Environment and Planning A 2009, volume 41, pages 722 – 742

doi:10.1068/a40256

Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso

David C Wheeler
Department of Biostatistics, 1518 Clifton Road, NE Third Floor, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA; e-mail: dcwheel@sph.emory.edu
Received 2 October 2007; in revised form 13 December 2007; published online 3 December 2008

Abstract. In the field of spatial analysis, the interest of some researchers in modeling relationships between variables locally has led to the development of regression models with spatially varying coefficients. One such model that has been widely applied is geographically weighted regression (GWR). In the application of GWR, marginal inference on the spatial pattern of regression coefficients is often of interest, as is, less typically, prediction and estimation of the response variable. Empirical research and simulation studies have demonstrated that local correlation in explanatory variables can lead to estimated regression coefficients in GWR that are strongly correlated and, hence, problematic for inference on relationships between variables. The author introduces a penalized form of GWR, called the 'geographically weighted lasso' (GWL) which adds a constraint on the magnitude of the estimated regression coefficients to limit the effects of explanatory-variable correlation. The GWL also performs local model selection by potentially shrinking some of the estimated regression coefficients to zero in some locations of the study area. Two versions of the GWL are introduced: one designed to improve prediction of the response variable, and one more oriented toward constraining regression coefficients for inference. The results of applying the GWL to simulated and real datasets show that this method stabilizes regression coefficients in the presence of collinearity and produces lower prediction and estimation error of the response variable than does GWR and another constrained version of GWR—geographically weighted ridge regression.

1 Introduction

In the field of spatial analysis the interest of some researchers in modeling relationships between variables locally has led to the development of regression models with spatially varying coefficients. This is evidenced by the spatial expansion method (Casetti, 1992), geographically weighted regression (GWR) designed to model spatial parametric nonstationarity (Brunsdon et al, 1996; Fotheringham et al, 2002), and GWR designed to model variance heterogeneity (Páez et al, 2002). Of these, GWR as a model for spatial parametric nonstationarity has experienced the widest application to date, due at least in part to the ready availability of software for this technique. One can see the similarities of GWR with nonparametric local, or locally weighted, regression models which were first developed in the field of statistics (Cleveland, 1979; see also Hastie et al, 2001; Loader, 1999, for more details). A clear methodological link between local regression and GWR is found in the similarity of the estimation procedures for loess smoothing, which is synonymous with local regression (Martinez and Martinez, 2002, pages 292 – 293) and the GWR model of Fotheringham et al (2002), which suggests GWR be viewed as a local smoothing method. A key difference between GWR and locally weighted regression is that in GWR weights arise from a spatial kernel function applied to observations in a series of related local weighted regression models across the study area, whereas the weights in locally weighted regression are from a kernel function applied in variable space. Historically, GWR is based on the replacement of attribute space in locally weighted regression for curve fitting with geographical space in locally weighted regression for modeling potentially spatially

varying relationships. GWR also differs from local regression in the focus of its typical applications. Most published applications of GWR are concerned with measuring statistically significant variation in estimated regression coefficients, and then visualizing and interpreting the varying regression coefficients—in line with the primary proposed benefit of GWR (Fotheringham et al, 2002). In contrast, local regression is concerned with fitting a curve to the response variable (Loader, 1999, page 19). This difference in objectives may be summarized as one of inference on relationships in GWR and estimation and prediction of the response variable in local regression. The discrepancy between the principal applied focus of GWR and its methodological origins appears to be a noteworthy one, and perhaps a seemingly more appropriate use of GWR, in line with its theoretical statistical origins, is in the estimation and prediction of the response variable.

One issue of concern with GWR models expressed in the literature concerns correlation in the estimated coefficients, due at least in part 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 collinearity. The collinearity in explanatory variables can apparently be increased by the GWR spatial kernel weights, and moderate collinearity of locally weighted explanatory variables can lead to potentially strong dependence in the local estimated coefficients (Wheeler and Tiefelsdorf, 2005), which makes interpreting individual coefficients problematic. As an additional example, Wheeler (2007) applies collinearity diagnostic tools in a Columbus, Ohio, crime dataset to clearly link local collinearity to strong GWR coefficient correlation and increased coefficient variability for two covariates at numerous data locations with counterintuitive regression coefficient signs.

