Environment and Planning A 2007, volume 39, pages 2464 – 2481

DOI:10.1068/a38325

Diagnostic tools and a remedial method for collinearity in geographically weighted regression

David C Wheeler
Department of Biostatistics, Emory University, Atlanta, GA 30322, USA;
e-mail: dcwheel@sph.emory.edu
Received 24 August 2005; in revised form 12 January 2006

Abstract. Geographically weighted regression (GWR) is drawing attention as a statistical method to estimate regression models with spatially varying relationships between explanatory variables and a response variable. Local collinearity in weighted explanatory variables leads to GWR coefficient estimates that are correlated locally and across space, have inflated variances, and are at times counterintuitive and contradictory in sign to the global regression estimates. The presence of local collinearity in the absence of global collinearity necessitates the use of diagnostic tools in the local regression model building process to highlight areas in which the results are not reliable for statistical inference. The method of ridge regression can also be integrated into the GWR framework to constrain and stabilize regression coefficients and lower prediction error. This paper presents numerous diagnostic tools and ridge regression in GWR and demonstrates the utility of these techniques with an example using the Columbus crime dataset.

1 Introduction

The recognition in the social and health sciences that relationships between explanatory variables and a response variable in a regression model are not always constant across a study area has led researchers to develop models that allow for spatially varying coefficients. One relatively new method in the field of geography for modeling spatially varying coefficients is geographically weighted regression (GWR). Fotheringham et al (2002) provide a comprehensive introduction to GWR. GWR is similar to the idea of local linear regression models found in the statistics literature (Hastie et al, 2001; Loader, 1999) in that it uses a kernel function to calculate weights that are applied to observations in a series of local weighted regression models. It differs in that the kernel function is applied in geographic space instead of in explanatory variable space, as it is in local regression models. It also differs in its focus, as GWR is concerned with measuring statistically significant variation in the regression coefficients and providing an interpretation of the coefficients, while local regression is concerned with the smoothing of data with a fitted curve. When comparing local regression to traditional linear regression, Loader (1999, page 19) states, "Instead of concentrating on the coefficients, we focus on the fitted curve." The goal of GWR is to allow spatial data analysts to visualize the spatial variation in relationships of explanatory variables to a response variable by way of the estimated regression coefficients from each calibration location in the study area. One unresolved problem with spatially varying coefficient regression models is with correlation in the estimated coefficients, at least partly due to collinearity in the explanatory variables of each local model. Wheeler and Tiefelsdorf (2005) show that, while GWR coefficients can be correlated when there is no explanatory variable collinearity, the coefficient correlation increases systematically with increasing variable collinearity. The collinearity in explanatory variables can be exacerbated by the kernel weights. Intuitively speaking, one is using values of a variable for each local model that are similar because they are close in space, and then applying similar weights to these nearby observations, thus intensifying the similarity in these values. Moderate collinearity of locally weighted explanatory variables can lead to potentially strong dependence in the local estimated coefficients (Wheeler and Tiefelsdorf, 2005). This strong dependence in estimated coefficients makes interpretation of individual coefficients tenuous at best, and highly misleading at worst.

In the literature on GWR, there has been little focus placed on diagnostic tools for collinearity effects in models. One exception is Wheeler and Tiefelsdorf (2005), who recommend the use of bivariate scatter plots of estimated regression coefficients and local parameter correlation maps to diagnose collinearity effects on the regression coefficients. These methods highlight both the local collinearity effects on estimated coefficients and the global pattern of correlated regression coefficients across the study area. Other methods to detect explanatory variable collinearity are variance decomposition using singular-value decomposition (SVD), as described in Belsley (1991), and variance inflation factors (VIF), as outlined in Neter et al (1996). However, these have not yet been applied to GWR models systematically. Two remedial methods used to overcome collinearity effects in regression models with unvarying coefficients are ridge regression and the lasso (see Hastie et al, 2001). Both of these are shrinking methods that place a constraint on the regression coefficients. Ridge regression penalizes the size of regression coefficients and decreases the influence in the model of variables with relatively small variance. The lasso also shrinks the regression coefficients but shrinks the smaller variance variable coefficients to zero, thereby simultaneously performing model selection and coefficient penalization. However, neither of these remedial methods has been implemented in any spatially varying coefficient models, including GWR. Seifert and Gasser (2000) suggest using ridging to improve the performance of local polynomial regression models by shrinking the local polynomial estimate towards the origin when the estimation location is far from the mean. Their method strikes a compromise between the variance and bias of the local linear estimator. Note that Seifert and Gasser were working in the setting of local regression models, so their kernels are applied in variable space, and they are not concerned with the interpretation of the parameters in the local polynomial fitting. The idea of balancing variance and bias in estimators is a common thread in the statistical learning literature, and is mentioned briefly in the context of GWR by Fotheringham et al (2002, pages 62 – 63). As with local regression models, in GWR the variance of the regression coefficient estimator should generally decrease as the kernel increases, but the bias of the coefficient should increase. Conversely, the bias should decrease as the kernel size decreases, but at the expense of increased variance of the coefficient. A common approach to balance bias and variance of an estimator is to use cross-validation (CV), which minimizes a function of prediction error. CV is used both in local regression models and in GWR to estimate the kernel bandwidth. It can also be used to estimate the ridge parameter in ridge regression (Golub et al, 1979; Welsch, 2000).

The objective of this paper is to present diagnostics for collinearity in GWR and apply ridge regression in the GWR framework to decrease the effects of collinearity in explanatory variables on estimated regression coefficients. The goal is to produce regression coefficients that are more useful for interpretation by constraining the estimated coefficients. I first review the GWR framework and ridge regression in section 2 and then present the formulation for geographically weighted ridge regression (GWRR) and the diagnostic tools of VIFs and variance – decomposition proportions. Next, I demonstrate the problem of collinearity in GWR with an example dataset and report the results of applying GWRR to the same dataset in section 3. Finally, I summarize the findings and suggest future research ideas in section 4.

2 Methods

GWR models relationships between variables in data that are essentially geostatistical in nature. The variables are utilized at fixed points that have spatial coordinates and the variable values are usually mean measures of aggregate data. The spatial coordinates are used for calculating distances that are used in the kernel function. Typically, a model is fitted for each point location in the dataset. The basic GWR model for each calibration location is

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} x_{ik} + \varepsilon_i , \qquad i = 1, ..., n , \tag{1}$$

where y_i is the dependent variable at location i, β_{i0} is the intercept parameter at location i, β_{ik} is the local regression coefficient for the kth explanatory variable at location i, x_{ik} is the value of the kth explanatory variable at location i, and ε_i is the random error at location i. The regression coefficients are estimated for each calibration location independently by weighted least squares. The matrix calculation for the estimated regression coefficients is

$$\hat{\boldsymbol{\beta}}(i) = [\mathbf{X}^{\mathrm{T}}\mathbf{W}(i)\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(i)\boldsymbol{y} , \tag{2}$$

where $\mathbf{X}$ is the matrix of exogenous variables with a column of 1s for the intercept, $\mathbf{W}(i) = \mathrm{diag}[w_1(i), ..., w_n(i)]$ is the diagonal weights matrix that varies by calibration location i, $\boldsymbol{y}$ is the vector of dependent variables, $\hat{\boldsymbol{\beta}}(i) = (\hat{\beta}_{i0}, \hat{\beta}_{i1}, ..., \hat{\beta}_{ip})^{\mathrm{T}}$ is the vector of $(p+1)$ local regression coefficients at location i, and the superscript T indicates the matrix transpose. The weights matrix is calculated from a kernel function that places more weight on observations that are a smaller distance away from the calibration location i. A commonly used kernel function in the literature is the bi-square nearest neighbor function

$$w_j(i) = \begin{cases} \left[1 - \left(\dfrac{d_{ij}}{b}\right)^2\right]^2 , & \text{if } j \in \{N_i\}, \\ 0, & \text{if } j \notin \{N_i\}, \end{cases} \tag{3}$$

where d_{ij} is the distance between the calibration location i and location j, b is the distance to the Nth nearest neighbor, and the set $\{N_i\}$ contains the observations that are within the distance of the Nth nearest neighbor. The weights for observations beyond the Nth nearest neighbor distance are zero and the weight for observation i is 1. In fitting the GWR model, the kernel bandwidth is first estimated by CV using all the calibration locations; next the weights are calculated using equation (3); then the regression coefficients are estimated at each calibration location using equation (2); and finally the response estimates are made at each location by $\hat{\boldsymbol{y}}_{(i)} = \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}$. These steps are similar to the steps in fitting local linear regression models (see Hastie et al, 2001).

