Therefore the weighting matrix $\mathbf{W}(i)$ equal $\mathbf{I}$, an identity matrix of order n, at any location i. The estimates of the parameters at any location i are identical and coincide with the estimates of the parameters by ordinary least squares method in the OLR model. So, we have $\mathrm{RSS_g} = \mathrm{RSS_o}$, and $V_0^2 = V_1^2 = 0$ in this case. If the proposed test statistics are used, we have $\delta_1 = \delta_2 = n - 2$, and $F_1 = 1$, $F_3(0) = F_3(1) = 0$. The corresponding p-values are $p_1 = 0.5$, $p_3(0) = p_3(1) = 1$, respectively, and therefore we can also conclude from the p-values that the GWR model and the OLR model describe the generated data set equally well and that the intercept and the slope of the model do not vary significantly. This correctly reflects the properties of the true model, that is, the OLR model with constant intercept and slope. But in the case of $\theta = 0$, the statistic F_2 is invalid because $v_1 = 0$.

In the other cases and for all $m = 6, 7, 8, 9, 10$, the simulations we have done show that, if variations in the parameters do exist, the p-values of the corresponding test statistics are generally very small, and, if the intercept or the slope is constant, the p-values of the corresponding test statistics are generally large compared with the p-values for the situations of varying parameters. Generally speaking, when variations do exist, the p-values of the statistics become smaller when m becomes larger. The sensitivity of the statistics proposed to explore spatial variations in the parameters and the test power are rather high. Also, from a large number of the simulation runs, it seems that the sensitivity of F_1 for exploring spatial variations of the parameters is higher than that of F_2. Because of the limitation of space, we report only the simulation results for $m = 7$ (the same spatial grid as in Foster and Gorr, 1986) for cases 2 to 7

Table 1. Simulation results of the 7 × 7 lattice-point experiment.

Case	Statistic	Value	NDF[a]	DDF[a]	p-value
2	F_1	0.3263	37.71	47	0.0003
	F_2	2.3422	23.76	47	0.0064
	$F_3(0)$	6.0615	5.75	37.71	0.0002
	$F_3(1)$	1.2303	5.88	37.71	0.3130
3	F_1	0.3736	36.97	47	0.0013
	F_2	2.0999	25.60	47	0.0136
	$F_3(0)$	0.7491	6.27	36.97	0.6193
	$F_3(1)$	5.2210	6.41	36.97	0.0004
4	F_1	0.1271	32.42	47	0.0000
	F_2	1.7125	36.16	47	0.0416
	$F_3(0)$	4.6739	10.43	32.42	0.0003
	$F_3(1)$	4.4095	10.40	32.42	0.0005
5	F_1	0.5654	41.23	47	0.0322
	F_2	2.6808	15.20	47	0.0049
	$F_3(0)$	3.6475	3.85	41.23	0.0133
	$F_3(1)$	1.1953	3.93	41.23	0.3271
6	F_1	0.4605	38.49	47	0.0074
	F_2	2.2330	21.81	47	0.0106
	$F_3(0)$	0.4711	5.25	38.49	0.8036
	$F_3(1)$	5.0106	5.37	38.49	0.0010
7	F_1	0.2702	35.29	47	0.0001
	F_2	1.9654	29.74	47	0.0186
	$F_3(0)$	1.8988	7.65	35.29	0.0939
	$F_3(1)$	3.6154	7.78	35.29	0.0039

[a] NDF and DDF represent the degrees of freedom of the numerator and denominator, respectively, of the corresponding F-distributions.

in table 1. The results for m other than 7 are similar and generally become better with increasing m.

The p-values in table 1 correctly reflect the properties of the true models for the significance level $\alpha = 0.05$, except for case 7 where the p-value of $F_3(0)$ is 0.0939. We see that the p-values of F_1 and F_2, which are used to test whether the GWR model fits the data set significantly better than the OLR model does, are all smaller than 0.05 in each case. It correctly reflects the real characteristics of the true models that each case indeed includes at least one set of spatially varying parameters. The p-values of $F_3(0)$ and $F_3(1)$, which are respectively used to explore the variations in the intercept and slope, are all larger than 0.3 when the corresponding parameters are constant and are all smaller than 0.02, except for the p-value of $F_3(0)$ in case 7[(1)], when the corresponding parameters do vary. The simulations show that the proposed statistics can correctly reflect the true properties of the model with rather high test power even with moderate sample size.