Another issue in GWR is with the customary standard-error calculations associated with regression-coefficient estimates. The standard-error calculations in GWR are only approximate because of the reuse of the data for estimation at multiple locations (Congdon, 2003; LeSage, 2004) and because the data are used to estimate both the kernel bandwidth and the regression coefficient (Wheeler and Calder, 2007). In addition, local collinearity can increase variances of estimated regression coefficients in the general regression setting (Neter et al, 1996). The issue with the standard errors implies that the confidence intervals for estimated GWR coefficients are only approximate, and are not entirely reliable for local model selection via significance tests. An issue related to inference on the regression coefficients is that of multiple testing in GWR, where tests of coefficient significance are carried out at many locations by use of the same data.

There are methods in the statistical literature which attempt to circumvent collinearity in traditional linear regression models with constant coefficients. These methods include ridge regression, the lasso, principal components regression, and partial least squares. Hastie et al (2001) and Frank and Friedman (1993) independently provided performance comparisons of these methods. Ridge regression and the lasso are both penalization, or regularization, methods which place a constraint on the regression coefficients; and principal components regression and partial least squares are both variable subset selection methods which use linear combinations of the explanatory variables in the regression model. Ridge regression was designed specifically to reduce collinearity effects by penalizing the size of regression coefficients and decreasing the influence in the model of variables with relatively small variance in the design matrix. The lasso is a more recent development which also shrinks the regression coefficients, but shrinks the least significant variable coefficients to zero, thereby simultaneously performing coefficient penalization and model selection. The name of the lasso technique is derived from its function as a "least absolute shrinkage and selection operator"

(Tibshirani, 1996). Ridge regression and the lasso are deemed better candidates than principal components regression and partial least squares for addressing collinearity in local spatial regression models because they more directly reduce the variance in the regression coefficient while retaining interpretability of covariate effects.

To address the issue of collinearity in the GWR framework, Wheeler (2007) implemented a ridge-regression version of GWR, called GWRR, and found it was able to constrain the regression coefficients to counter local correlation present in an existing dataset. Another finding was a reduced prediction error for the response variable in GWRR compared with that from GWR. The lasso has not yet been introduced into the GWR framework in the literature, and its implementation in GWR is my goal in this paper. The lasso is appealing in the GWR framework because of its ability to carry out coefficient shrinkage and local model selection, as well as for its potential to improve on the performance of GWR for estimating the response variable—in terms of lower prediction and estimation errors. While ridge regression in GWR has the potential to control the variability in estimated regression coefficients, the lasso, in theory, should be able to constrain the coefficients and, additionally, perform local model selection by eliminating covariates from individual local models. Thus, the lasso offers a key advantage to ridge regression in the GWR framework and should lessen the reliance on approximate confidence intervals in GWR for the identification of insignificant local effects. In this paper I first review the GWR and lasso methods and then introduce the lasso in the GWR framework. I then demonstrate the benefit of using the geographically weighted lasso (GWL) through a comparative analysis with GWR and GWRR of two existing crime datasets and simulated data.

2 Methods

2.1 Geographically weighted regression

In the application of GWR, data are often mean measures of aggregate data at fixed points with associated spatial coordinates [for example, see the Georgia County example in Fotheringham et al (2002)], although this need not be the case. The spatial coordinates of the data are used in calculation of distances that are input into a kernel function to determine weights for spatial dependence between observations. Typically, a regression model is fitted at each data point, called a model calibration location. Local regression models are related through sharing data, but the dependence between regression coefficients at different model calibration locations is not specified in the model. For each calibration location, $i = 1, \ldots, n$, the GWR model at location i is

$$y(i) = \boldsymbol{X}(i)\boldsymbol{\beta}(i) + \varepsilon(i) \ , \tag{1}$$

where $y(i)$ is the dependent variable at location i, $\boldsymbol{X}(i)$ is the row vector of explanatory variables at location i, $\boldsymbol{\beta}(i)$ is the column vector of regression coefficients at location i, and $\varepsilon(i)$ is the random error at location i. The vector of estimation regression coefficients at location i is