Shrinkage methods, such as ridge regression, place a constraint on the regression coefficients. The ridge regression coefficients minimize the residual sum of squares along with a penalty on the size of the squared coefficients as

$$\hat{\boldsymbol{\beta}}^{\mathrm{R}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{p} x_{ik}\beta_k \right)^2 + \lambda \sum_{k=1}^{p} \beta_k^2 \right\} , \tag{4}$$

where λ is the ridge regression parameter that controls the amount of shrinkage in the regression coefficients. As Hastie et al (2001) point out, an equivalent way to write the ridge regression problem that explicitly defines the constraint is

$$\hat{\boldsymbol{\beta}}^{\mathrm{R}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{p} x_{ik}\beta_k \right)^2 , \tag{5}$$

subject to

$$\sum_{k=1}^{p} \beta_k^2 \leqslant s\,,$$

where there is one-to-one correspondence between the parameters λ and s. The intercept is not constrained by the ridge parameter and the solutions are not invariant to scaling, so the input variables are typically standardized before estimating λ. Hastie et al (2001) effectively remove the intercept from ridge regression by centering the x variables and estimating β_0 by the mean of y, thereby leaving only the p variable coefficients to constrain. The ridge regression solutions are

$$\hat{\boldsymbol{\beta}}^{\mathrm{R}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\,, \tag{6}$$

where $\mathbf{I}$ is the $(p \times p)$ identity matrix. The ridge regression parameter can be estimated before estimating the ridge expression coefficients by CV or generalized CV (Golub et al, 1979; Welsch, 2000) by minimizing the squared prediction error.

To include ridge regression in GWR, it is necessary to remove or isolate the intercept term that is customarily included in these models. There are two approaches considered here to remove the intercept using centering of the variables, one using global centering and one using local entering. Using global centering, one first centers the x variables to remove the portion of the intercept when $x = 0$, leaving the global mean of y. Next, one centers the response variable to remove the global y mean and then scales the x and y variables. Then, one removes the local x and y mean deviations from the global means to get an intercept of 0 for each local model. A convenient, albeit inefficient, way to do this is to fit a GWR model to the globally centered data and then subtract the fitted intercept from the local y values. At this point, the intercept term is effectively removed from the ridge regression constraint and the penalized coefficients can be estimated. The approach allows one to compare the GWR estimates to the standardized global regression coefficients because the centering is the same, but the incremental estimation to remove the intercept results in additional bias in the ridge regression adjusted coefficients. It is advisable to scale the x and y variables by their respective standard deviations because the ridge regression solutions are scale dependent. If one does not scale the x variables, the ridge regression solution will be more influenced by variables with large variance. In other words, coefficients associated with variables of small scale should shrink more than those of larger scale.

The formula to estimate the GWRR coefficients with global centering is

$$\hat{\boldsymbol{\beta}}(i) = [\mathbf{X}^{*\mathrm{T}}\mathbf{W}(i)\mathbf{X}^{*} + \lambda\mathbf{I}]^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{W}(i)\boldsymbol{y}^{*}\,, \tag{7}$$

where $\mathbf{X}^{*}$ is the matrix of standardized explanatory variables, $\boldsymbol{y}^{*}$ is the standardized response variable, and other terms are as previously defined. Note that when the ridge parameter is 0, the estimated GWR and GWRR coefficients are the same. As with ridge regression, one first must estimate the ridge parameter λ before calculating the regression coefficients for each model. If one elects to use only a single λ for the entire study area, then there are now two global parameters to estimate before fitting the local models. Once the GWRR coefficients have been estimated, the response variable predictions are calculated after adjusting for the intercept term. To do so, the local mean deviation from the global $\bar{y}$ must be added back to $\hat{y}^{*}_{(i)} = \mathbf{X}^{*}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}$. To get the estimated y in the original units, the transformation $\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}^{*}\mathrm{std}(y) + \bar{y}$ must be performed. Bootstrapping can be used to perform inference on the regression coefficients. The bootstrap procedure to accomplish this is currently undeveloped and is left for future research.

Alternatively, one can use locally centered and globally scaled x and y values to effectively remove the local intercept. The estimation procedure is more straightforward than with global centering, but it requires centering the data for each model. The x and y variables are first globally scaled and then for each local model the x and y variables are locally centered by first calculating the weighted mean for each variable using the square root of the kernel weights $\mathbf{W}^{1/2}(i)$ and then subtracting the weighted mean from each variable. The square root of the weight is taken to correspond to the weighting of the x and y variables in equation (7). The weights $\mathbf{W}^{1/2}(i)$ are then applied to the centered values and this changes the coefficient estimation in equation (7) to

$$\hat{\boldsymbol{\beta}}(i) = (\mathbf{X}_{\mathrm{w}}^{\mathrm{T}}\mathbf{X}_{\mathrm{w}} + \lambda\mathbf{I})^{-1}\mathbf{X}_{\mathrm{w}}^{\mathrm{T}}\boldsymbol{y}_{w} , \tag{8}$$

where $\mathbf{X}_{\mathrm{w}}$ is the matrix of weighted, locally centered explanatory variables, $\boldsymbol{y}_{\mathrm{w}}$ is the vector of weighted, locally centered responses, and other terms are as previously defined. After estimating the coefficients, the response variable predictions are calculated by adding the local mean $\bar{y}_{\mathrm{w}}$ to $\hat{\boldsymbol{y}}_{\mathrm{w}}(i) = \mathbf{X}_{\mathrm{w}}(i)\hat{\boldsymbol{\beta}}(i)$. To transform the estimated y to the original units, the estimate is scaled by the standard deviation of y, in other words $\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}_{\mathrm{w}}\,\mathrm{std}(y)$. The estimated coefficients from this approach are not as directly comparable to the standardized global regression model as the global centering results due to the different variable centering, but the local centering should not produce coefficients that are largely dissimilar from the global centering. An advantage of this approach is that it introduces less bias in the coefficients than with the global centering approach. The local centering approach has the property that locally centred x variables will not have exactly equal scale, which means not all local models will have equal impact in the estimation of one global ridge regression parameter. In the analysis presented later, locally centered and globally scaled variables are primarily used for the ridge parameter estimation to reduce the estimation bias in the GWRR coefficients. It is recommended that one consider the global centering approach only if a direct comparison to the global standardized regression model and traditional GWR results is of interest.