5.2 Stability of the test statistics

In order to confirm the stability of the test statistics for exploring spatial variations, we generated several sets of observations of the dependent and independent variables in each case and for different ms to perform the simulations. Although the observed

[(1)] It might be caused by the randomness of the observations or the extent of variation of the intercept compared with that of the slope. This p-value is 0.0513 for $m = 6$; 0.0187 for $m = 8$; 0.0001 for $m = 9$; and 0.0000 for $m = 10$, respectively.

values of the test statistics, the degrees of freedom of the corresponding approximated distributions, and the p-values vary from one set of observations to another, the results are all similar to those reported in table 1, and the difference between the p-values of the statistics corresponding to the varying parameters and nonvarying parameters situations is large enough to support our claim that they correctly reflect the properties of the true models under some specified significance levels.

5.3 Robustness of the p-values

In the simulations, we also examine the robustness of the p-values of the statistics with respect to the variation of the parameter θ in the weighting function. We find that the p-values of the statistics are rather robust when the parameter θ varies around its 'best value' determined by the cross-validation approach. In the stationary case (case 1) of $\beta_{i0} = 10$ and $\beta_{i1} = 15$, all the p-values of the statistics are very robust to the variation of θ. As previously mentioned, the best value of θ is zero. Even if we take $\theta = 1.1$ or larger, all the p-values of the statistics are still large enough for us to conclude that the GWR model fits the data as well as the OLR model, and the parameters do not vary significantly across the study region, which are the real characteristics of the true model. When there indeed exist variations in the model parameters (all other cases), the p-values of the statistics, except for F_2, are still very robust to the variation of θ. The simulation runs have shown that the p-value of F_2 increases very fast with increasing θ when variations in model parameters do exist. If θ is too far away from the best value, the p-value of F_2 may change dramatically and sometimes a contrary conclusion may be drawn from the p-value. Because of the limitation of space, we only list in table 2 the p-values of each statistic with different values of θ in cases 1 and 2 to show their robustness.

From the simulation experience, we can say that, because the p-values are robust to the variation of θ, it may then be unnecessary to use the exact best value of θ when we perform the test in practice. For example, if we want to save computation time, we can take a larger step size to determine the best value of θ or we can even take a moderately large value as an estimate of θ to perform the test. But it is important to choose a θ which is close to the best value when we use F_2 as a test statistic, especially when variations in model parameters are likely.

Table 2. p-values of the statistics for different values of θ in the 7 × 7 lattice-point experiment for cases 1 and 2.

Case	Statistic	Best value[a]	$\theta = 0.1$	$\theta = 0.3$	$\theta = 0.5$	$\theta = 0.7$	$\theta = 0.9$	$\theta = 1.1$
1		$\theta = 0.0$						
	F_1	0.5	0.5791	0.5968	0.5722	0.5420	0.5141	0.4881
	F_2	na	0.8239	0.6772	0.5750	0.5231	0.4995	0.4891
	$F_3(0)$	1	0.8125	0.7911	0.6869	0.5849	0.5144	0.4695
	$F_3(1)$	1	0.7222	0.6451	0.5422	0.4544	0.4014	0.3689
2		$\theta = 0.34$						
	F_1	0.0003	0.0064	0.0003	0.0002	0.0002	0.0002	0.0002
	F_2	0.0064	0.0001	0.0040	0.0261	0.0767	0.1470	0.2207
	$F_3(0)$	0.0002	0.0017	0.0002	0.0002	0.0005	0.0011	0.0021
	$F_3(1)$	0.3130	0.6185	0.3345	0.2689	0.2389	0.2177	0.2003

[a] The 'best value' of θ determined by the cross-validation method.
na means the value is not available.

6 Prediction

Although the GWR technique is used mainly for exploring spatial nonstationarity of the parameters, prediction is still an important aspect in regression analysis. For the GWR model, solution of the prediction problem is straightforward. In this section, the prediction method based on the GWR model is discussed and an approximated confidence interval for the true value of the dependent variable at a new location is established.