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

where $\boldsymbol{X} = [\boldsymbol{X}^{\mathrm{T}}(1); \boldsymbol{X}^{\mathrm{T}}(2); \ldots; \boldsymbol{X}^{\mathrm{T}}(n)]^{\mathrm{T}}$ is the design matrix of explanatory variables, which typically includes a column of 1s for the intercept, $\mathbf{W}(i) = \mathrm{diag}[w_1(i), \ldots, w_n(i)]$ is the diagonal weights matrix that is calculated for each calibration location i and applies weights to observations $j = 1, \ldots, n$, with typically more weight applied to proximate, or neighboring, observations; $\boldsymbol{y}$ is the $n \times 1$ vector of dependent variable values; and $\hat{\boldsymbol{\beta}}(i) = (\hat{\beta}_{i0}, \hat{\beta}_{i1}, \ldots, \hat{\beta}_{ip})^{\mathrm{T}}$ is the vector of $p + 1$ local regression coefficients at location i for p explanatory variables and an intercept term.

The weights matrix, $\mathbf{W}(i)$, is calculated from a kernel function that places more emphasis on observations that are closer to the model calibration location *i*. There are numerous choices for the kernel function, including the Gaussian function, the bi-square nearest-neighbor function, and the exponential function. The exponential kernel function is utilized in this paper. The weight from the exponential kernel function between any location j and the model calibration location i is calculated as

$$w_j(i) = \exp\left(-\frac{d_{ij}}{\phi}\right) , \tag{3}$$

where d_{ij} is the distance between the calibration location i and location j, and ϕ is the kernel bandwidth parameter.

To fit the GWR model, the kernel bandwidth is first estimated, often in practice by leave-one-out cross-validation (CV) across all the calibration locations. CV is an iterative process that finds the kernel bandwidth with the lowest associated prediction error of all the responses $y(i)$. For each calibration location i, it removes the data for observation i in the model calibration at location i and predicts $y(i)$ from the other data points and the kernel weights associated with the current bandwidth. An alternative to CV in kernel-bandwidth estimation is the Akaike information criterion (AIC), as discussed by Fotheringham et al (2002). CV and the AIC are tools used in model selection, and more general information on the AIC and model selection are available elsewhere (Burnham and Anderson, 2004). It is currently unclear whether CV or AIC will generally return the same solution or whether one method should be favored in certain situations. The need for more research in this area is stressed by Farber and Páez (2007). Next, the kernel weights are calculated at each calibration location from the estimated bandwidth in the kernel function. Then, the regression coefficients are estimated at each model calibration location, and, finally, the responses are estimated by the expression $\hat{y}(i) = \boldsymbol{X}(i)\hat{\boldsymbol{\beta}}(i)$.

2.2 The lasso

Shrinkage methods, such as ridge regression and the lasso, introduce a constraint on the regression coefficients. The ridge-regression coefficients minimize the sum of a penalty on the size of the squared coefficients and the residual sum of squares (for details see Wheeler, 2007). The lasso takes the shrinkage of ridge regression a step further by potentially shrinking the regression coefficients of some variables to zero. The lasso specification is similar to that of ridge regression, but it has an absolute value coefficient penalty in place of the ridge squared coefficient penalty.

The lasso is defined as

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

subject to

$$\sum_{k=1}^{p} |\beta_k| \leqslant s , \tag{5}$$

where s is a parameter that controls the amount of regression coefficient shrinkage. Tibshirani (1996) notes that the lasso constraint $\sum_k |\beta_k| \leqslant s$ is equivalent to adding the penalty term $\lambda \sum_k |\beta_k|$ to the residual sum of squares; hence there is a direct correspondence between the parameters s and λ which control the amount of shrinkage of the regression coefficients. The equivalent statement for the lasso

coefficients is

$$\hat{\boldsymbol{\beta}}^{\mathrm{R}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{p} x_{ik}\beta_k \right)^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right\} . \quad (6)$$

The absolute value constraint on the regression coefficients makes the problem nonlinear and a typical way of solving this type of problem is by the use of quadratic programming.