There are numerous possible schemes for estimating the kernel and ridge parameters with CV: (1) estimate the kernel bandwidth first and then the ridge parameter; (2) estimate the kernel bandwidth, then estimate the ridge parameter, and then repeat using previous values until the parameters converge; (3) perform a search for the kernel bandwidth and perform a search for the ridge parameter at each evaluated value of the kernel bandwidth; or (4) estimate the kernel bandwidth and ridge parameter simultaneously with constrained optimization techniques. Preliminary results show that there is some interaction between the two parameters, although the kernel bandwidth dominates the squared prediction error, so scheme 1 will generally not produce optimal solutions. Scheme 2 also tends to result in a suboptimal solution because the kernel bandwidth tends to dominate the solution and it is unlikely to move to another bandwidth in the solution by changing the ridge parameter. Scheme 3 is less efficient and scheme 4 is more complicated than the first two schemes, but these schemes will generally produce better solutions. Schemes 1 and 2 will generally produce similar solutions and schemes 3 and 4 should produce similar solutions if scheme 4 treats the kernel bandwidth as an integer variable, as it is handled in scheme 3. While schemes 1 and 2 will not always find the minimum solution, they should yield good solutions that are almost as good as those from schemes 3 and 4 and will do so with less computational effort. Scheme 3 is a compromise between the computational ease of schemes 1 and 2 and the complexity of scheme 4 and is therefore viewed as the best scheme for this research. Scheme 1 has some conceptual appeal in addition to its computational ease in that it takes the best kernel bandwidth found from the standard

GWR estimation procedure and then effectively applies the ridge parameter to that solution. This makes the effect of the ridge parameter on GWR clearer, and for this reason the scheme-1 solution will at times be compared to the scheme-3 solution in the analysis that follows. A golden section search can be used for schemes 1 and 2, while a constrained optimization algorithm is appropriate for scheme 4. A nested golden section search algorithm can be used for the two components of the estimation in scheme 3. Conveniently, Matlab software, among others, provides functions for constrained optimization and the golden section search.

One problem with the golden section search algorithm is that it can terminate in a local optimum and may need to be adjusted when estimating the kernel bandwidth, as the squared prediction error function can flatten out and become stable as N increases with some datasets. This was an issue in the dataset analyzed in this paper and was addressed by running the golden section search routine in Matlab two times and truncating the bounds of the search space in the second run using the solution from the first run as the upper bound. A more conservative approach is to evaluate all possible values of the kernel bandwidth in the cross-validation and then select the best bandwidth by inspection. This, however, may be computationally expensive for large datasets. There are other possible methods to estimate the kernel bandwidth. Fotheringham et al (2002) describe a generalized cross-validation criterion for GWR that is adapted from local linear regression and also define an Akaike information criterion for the GWR framework. Páez et al (2002) present an alternative GWR model that can estimate local kernel bandwidths at each model calibration location by using maximum likelihood estimation and calculating the spatial weights as part of a model for variance. More attention is needed to determine the most appropriate kernel bandwidth estimation method. It is worth mentioning that the type of CV used here is leave-one-out because for local regression models it is not readily justifiable to remove more than one observation for each local model. Removing anything but the ith observation for the prediction at observation i seems arbitrary.

There is only a modest increase in computational complexity to include the ridge regression parameter in GWR. The main computational burden in the GWR version implemented here is the CV estimation of the kernel bandwidth. The number of calculations in the CV estimation is dominated by the calculation of the kernel weights and matrix inverse for the regression coefficients at each location, not the number of iterations of the golden section search routine. An estimate of the total time required for the CV estimation of the bandwidth in GWR is $\mathrm{O}[\ln n(n^2+np^3)]$, where the number of iterations of the search routine is of the order of $\ln n$ and there are n calculations of the kernel weights and matrix inverse taking $(n+p^3)$ calculations. Under estimation scheme 3, the CV estimation of λ in GWRR is nested within the CV estimation of N and this transfers the $\mathrm{O}(n^2+np^3)$ time from the N estimation to the λ estimation. The only additional computation is with the number of golden section search iterations needed to find λ at each value of N. GWRR model calibrations for four different sized datasets show that the number of search iterations needed for λ is approximately the same as for N. Therefore, including the ridge parameter in GWR effectively doubles the number of iterations needed in the search routine to estimate the parameters and the computational complexity of the CV estimation in GWRR is $\mathrm{O}[(\ln n)^2(n^2+np^3)]$. GWR and GWRR are both polynomial-time algorithms.

There are diagnostic tools one can use to evaluate whether GWRR is worthwhile using for a certain dataset. In addition to scatter plots of local regression coefficients and maps of local regression coefficient correlations (see Wheeler and Tiefelsdorf, 2005), one can use VIFs and variance–decomposition proportions with weighted design matrix condition numbers. Of course, one can always fit the GWRR model and if the estimated ridge parameter is zero then constraining the regression coefficients does not lower the prediction error and the GWR model results. When using GWR models, it is possible to calculate VIF values for each explanatory variable in each local model. The VIF for a variable at location *i* is

$$\mathrm{VIF}_k(i) = \frac{1}{1 - R_k^2(i)}, \tag{9}$$

where $R_k^2(i)$ is the coefficient of determination when x_k is regressed on the other explanatory variables at model calibration location *i*. The kernel size for these models is the same as in the GWR model to ensure we are diagnosing collinearity at the scale of the GWR model. For models with more than two explanatory variables, a weighted local regression of each variable on all the other variables would give the $R_k^2(i)$ needed to calculate the VIFs for the *p* variables. Naturally, a more efficient method to calculate the VIFs in this situation would be computationally beneficial. In a model with only two variables, the VIF is the same for both variables and is straightforward to calculate using the weighted correlation coefficient between the variables [see Fotheringham et al (2002) for a discussion of weighted moment-based statistics]. The geographically weighted correlation coefficient for two variables is

$$r_{k,l}(i) = \frac{\displaystyle\sum_{j=1}^{n} w_{ij}^*(x_{kj} - \bar{x}_{ki})(x_{lj} - \bar{x}z_{li})}{\left[\displaystyle\sum_{j=1}^{n} w_{ij}^*(x_{kj} - \bar{x}_{ki}) \sum_{j=1}^{n} w_{ij}^*(x_{lj} - \bar{x}_{li})\right]^{1/2}}, \tag{10}$$

where

$$\bar{x}_{li} = \sum_{j=1}^{n} w_{ij}^* x_{lj} \tag{11}$$

is the weighted mean for explanatory variable *l* at location *i* (similarly for variable *k*) and w_{ij}^* is the standardized square root of the kernel weight between locations *i* and *j*, where the weights are standardized to sum to one. Setting $R_k^2(i) = r_{k,l}^2(i)$ allows calculation of the VIF using equation (9).