If a GWR model can describe a given data set well, we can use it to predict the true value of the dependent variable given the observations of the independent variables at a new point in the study area. Let $(x_{01}\ x_{02}\ \dots\ x_{0p})$ be an observation of the independent variables at new location p_0 and $\boldsymbol{x}_0^{\mathrm{T}} = (1\ x_{01}\ x_{02}\ \dots\ x_{0p})$. Then the predicted value for y_0 (the true value of the dependent variable at p_0) is

$$\hat{y}_0 = \boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}(p_0)\,, \tag{69}$$

where

$$\hat{\boldsymbol{\beta}}(p_0) = [\mathbf{X}^{\mathrm{T}}\mathbf{W}(p_0)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(p_0)\boldsymbol{Y}, \tag{70}$$

and $\mathbf{W}(p_0)$ is known at this moment because the parameter in it has been determined in the model calibration process and the distances from p_0 to other points where the data are observed are also known once the point p_0 is given.

The next stage is to construct a confidence interval for y_0. As pointed out previously, a local fitting technique tends to find a low-bias estimate. We assume that $\hat{y}_0$ is an unbiased estimator of $\mathrm{E}(y_0)$, that is,

$$\mathrm{E}(\hat{y}_0) = \mathrm{E}(y_0)\,. \tag{71}$$

We know from equation (69) and (70) that $\hat{y}_0$ is not related to y_0. Therefore, $\hat{y}_0$ and y_0 are independent and we have

$$\begin{aligned}\mathrm{var}(\hat{y}_0 - y_0) &= \mathrm{var}(\hat{y}_0) + \mathrm{var}(y_0)\\ &= \mathrm{var}\{\boldsymbol{x}_0^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(p_0)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(p_0)\boldsymbol{Y}\} + \sigma^2\\ &= \boldsymbol{x}_0^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(p_0)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(p_0)\mathrm{var}(\boldsymbol{Y})\mathbf{W}(p_0)\mathbf{X}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(p_0)\mathbf{X}]^{-1}\boldsymbol{x}_0 + \sigma^2\\ &= \{1 + \boldsymbol{x}_0^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(p_0)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^2(p_0)\mathbf{X}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(p_0)\mathbf{X}]^{-1}\boldsymbol{x}_0\}\sigma^2\,.\end{aligned} \tag{72}$$

For simplicity, let

$$\mathrm{S}(p_0) = \boldsymbol{x}_0^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(p_0)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^2(p_0)\mathbf{X}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(p_0)\mathbf{X}]^{-1}\boldsymbol{x}_0\,. \tag{73}$$

Under the assumption made in equation (71) and the fact that $\hat{y}_0$ and y_0 are normally distributed, we obtain

$$\frac{\hat{y}_0 - y_0}{\sigma[1 + \mathrm{S}(p_0)]^{1/2}} \sim \mathrm{N}(0,\,1)\,. \tag{74}$$

Also, from section 3.1 we know that the distribution of the statistic $\delta_1^2\hat{\sigma}^2/\delta_2\sigma^2$ can be approximated by a χ^2 distribution with δ_1^2/δ_2 degrees of freedom. So, the distribution of the statistic

$$T = \frac{\hat{y}_0 - y_0}{\hat{\sigma}[1 + \mathrm{S}(p_0)]^{1/2}} \tag{75}$$

can be approximated by a t-distribution with δ_1^2/δ_2 degrees of freedom, where $\hat{\sigma}^2$ is defined by equation (21). Given a confidence level α, we write $t_{\frac{\alpha}{2}}(\delta_1^2/\delta_2)$ as the upper $100(\frac{\alpha}{2})$ percentage point. Thus an approximated confidence interval of y_0 with confidence $1 - \alpha$ is obtained as follows:

$$\hat{y}_0 \pm \hat{\sigma}[1 + \mathrm{S}(p_0)]^{1/2} t_{\frac{\alpha}{2}}\left(\frac{\delta_1^2}{\delta_2}\right) . \tag{76}$$

7 Conclusion

The GWR model appears to be a useful means to explore variation of the slope between the dependent variable and the independent variables over space. To fill the theoretical gap in the literature, we have proposed in this paper some statistical testing methods by which the goodness of fit of the GWR model can be evaluated in comparison with the OLR model, and the variation of the parameters over space can be determined. We have also formulated a procedure for choosing significant independent variables from a large number of candidate independent variables for the model. Based on the theoretical results, we further investigated the prediction issue of the GWR model and constructed an approximate confidence interval of the true value of the dependent variable.