There are, however, ways to estimate the lasso coefficients outside of the mathematical programming framework. Tibshirani (1996) provides an algorithm that finds the lasso solutions by treating the problem as a least-squares problem with 2^p inequality constraints, one for each possible sign of the β_ks, and applying the constraints sequentially. An even more attractive way of solving the lasso problem is proposed by Efron et al (2004a), who solve it with a small modification to the least angle regression (LARS) algorithm, which is a variation of the classic forward-selection algorithm in linear regression. The modification ensures that the sign of any nonzero estimated regression coefficient is the same as the sign of the correlation coefficient between the corresponding explanatory variable and the current residuals. Grandvalet (1998) shows that the lasso is equivalent to adaptive ridge regression, and develops an expectation-maximization algorithm to compute the lasso solution.

It is worthwhile describing in more detail the LARS and lasso algorithms of Efron et al (2004a) because these methods have not previously appeared in the geography literature at the time of writing. The LARS algorithm is similar in spirit to forward stepwise regression, which I now describe. The forward stepwise regression algorithm is:

(1) Start with all coefficients β_k equal to zero and set $\boldsymbol{r} = \boldsymbol{y}$, where $\boldsymbol{r}$ is the residual vector and $\boldsymbol{y}$ is the dependent variable vector.
(2) Find the predictor x_k most correlated with the residuals $\boldsymbol{r}$ and add it to the model.
(3) Calculate the residuals $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}$.
(4) Continue steps (2)–(3) until all predictors are in the model.

While the LARS algorithm is described in detail algebraically by Efron et al (2004a), Efron et al (2004b) restate the LARS algorithm as a purely statistical one with repeated fitting of the residuals, similar to the forward stepwise regression algorithm. The statistical statement of the LARS algorithm is:

(1) Start with all coefficients β_k equal to zero and set $\boldsymbol{r} = \boldsymbol{y}$.
(2) Find the predictor x_k most correlated with the residuals $\boldsymbol{r}$.
(3) Increase the coefficient β_k in the direction of the sign of its correlation with $\boldsymbol{r}$, calculating the residuals $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ at each increase, and continue until some other predictor x_m has as much correlation with the current residual vector $\boldsymbol{r}$ as does predictor x_k.
(4) Update the residuals and increase (β_k, β_m) in the joint least squares direction for the regression of $\boldsymbol{r}$ on (x_k, x_m) until some other predictor x_j has as much correlation with the current residual $\boldsymbol{r}$.
(5) Continue steps (2)–(4) until all predictors are in the model. Stop when $\mathrm{corr}(\boldsymbol{r}, x_j) = 0\ \forall j$, which is the ordinary least squares (OLS) solution.

As with ridge regression, typically the response variable is centered and the explanatory variables are centered and scaled to have equal (unit) variance prior to starting the LARS algorithm. In other words,

$$\sum_{i=1}^{n} y_i = 0, \qquad \sum_{i=1}^{n} x_{ij} = 0, \qquad \sum_{i=1}^{n} x_{ij}^2 = 1, \qquad \forall j = 1, \ldots, m . \quad (7)$$

Efron et al (2004a) show that a small modification to the LARS algorithm yields the lasso solutions. In a lasso solution the sign of any nonzero coefficient β_k must agree with the sign of the current correlation x_k and the residual. The LARS algorithm does not enforce this, but Efron and coauthors modify the algorithm to do so by removing β_k from the lasso solution if it changes in sign from the sign of the correlation of x_k and the current residual. This modification means that in the lasso solution the active set of variables in the solution does not necessarily increase monotonically as the routine progresses. Therefore, the LARS algorithm typically takes less iterations than does the lasso algorithm. The modified LARS algorithm produces the entire range of possible lasso solutions—from the initial solution with all coefficients equal to zero, to the final solution, which is also the OLS solution.