Two drawbacks of the VIF as a collinearity diagnostic are that it does not consider collinearity with the constant term and does not illuminate the nature of the collinearity. Belsley (1991) suggests another diagnostic tool for collinearity that uses SVD of the design matrix **X** to form condition indexes of this matrix and variance – decomposition proportions of the coefficient covariance matrix. The SVD of the design matrix in the GWR framework is

$$\mathbf{W}^{1/2}(i)\mathbf{X} = \mathbf{UDV}^{\mathrm{T}}, \tag{12}$$

where **U** and **V** are orthogonal $n \times (p+1)$ and $(p+1) \times (p+1)$ matrices, respectively, **D** is a $(p+1) \times (p+1)$ diagonal matrix of singular values of decreasing value down the diagonal starting at position (1, 1), **X** is the column scaled matrix (by its norm) of explanatory variables including the constant, and $\mathbf{W}^{1/2}(i)$ is the square root of the diagonal weight matrix for calibration location *i* calculated from the kernel function. Using the SVD, the variance – covariance matrix of the regression coefficients is

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{VD}^{-2}\mathbf{V}^{\mathrm{T}}, \tag{13}$$

and the variance of the *k*th regression coefficient is

$$\mathrm{var}(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^{p} \frac{v_{kj}^2}{d_j^2}, \tag{14}$$

where the d_j are the singular values and the v_{kj} are elements of the $\mathbf{V}$ matrix. The variance–decomposition proportion is the proportion of the variance of the kth regression coefficient affiliated with the jth component of its decomposition. Following from Belsley (1991), the variance–decomposition proportions are

$$\pi_{jk} = \frac{\phi_{kj}}{\phi_k}, \tag{15}$$

where

$$\phi_{kj} = \frac{v_{kj}^2}{d_j^2}, \tag{16}$$

$$\phi_k = \sum_{j=1}^{p} \phi_{kj} . \tag{17}$$

The condition index for column $j = 1, ..., p+1$ of $\mathbf{W}^{1/2}(i)\mathbf{X}$ is

$$\eta_j = \frac{d_{\max}}{d_j}, \tag{18}$$

where d_j is the jth singular value of $\mathbf{W}^{1/2}(i)\mathbf{X}$.

Belsley recommends using condition indexes greater than or equal to 30 for a column-scaled $\mathbf{X}$ and variance proportions greater than 0.5 for two or more coefficients for each variance component as an indication of collinearity in a regression model. The larger the condition number, the stronger is the collinearity among the columns of $\mathbf{X}$. The critical value of 30 is a general and conservative guideline from Belsley's experimentation and different values may be more appropriate for certain datasets. The presence of more than two variance proportions greater than 0.5 in one component of the variance–decomposition indicates that collinearity exists between more than two regression terms, which could include the constant. Belsley also recommends including the intercept in $\mathbf{X}$ and using uncentered explanatory variables in the SVD so as not to disguise any collinearity with the intercept.

3 Results

The dataset for the analysis in this paper is the Columbus crime dataset analyzed previously by Anselin (1988). This dataset contains crime rates in forty-nine planning neighbourhoods, closely related to census tracts, in Columbus, OH. The data consist of variables for mean housing value, household income, x and y spatial coordinates of neighborhood centroids, and residential and vehicle thefts combined per thousand people for 1980. Figure 1 is a map of the study area with observation identifiers as labels. The regression models in this paper are limited to only two explanatory variables for ease of exposition and clarity of the methods and results. Of course, nothing in the method limits models to only two variables. In fact, the method proposed here should be more useful as the number of potentially collinear explanatory variables in the regression model increases. The dependent variable here is residential and vehicle thefts per one thousand people and is referred to as crime rate. The two explanatory variables in the model are mean housing value and mean household income, and they exhibit moderate positive correlation ($r = 0.50$) with one another and have clear intuitive and statistical negative relationships with crime rate ($r = -0.57$ and $r = -0.70$, respectively). One would naturally expect a negative relationship between income and crime and also between housing value and crime, as more affluent neighborhoods generally have lower crime rates.

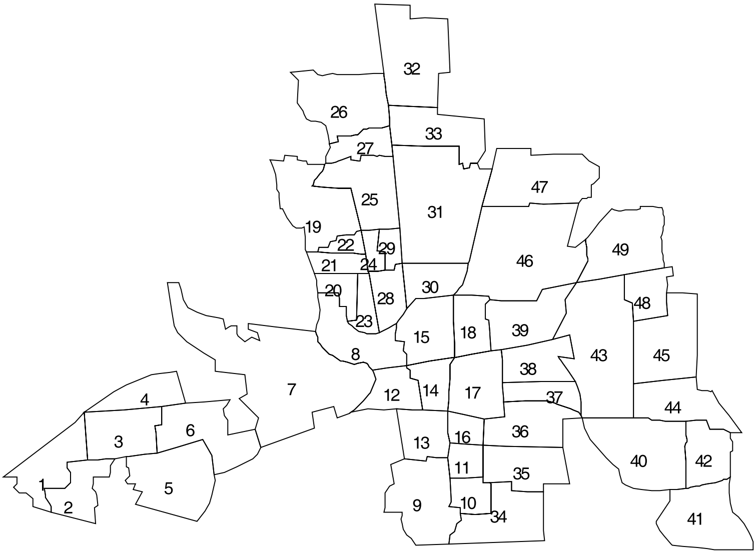

Figure 1. Columbus, OH, 1980 crime rate neighborhood areas with identifiers.

The global regression model estimates for the unstandardized and standardized variables are listed in table 1. The results show that collinearity is not a problem in the global model, as indicated by the low VIF value (1.33) for the two variables, parameter correlation of −0.50 and the intuitive negative signs for the variable coefficients. The GWR model mean coefficient estimates and overall fit are listed in table 2. One of the features of GWR models is typically a large increase in R^2 over the global regression model, and this is the case in tables 1 and 2, as the R^2 increases from 0.55 to 0.92. Another noticeable difference in the two tables is the large increase in VIF values, from 1.33 to a mean of 3.04 with GWR. Figure 2 displays the distribution of VIFs for the GWR model and shows that there are three local models with large VIFs over the conservative threshold value of 10, where VIFs greater than 10 correspond to variable correlation that considerably exceeds 0.90. These local models are at observations 37, 38, and 39, with observation 38 as the spatially connecting neighbor between 37 and 39. The local regression coefficient correlations for income and housing value at observations 37, 38, and 39 are −0.97, −0.99, and −0.98, respectively.

The mean GWR coefficient estimates in table 2 do not convey the amount of variation in the estimates and a scatter plot is beneficial to show this. Figure 3 shows the scatter of GWR estimated coefficients using local entering for housing value versus income, along with observation identifiers, and clearly demonstrates a strong linear

Table 1. Global regression model summary for unstandardized and standardized variables.

Parameter	Unstandardized					Standardized	
	estimate	standard error	p-value	VIF[a]	coefficient correlation	estimate	standard error
Intercept	68.619	4.735	0.000			0.000	0.000
Income	−1.597	0.334	0.000	1.333	−0.500	−0.544	0.114
House valve	−0.274	0.103	0.011	1.333	−0.500	−0.302	0.114
R^2	0.55						

[a] VIF—variance inflation factor.

Table 2. GWR (geographically weighted regression) model summary for unstandardized and standardized variables.

Parameter	Unstandardized				Standardized
	mean estimate	mean VIF[a]	mean parameter correlation	global parameter correlation	mean estimate
Intercept	62.670				0.000
Income	−1.398	3.040	−0.582	−0.796	−0.477
House value	−0.153	3.040	−0.582	−0.796	−0.169
R^2	0.92				

[a] VIF—variance inflation factor.