Because of the complexity of parameter estimation in the GWR model, it seems to be a difficult task to find the exact distributions of the proposed statistics. Also, if we take account of the fact that the parameter in the weighting function is determined by the cross-validation approach and is, therefore, related to the observations, this will make it more difficult to find the exact distributions of the testing statistics. So, from a practical point of view, finding approximated distributions is a feasible way to solve this kind of problem. We have thus chosen one simple but accurate method to approximate the distribution of the quadratic form with normal variables and positive semidefinite matrix, and the distributions of the other proposed statistics. The simplicity in computation is apparent. The simulation results have shown that the test power of the proposed statistics is rather high and their p-values are rather robust to the variation of the parameter in the weighting matrix. Nevertheless, more simulations with certain structures in the x-variables still need to be performed in further studies in order to understand the generality of the proposed approximation methods in more complex environments. Furthermore, although the proposed approximations are sensible, it is also worthwhile carrying out further research to find the exact distributions of the proposed test statistics. Based on the exact distribution theory of quadratic forms in normal variables (for example, see Imhof, 1961) which has recently been used in the test for spatial autocorrelation (Tiefelsdorf and Boots, 1995; Hepple, 1998), it might be possible to use this line of thinking to derive the exact p-values for the proposed statistics in this paper. If this possibility can be confirmed, it could then provide a useful way of checking the validity of our approximation methods. However, it is obvious that the computation overhead will be considerable.

Although some further studies are still needed on this topic, with the present theoretical investigation, not only does the GWR model allow complex spatial variations in parameters to be identified, mapped, and modeled, but we can also make it possible to perform the significance test in a conventional statistical manner.

Acknowledgement. This project is supported by the earmarked grant CUHK 4037/97H of the Hong Kong Research Grants Council.

References

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

Anselin L, 1990, "Spatial dependence and spatial structural instability in applied regression analysis" *Journal of Regional Science* **30** 185 – 207

Bowman A W, 1984, "An alternative method of cross-validation for the smoothing of density estimate" *Biometrika* **71** 353 – 360

Brunsdon C F, 1995, "Estimating probability surfaces for geographical point data: an adaptive kernel algorithm" *Computers and Geosciences* **21** 877 – 894

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

Brunsdon C, Fotheringham A S, Charlton M, 1997, "Geographical instability in linear regression modelling—a preliminary investigation", in *New Techniques and Technologies for Statistics II* (IOS Press, Amsterdam) pp 149 – 158

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

Casetti E, 1982, "Drift analysis of regression analysis: an application to the investigation of fertility development relations" *Modeling and Simulation* **13** 961 – 966

Casetti E, 1986, "The dual expansion method: an application for evaluating the effects of population growth on development" *IEEE Transactions on Systems, Man and Cybernetics* **16** 29 – 39

Charlton M, Fotheringham A S, Brunsdon C, 1997, "The geography of relationships: an investigation of spatial non-stationarity", in *Spatial Analysis of Biodemographic Data* Eds J-P Bocquet-Appel, D Courgeau, D Pumain (John Libbey Eurotext, Montrouge) pp 23 – 47

Cleveland W S, 1979, "Robust locally weighted regression and smoothing scatter-plots" *Journal of the American Statistical Association* **74** 829 – 836

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

Cleveland W S, Devlin S J, Grosse E, 1988, "Regression by local fitting: methods, properties and computational algorithms" *Journal of Econometrics* **37** 87 – 114

Dobson A J, 1990 *An Introduction to Generalized Linear Models* (Chapman and Hall, London)

Foster S A, Gorr W L, 1986, "An adaptive filter for estimating spatially-varying parameters: application to modelling police hours spent in response to calls for service" *Management Science* **32** 878 – 889

Fotheringham A S, 1997, "Trends in quantitative methods I: stressing the local" *Progress in Human Geography* **21** 88 – 96

Fotheringham A S, Pitts T C, 1995, "Directional variation in distance-decay" *Environment and Planning A* **27** 715 – 729

Fotheringham A S, Charlton M, Brunsdon C, 1996, "The geography of parameter space: an investigation into spatial non-stationarity" *International Journal of Geographical Information Systems* **10** 605 – 627

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

Fotheringham A S, Charlton M, Brunsdon C, 1997b, "Two techniques for exploring non-stationarity in geographical data" *Geographical Systems* **4** 59 – 82

Gorr W L, Olligschlaeger A M, 1994, "Weighted spatial adaptive filtering: Monte Carlo studies and application to illicit drug market modelling" *Geographical Analysis* **26** 67 – 87

Hepple L W, 1998, "Exact testing for spatial autocorrelation among regression residuals" *Environment and Planning A* **30** 85 – 109

Hocking R R, 1996 *Methods and Applications of Linear Models* (John Wiley, New York)

Imhof J P, 1961, "Computing the distribution of quadratic forms in normal variables" *Biometrika* **48** 419 – 426

Johnson N L, Kotz A B, 1970 *Continuous Univariate Distributions* (John Wiley, New York)

Jones J P, Casetti E, 1992 *Applications of the Expansion Method* (Routledge, London)

Neter J, Wasserman W, Kutner M H, 1989 *Applied Linear Regression Models* 2nd edition (Irwin, Homewood, IL)

Openshaw S, 1993, "Exploratory space – time-attribute pattern analysis", in *Spatial Analysis and GIS* Eds A S Fotheringham, P A Rogerson (Taylor and Francis, London) pp 147 – 163

Solomon H, Stephens M A, 1977, "Distribution of a sum of weighted chi-square variables" *Journal of the American Statistical Association* **72** 881 – 885

Solomon H, Stephens M A, 1978, "Approximations to density functions using Pearson curves" *Journal of the American Statistical Association* **73** 153 – 160

Stuart A, Ord J K, 1994 *Kendall's Advanced Theory of Statistics. Volume 1, Distribution Theory* 6th edition (Edward Arnold, London)

Tiefelsdorf M, Boots B, 1995, "The exact distribution of Moran's *I*" *Environment and Planning A* **27** 985 – 999

Wand M P, Jones M C, 1995 *Kernel Smoothing* (Chapman and Hall, London)

APPENDIX A

The variance of $\mathrm{RSS_g}/\sigma^2$

From equation (22),

$$\frac{\mathrm{RSS_g}}{\sigma^2} = \left(\frac{\boldsymbol{\epsilon}}{\sigma}\right)^{\mathrm{T}}(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\frac{\boldsymbol{\epsilon}}{\sigma}, \tag{A1}$$

where $\boldsymbol{\epsilon}/\sigma \sim \mathrm{N}(\boldsymbol{0}, \mathbf{I})$, and $(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})$ is a real symmetric and positive semidefinite matrix. According to the theory of matrix algebra, there exists an orthogonal matrix $\mathbf{P}$ of order n such that

$$\mathbf{P}^{\mathrm{T}}(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\mathbf{P} = \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), \tag{A2}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix having the eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_n$, of the matrix $(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})$ in its main diagonal. Let

$$\boldsymbol{\eta} = (\eta_1\ \eta_2\ \ldots\ \eta_n)^{\mathrm{T}} = \mathbf{P}^{\mathrm{T}}\frac{\boldsymbol{\epsilon}}{\sigma}. \tag{A3}$$

According to the properties of multivariate normal distribution, $\eta_1, \eta_2, \ldots, \eta_n$ are independent identically distributed (iid) random variables with common distribution N(0, 1). On the other hand, under the orthogonal transformation (A3), we obtain

$$\begin{aligned}\frac{\mathrm{RSS_g}}{\sigma^2} &= \boldsymbol{\eta}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\mathbf{P}\boldsymbol{\eta} \\ &= \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\eta} \\ &= \sum_{i=1}^{n}\lambda_i\eta_i^2 .\end{aligned} \tag{A4}$$