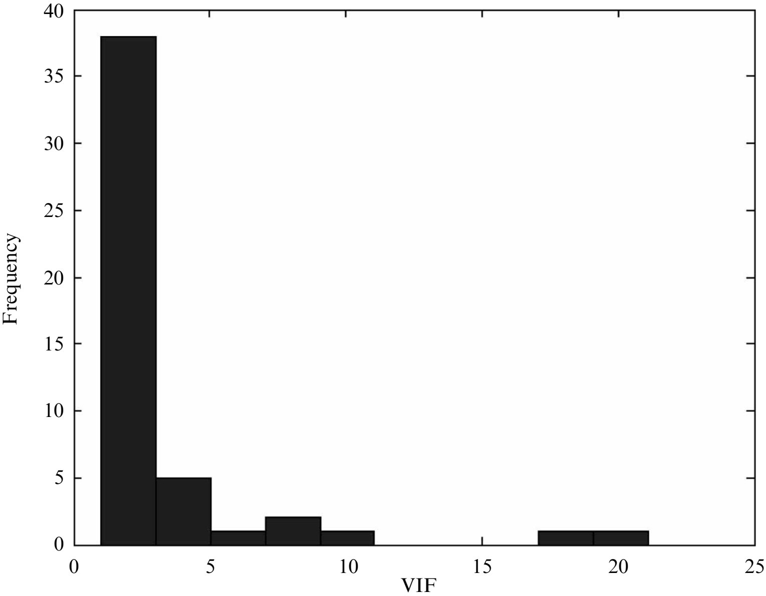

Figure 2. Distribution of GWR (geographically weighted regression), variance inflation factor (VIF) values in a regression model with two explanatory variables.

association between the coefficients across the study area. The correlation coefficient of the two sets of coefficients is −0.80. It appears from this figure that the local coefficient correlation is in a sense spilling over across the local models to show a trend of collinearity across the study area, as the global pattern is composed from the individual local model calibrations. In figure 3 it is clear that there are numerous estimated coefficients that are positive in sign, in contrast to the global model and intuition. This phenomenon is more pronounced for housing value (β_2) than for income (β_1). The three observations (37, 38, 39) with the largest VIFs are also the three observations in figure 3 that have the largest housing value regression coefficients and smallest income regression coefficients. The collinearity in the data at these model locations is increasing the variance in the estimated regression coefficients.

The variance–decomposition proportions and condition indexes indicate collinearity trouble with numerous local models. Table 3 shows the condition index and variance–decomposition proportions for the largest variance component for all observations with a

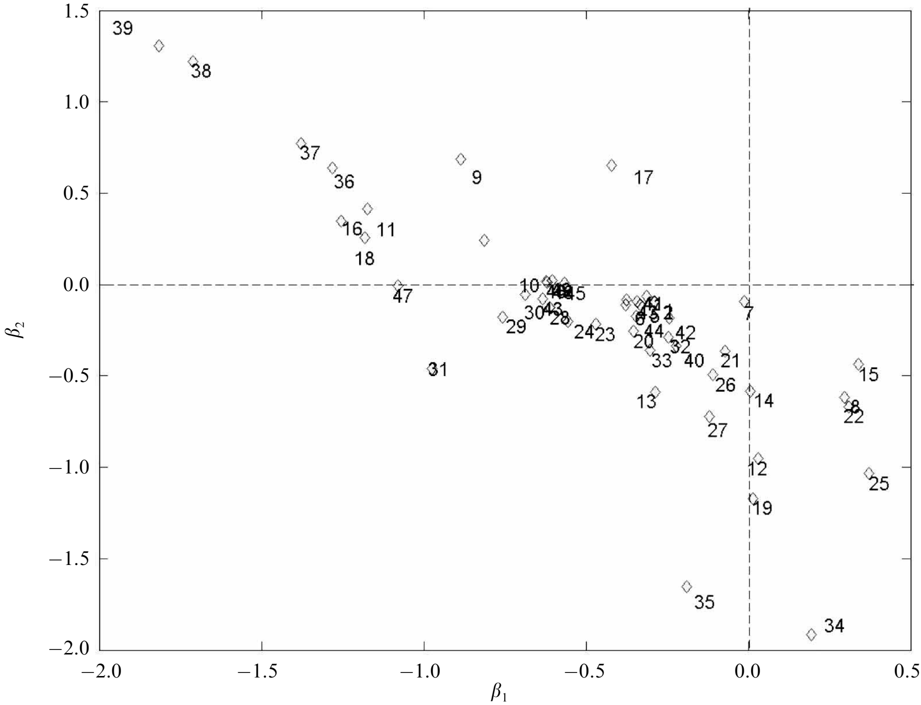

Figure 3. GWR (geographically weighted regression) estimated coefficients for housing value (β_2) versus income (β_1) with observation identifiers (see figure 1).

Table 3. Condition indexes, variance – decomposition proportions, and variance inflation factors (VIFs) for observations with a large condition index.

Identifier[a]	η_j	π_{j1}	π_{j2}	π_{j3}	VIF
10	28.804	0.065	0.909	0.966	2.827
34	46.472	0.388	0.948	0.992	5.602
35	41.748	0.493	0.930	0.994	4.525
39	25.999	0.053	0.988	0.981	17.815
40	25.358	0.100	0.981	0.955	8.306
41	31.024	0.038	0.975	0.956	8.382

[a] See figure 1.

condition index greater than 25, along with the VIFs. There are three observations (34, 35, 41) with a condition index over 30 and three observations (10, 39, 40) with a condition index between 25 and 30. Five of these six observations have a VIF well over 3, with observation 10 having a VIF just below 3. In all six of the observations, the variance proportions greater than 0.90 for the second and third variance proportion columns indicate that the collinearity is between the two explanatory variables (the intercept is the first variance proportion column). Observations 34 and 35 are among the observations with the smallest estimated regression coefficients for housing value and largest estimated regression coefficients for income in figure 3. Using the conservative VIF and variance – decomposition criteria outlined above, there are indications of collinearity in at least eight of the 49 observations in the study area.

These results for the GWR model suggest that collinearity is a problem with these data in the local regression models, even though it is not a problem in the global regression model. Based on previous experience, it appears to be a general result that lack of collinearity in the global regression model will be a poor indicator for absence of collinearity in GWR models. Therefore, ridge regression can help in the local regression case to control the variance of the regression coefficients when it is not needed in the global model. Based on the collinearity diagnostics, it appears worthwhile to apply this remedial method to the Columbus crime data in an attempt to correct the collinearity of the GWR model.

To fit the GWRR model using scheme 1, I first estimate the kernel bandwidth and then ridge parameter. The estimated kernel bandwidth $N = 11$ and the estimated ridge parameter $\lambda = 0.80$ with estimation scheme 1. The scheme-3 solution is the same as the scheme-1 solution in this case. The prediction error, as measured by the root mean squared prediction error (RMSPE), and the estimation error, as measured by the root mean squared error (RMSE), are plotted for the GWRR model in figure 4, along with the RMSPE for the GWR model, as functions of the kernel bandwidth. The prediction error is the error using CV and the estimation error uses all the observations in the model calibration. The estimation error is a global measure in that it uses the squared deviation of only the $\hat{y}(i)$ from $y(i)$ in each model at location i. The deviations from the $y(j)$ in each model at location i are not utilized. Figure 4 demonstrates that the estimation error is lower than the prediction error for GWRR, and this is also true for GWR. The higher prediction error is due to removing the maximally weighted observation, i, in the cross-validation for observation i, for this observation always has a weight of 1. When this observation is added back into the data used for estimating the coefficients in the model at location i, the fit naturally improves. Note that the behavior of the prediction error as N increases is stable in figure 4. This behaviour can be problematic for search routines and care is needed in the estimation of N to make sure a local solution with a larger N than is necessary is not selected. This is important

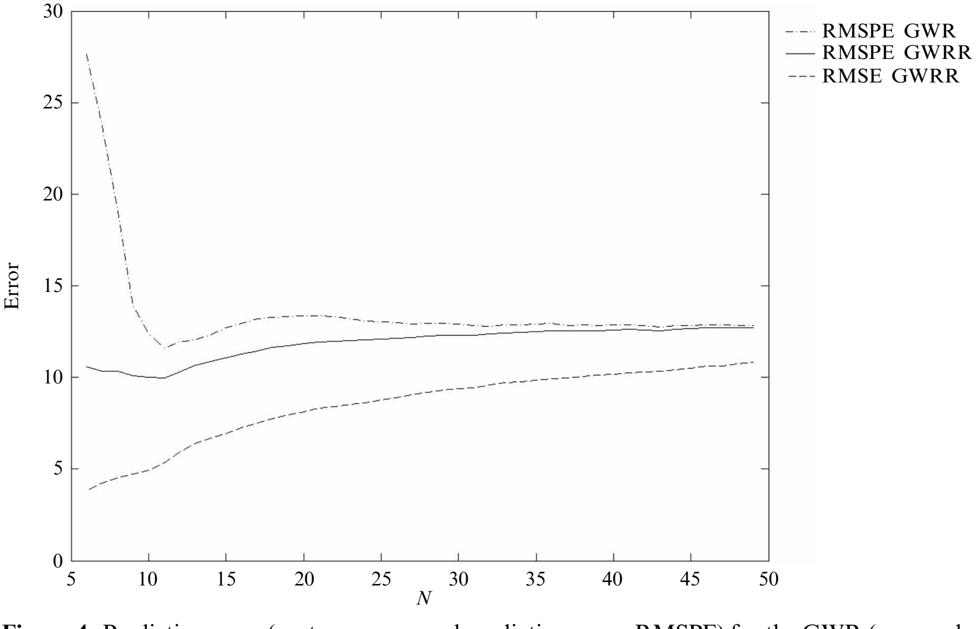

Figure 4. Prediction error (root mean squared prediction error, RMSPE) for the GWR (geographically weighted regression) ($\lambda = 0$) and GWRR (geographically weighted ridge regression) ($\lambda = 0.80$) solutions and the estimation error (root mean squared error, RMSE) for the GWRR solution as a function of N.

because initial results show that the overall GWR model fit decreases with increasing N. Note that in general there will be a perfect fit if $N \leqslant p$ because there are fewer observations than number of coefficients, and, for the bi-square nearest neighbor kernel function, the fit is perfect for $N \leqslant p+1$ because the weight of the Nth observation is 0. Figure 4 also shows that the GWRR solution has a lower prediction error than the GWR solution. This is congruous with the idea in the statistical learning literature that prediction error can usually be improved in global models by introducing some bias to reduce variability in the predictions.

Figure 5 shows the prediction error for a truncated range of kernel bandwidth values and four ridge parameter values. The value of $\lambda = 0.8$ corresponds to the GWRR solution. A value of 1.4 for λ serves as an upper bound reference, as no value larger than this for λ produces an optimal solution. A value of 0.0 for λ corresponds to the GWR model. The interaction of N and λ in the prediction error is apparent in the figure for $\lambda = 0.8$ and $\lambda = 1.4$. This figure shows that various values of λ improve on the GWR prediction error and that the best λ depends on N. In general, this should be the case. This evidence provides an argument for generally using scheme 3 to estimate N and λ simultaneously. However, as the scheme 1 and scheme 3 solutions are the same in this case, one might argue it is not worth the additional computational cost.

The GWRR model estimates are listed in table 4. The results show a decrease in the mean local coefficient correlation and the global coefficient correlation from the GWR model. The overall correlation coefficient between the two sets of estimated regression coefficients decreases from −0.80 to −0.53 and the mean local coefficient correlation decreases from −0.58 to −0.01. The overall model fit as measured by R^2 decreases only mildly from 0.92 to 0.90. Table 4 also shows that for the GWRR solution, the mean coefficient estimate for housing value becomes more negative and the

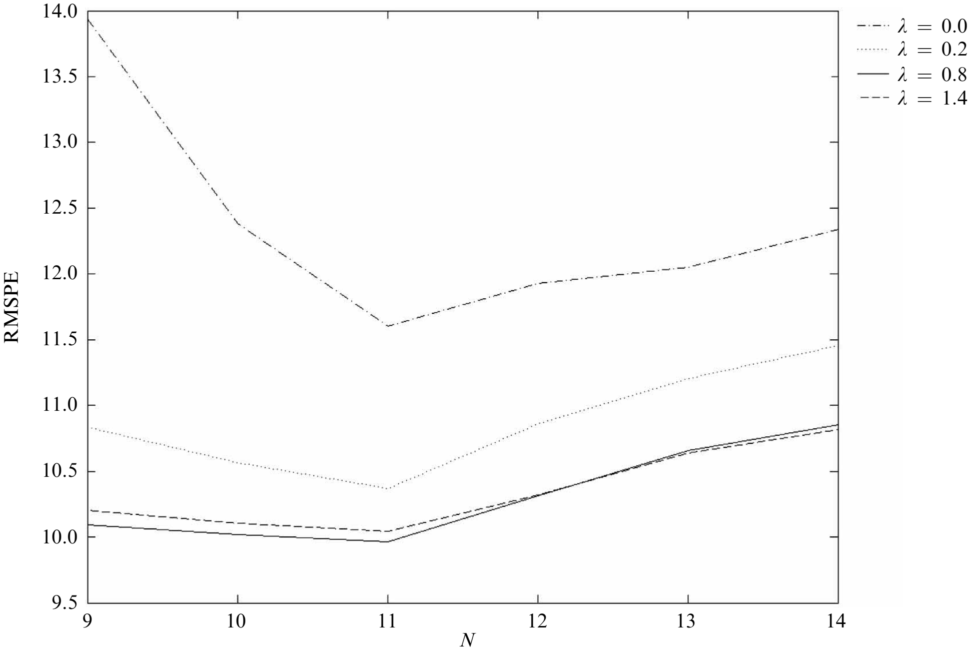

Figure 5. Prediction error (RMSPE, root mean squared prediction error) as a function of N and λ for a truncated range of N and selected values of λ. $\lambda = 0$ is the GWR (geographically weighted regression) solution and $\lambda = 0.8$ is the GWRR (geographically weighted ridge regression) solution. Two other λ values (0.2 and 1.4) illustrate the function behavior.

Table 4. GWRR (geographically weighted ridge regression) model summary.

Parameter	Unstandardized				Standardized
	mean estimate	mean VIF[a]	mean parameter correlation	global parameter correlation	mean estimate
Intercept	55.465				0.000
Income	−0.745		−0.012	−0.530	−0.254
House value	−0.186		−0.012	−0.530	−0.205
R^2	0.90				

[a] VIF—variance inflation factor.

mean coefficient estimate for income becomes less negative for income than with the GWR model. The Moran's *I* statistic is 0.054 for the GWR residuals and is 0.026 for the GWRR residuals. These statistics are not significant and indicate that there is no significant spatial autocorrelation in the GWR residuals in this case, and that the inclusion of the coefficient penalization does not significantly affect the spatial autocorrelation level in the model residuals.