It is well know that η_i^2 $(i = 1, 2, \ldots, n)$ are iid random variables with common χ^2 distribution with one degree of freedom. Therefore $\mathrm{var}(\eta_i^2) = 2$, and

$$\mathrm{var}\left(\frac{\mathrm{RSS_g}}{\sigma^2}\right) = \sum_{i=1}^{n}\lambda_i^2\mathrm{var}(\eta_i^2) = 2\sum_{i=1}^{n}\lambda_i^2 . \tag{A5}$$

We know from the theory of matrix algebra that $\lambda_1^2, \lambda_2^2, \ldots, \lambda_n^2$ are eigenvalues of the matrix $[(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})]^2$. Therefore

$$\mathrm{var}\left(\frac{\mathrm{RSS_g}}{\sigma^2}\right) = 2\,\mathrm{tr}[(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})]^2 = 2\delta_2 . \tag{A6}$$

APPENDIX B

The proofs of equations (28) and (29)

According to the definition of **L** and $\mathbf{Q} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, we have

$$\begin{aligned}
\mathbf{LQ} &= \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(1)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(1) \\ \boldsymbol{x}_2^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(2)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(2) \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(n)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(n) \end{pmatrix} \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \\
&= \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(1)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(1)\mathbf{X} \\ \boldsymbol{x}_2^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(2)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(2)\mathbf{X} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{W}(n)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(n)\mathbf{X} \end{pmatrix} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \\
&= \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \\
&= \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \\
&= \mathbf{Q}\,. \qquad \text{(B1)}
\end{aligned}$$

We know from equation (26) that $\mathbf{A} = (\mathbf{I}-\mathbf{Q}) - (\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})$. So,

$$\begin{aligned}
\mathbf{A}^2 &= (\mathbf{I}-\mathbf{Q})^2 - (\mathbf{I}-\mathbf{Q})(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L}) - (\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})(\mathbf{I}-\mathbf{Q}) \\
&\quad +[(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})]^2\,. \qquad \text{(B2)}
\end{aligned}$$

Because $\mathbf{I}-\mathbf{Q}$ is idempotent,

$$(\mathbf{I}-\mathbf{Q})^2 = \mathbf{I}-\mathbf{Q}\,. \qquad \text{(B3)}$$

Also, from equation (B1) and the symmetry of $\mathbf{I}-\mathbf{Q}$, we obtain

$$\begin{aligned}
(\mathbf{I}-\mathbf{Q})(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L}) &= [(\mathbf{I}-\mathbf{L})(\mathbf{I}-\mathbf{Q})]^{\mathrm{T}}(\mathbf{I}-\mathbf{L}) \\
&= (\mathbf{I}-\mathbf{Q}-\mathbf{L}+\mathbf{LQ})^{\mathrm{T}}(\mathbf{I}-\mathbf{L}) \\
&= (\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\,, \qquad \text{(B4)}
\end{aligned}$$

and

$$(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})(\mathbf{I}-\mathbf{Q}) = (\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})\,. \qquad \text{(B5)}$$

Substituting equations (B3), (B4), and (B5) into equation (B2), we obtain the second equation in expression (28), that is,

$$\mathbf{A}^2 = (\mathbf{I}-\mathbf{Q}) - 2(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L}) + [(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})]^2\,. \qquad \text{(B6)}$$

It is well known that

$$\mathrm{tr}(\mathbf{Q}) = \mathrm{tr}[\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}] = \mathrm{tr}[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}] = p+1\,. \qquad \text{(B7)}$$

Therefore

$$\begin{cases} \operatorname{tr}(\mathbf{A}) = \operatorname{tr}[(\mathbf{I}-\mathbf{Q}) - (\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})] \\ \qquad = n-p-1-\delta_1 \,, \\ \operatorname{tr}(\mathbf{A}^2) = \operatorname{tr}\{(\mathbf{I}-\mathbf{Q}) - 2(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L}) + [(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})]^2\} \\ \qquad = n-p-1-2\delta_1+\delta_2 \,, \end{cases} \tag{B8}$$

where $\delta_i = \operatorname{tr}[(\mathbf{I}-\mathbf{L})^{\mathrm{T}}(\mathbf{I}-\mathbf{L})]^i \, (i = 1, 2)$.