The regression coefficients for the GWRR models using local centering and global centering under scheme-1 estimation are plotted in figure 6. The pattern of coefficients is similar for the two solutions, although there is overall more coefficient shrinkage in the local centering solution with a smaller ridge parameter value. The observations with the large VIFs and estimated GWR coefficients in the upper left corner of figure 3 have been penalized in the GWRR solutions to now reside in the main grouping of

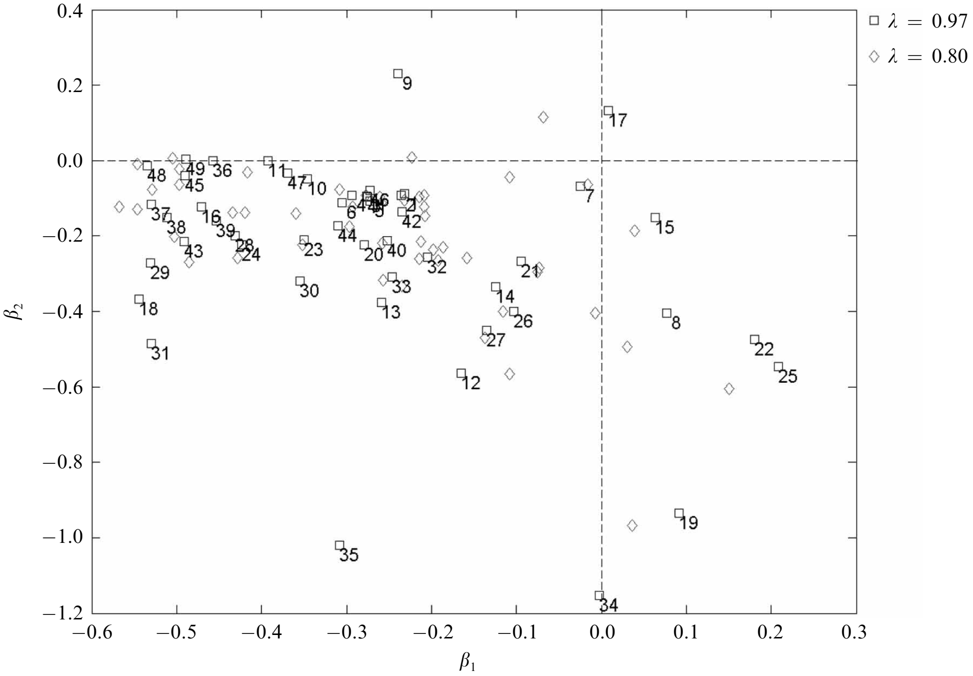

Figure 6. Estimated regression coefficients for the GWRR (geographically weighted ridge regression) local centered (λ = 0.80) and global centered (λ = 0.97) solutions with observation identifiers (see figure 1).

observations that have intuitively signed coefficients. The regression coefficients for the GWR model and the GWRR local centering model are plotted in figure 7. The effect of the ridge parameter on the estimated coefficients is more clear in this figure. The coefficients have been reduced away from the positive values they had in the GWR model, especially for the housing value variable (β_2). The coefficients now have more intuitive signs considering the response variable of crime. The regression coefficients for the locally centered GWR model are mapped in figure 8 and the corresponding GWRR coefficients are mapped in figure 9. The dependence in the GWR regression coefficients in the form of negative association is clear in figure 8. The areas with counterintuitive positive regression coefficients for income are not the same areas with counterintuitive positive regression coefficients for housing value. The two maps in figure 9 show less of a complementary pattern than those in figure 8, with fewer

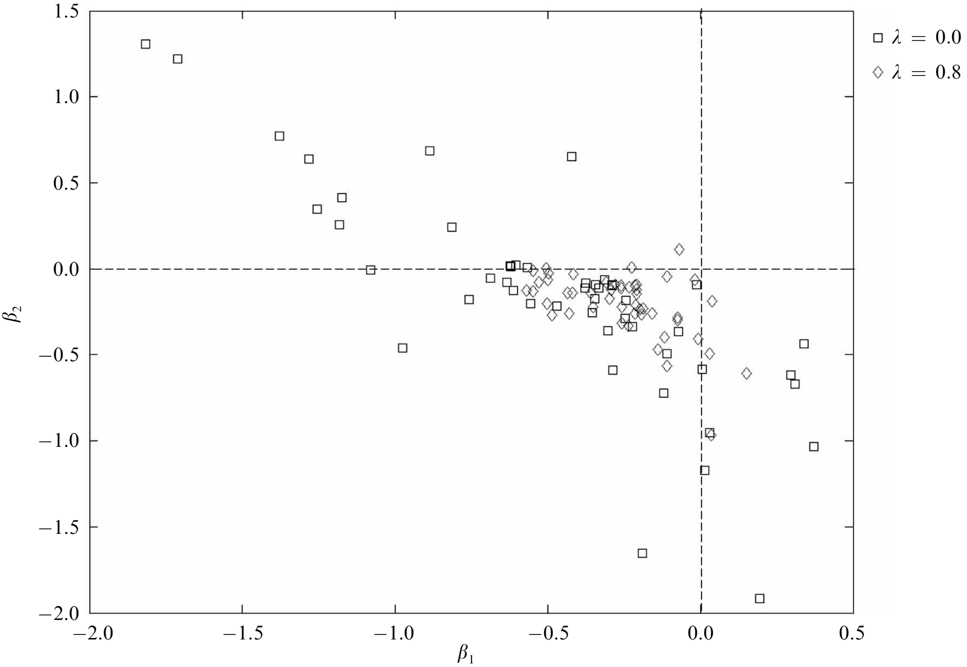

Figure 7. GWR (geographically weighted regression) ($\lambda = 0.0$) and GWRR (geographically weighted ridge regression) ($\lambda = 0.8$) estimated regression coefficients using local centering.

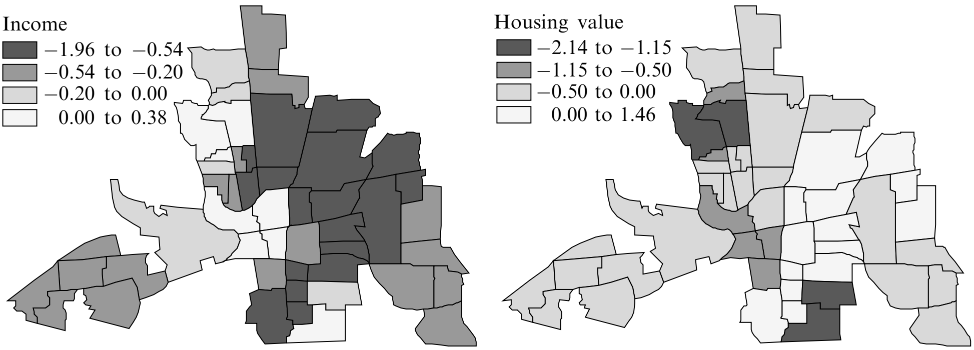

Figure 8. Estimated regression coefficients for the GWR (geographically weighted regression) model.

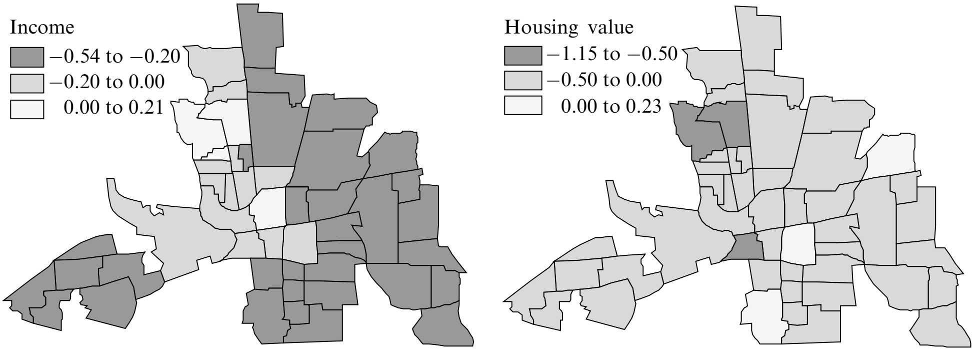

Figure 9. Estimated regression coefficients for the GWRR (geographically weighted ridge regression) model.

areas that have light shaded values for the housing value parameter when the income parameter is dark-shaded (most negative), and vice versa. The strong negative association in the GWR coefficients in the east-central portion of the study area in figure 8 has been especially reduced in the GWRR coefficients in figure 9. Taken together, the results in this section show that it is fruitful to diagnose and adjust for collinearity in GWR at the local model level.

A preliminary experiment to evaluate how the GWRR model responds to increasing collinearity in the explanatory variables shows the model to be quite robust to extremely collinear variables. The experiment involved increasing the level of collinearity in the Columbus crime model from the original level by replacing the standardized housing value variable in the model by a weighted combination of the standardized income and housing value variables. The new variable, x_2', is calculated from the standardized variables x_1^* and x_2^* as

$$x_2' = a x_1^* + (1 - a) x_2^*, \qquad (19)$$

Table 5. Mean local regression coefficient correlation and global regression coefficient correlation in the GWRR (geographically weighted ridge regression) and GWR (geographically weighted regression) models at various levels of correlation in the explanatory variables. The weight determines the amount of variable correlation and $\lambda = 0$ corresponds to the GWR model.

Weight	Variable correlation	λ	Mean coefficient correlation	Global coefficient correlation	Mean $\hat{\beta}_1$	Mean $\hat{\beta}_2$
0.00	0.50	0.00	−0.58	−0.80	−0.48	−0.17
		0.80	−0.01	−0.53	−0.25	−0.21
0.40	0.63	0.00	−0.68	−0.86	−0.44	−0.19
		0.84	−0.02	−0.55	−0.23	−0.22
0.60	0.74	0.00	−0.76	−0.91	−0.40	−0.22
		0.91	0.10	−0.57	−0.20	−0.24
0.80	0.89	0.00	−0.88	−0.97	−0.27	−0.33
		1.16	0.35	−0.48	−0.16	−0.24
0.90	0.97	0.00	−0.96	−0.99	0.01	−0.57
		1.51	0.68	0.05	−0.15	−0.21
0.95	0.99	0.00	−0.99	−1.00	0.51	−1.08
		1.77	0.89	0.72	−0.15	−0.18

where a is a weighting scalar between 0 and 1 that controls the amount of correlation in the standardization explanatory variables. Table 5 contains the summary results of the experiment. The correlation in the variables ranges from 0.50 to 0.99, and the results correspond to the GWR model when $\lambda = 0$ and correspond to the GWRR model for nonzero λ. The kernel bandwidth is fixed at $N = 11$ for all values of λ to eliminate a source of variation in the experiment even though the optimal N would likely change with λ at different levels of variable correlation. The table shows that, for strongly collinear variables, the GWR model behaves in such a way that it pushes the coefficients apart from one another while the GWRR model properly reflects the relationship of the variables in that the mean coefficients are still negative and are about equal, which reflects the association structure between the explanatory variables and the response variable. Naturally, λ increases as the variable correlation increases to reduce the coefficient variances. While the correlation levels at the bottom of the table are admittedly extreme, they are helpful in revealing the behavior of the two methods and showing the benefit of using GWRR with collinear variables, even when only two explanatory variables are included in the model.

4 Conclusions

While geographically weighted regression models offer the potential of increased understanding of the nature of varying relationships between variables across space, collinearity in the weighted explanatory variables produces statistical artifacts in the local regression coefficients that distort and potentially invalidate conclusions about the relationships based on the estimated coefficients. In a paper that is related in its general theme but not its topic, Gelman and Price (1999) show that different statistical artifacts are present in estimated disease rates across a study area based on the statistical method used. The paper presented here makes a contribution to the geography literature because it is the first work to both document the issue of collinearity in geographically weighted regression models using the diagnostic tools of variance inflation factors and variance–decomposition and also suggest a viable alternative while retaining the GWR framework. The results presented here show that it is possible to use geographically weighted regression models with a ridge regression parameter to reduce the effect of local variable collinearity on the model while producing more intuitively signed coefficients and surrendering only a modest amount of overall model fit. In addition, the GWRR model has lower prediction error through the stabilized variance of the parameters. While the method presented here may not address all potential statistical artifacts inherent in GWR, it is a viable tool for spatial data analysts who wish to investigate spatially varying relationships between variables in a regression setting and consider at the same time certain model complications arising from collinearity. There is natural appeal in explaining spatial variation in relationships through estimated regression coefficients, and this work should contribute to making that a more reliable exercise. However, there remain research opportunities to distinguish clearly spatial variation in the effect of an explanatory variable from any latent spatial autocorrelation in the pattern of GWR coefficients.

Even though the results presented here for the Columbus crime dataset are encouraging, more experimentation is needed to verify that they generalize for larger models and larger datasets. The GWRR model as presented here, with one ridge parameter, may not be as corrective for larger study areas and datasets. More research is possible to study the benefit of adding multiple ridge parameters. In the relatively small model analyzed here, one ridge parameter was adequate to substantially correct the collinearity present in the model. More ridge parameters may be needed in other datasets with varying levels of collinearity. As many as one ridge parameter for each local model could be added, or a ridge parameter could be added for a group of observations, where the observation groups could be determined endogenously. Generalized cross-validation and an Akaike information criterion could prove useful in estimating multiple ridge parameters, as cross-validation could be computationally demanding for large datasets. Preliminary results from the Columbus data indicate that individual model ridge parameters are not required and would lead to overfitting and unnecessary complexity, but more comprehensive research is required for a general conclusion. Clearly, formal statistical tests would be beneficial to determine the number of ridge parameters to include in the regression model.

Acknowledgements. I would like to thank Dr Prem Goel for his useful suggestions during discussions of this paper. I would also like to thank three anonymous reviewers and Dr Eric Sheppard for their helpful comments on an earlier draft of this paper. The author alone is responsible for any errors or omissions in this work.

References

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

Belsley D A, 1991 *Conditioning Diagnostics: Collinearity and Weak Data in Regression* (John Wiley, New York)

Fotheringham A S, Brunsdon C, Charlton M, 2002 *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* (John Wiley, Chichester, Sussex)

Gelman A, Price P N, 1999, "All maps of parameter estimates are misleading" *Statistics in Medicine* **18** 3221 – 3234

Golub G H, Heath M, Wahba G, 1979, "Generalized cross-validation as a method for choosing a good ridge parameter" *Technometrics* **21**(2) 215 – 223

Hastie T, Tibshriani R, Friedman J, 2001 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York)

Loader C, 1999 *Local Regression and Likelihood* (Springer, New York)

Neter J, Kutner M H, Nachtsheim C J, Wasserman W, 1996 *Applied Linear Regression Models* (Irwin, Chicago, IL)

Páez A, Uchida T, Miyamoto K, 2002, "A general framework for estimation and inference of geographically weighted regression models: 1. Location-specific kernel bandwidths and a test for locational heterogeneity" *Environment and Planning A* **34** 733 – 754

Seifert B, Gasser T, 2000, "Data adaptive ridging in local polynomial regression" *Journal of Computational and Graphical Statistics* **9** 338 – 360

Welsch R, 2000, "Is cross-validation the best approach for principal component and ridge regression?" *Computing Science and Statistics* **32** 356 – 361

Wheeler D, Tiefelsdorf M, 2005, "Multicollinearity and correlation among local regression coefficients in geographically weighted regression" *Journal of Geographical Systems* **7** 161 – 187