In some of the lasso algorithms, such as the modified LARS algorithm and the algorithm which Tibshirani describes, the shrinkage parameter (s) must be estimated before finding the lasso solutions. Hastie et al (2001) estimate the shrinkage parameter through tenfold cross-validation. Tibshirani (1996) uses fivefold cross-validation, generalized cross-validation, and a risk minimizer to estimate the shrinkage parameter, with the computational cost of the three methods decreasing in the same order. Efron et al (2004a) also recommend use of cross-validation to estimate the lasso parameter. One can define the lasso shrinkage parameter as

$$s = \frac{\sum_{k=1}^{p} \left|\hat{\beta}_k\right|}{\sum_{k=1}^{p} \left|\hat{\beta}_k^{\mathrm{OLS}}\right|}, \tag{8}$$

and s ranges from 0 to 1, where 0 corresponds to the initial lasso solution with all regression coefficients shrunk to 0 and 1 corresponds to the final lasso solution, which is also the OLS solution. Then, s can be viewed as the fraction of the OLS solution that is the lasso solution. This is the definition of the lasso shrinkage parameter that I will use in the subsequent work in this paper.

2.3 Geographically weighted lasso

The lasso can be implemented in GWR relatively easily, and the result is called here the geographically weighted lasso (GWL). An efficient implementation of the GWL outlined here uses the lars function from the package of the same name written in the R language by Hastie and Efron (see the R Project website: http://cran.r-project.org/). The lars function implements the LARS and lasso methods, where the lasso is the default method, and details are described in Efron et al (2004a; 2004b). To make use of the lars function in the GWR framework, the x and y variables input to the function must be weighted by the kernel weights at each model calibration location. The lars function must be run at each model calibration location. This can be done in one of two ways: separate models with local scaling of the explanatory variables (GWL – local); or one model with global scaling of the explanatory variables (GWL – global). The first method, local scaling, requires n calls of the lars function, one for each location, and the weighted x and y are centered and the x variables are scaled by the norm in the lars function. This effectively removes the intercept and equates the scales of the explanatory variables to avoid the problem of different scales. The local scaling version estimates the lasso parameter to control the amount of coefficient shrinkage at each calibration location, so there is a shrinkage parameter s_i estimated at each location i. Since I am working here in the GWR framework, I will estimate the model shrinkage and kernel bandwidth parameters by means of leave-one-out CV while minimizing the

root mean square prediction error (RMSPE) of the response variable. Therefore, the n s_i parameters and the kernel bandwidth ϕ must be estimated in GWL with CV before the final lasso coefficient solutions are estimated. I have chosen to estimate these parameters simultaneously, as the lasso solution will likely depend on the kernel bandwidth. The algorithm to estimate the local scaling GWL parameters using CV validation is detailed below.

For each attempted bandwidth ϕ in the binary search for the lowest RMSPE:
(1) Calculate the $n \times n$ weights matrix $\mathbf{W}$ from an $n \times n$ interpoint distance matrix $\mathbf{D}$ and ϕ from equation (3). $\mathbf{W}$ has for its ith row $[w_1(i), \ldots, w_n(i)]$, the diagonal elements in the matrix $\mathbf{W}(i)$, defined earlier with equation (2).
(2) For each location i, $i = 1, \ldots, n$:
  (a) Set $\mathbf{W}^{1/2}(i)$ to the square root of $\mathbf{W}(i)$ and $\mathbf{W}^{1/2}(i)_{ii} = 0$, that is, set the (i, i) element of the square root of the diagonal weights matrix to 0 to effectively remove observation i.
  (b) Set $\boldsymbol{X}_w = \mathbf{W}^{1/2}(i)\boldsymbol{X}$, and $\mathbf{y}_w = \mathbf{W}^{1/2}(i)\boldsymbol{y}$ using the square root of the kernel weights $\mathbf{W}(i)$ at location i.
  (c) Call lars $(\boldsymbol{X}_w, \mathbf{y}_w)$, save the series of lasso solutions, find the lasso solution that minimizes the error for y_i, and save this solution.
(3) Stop when there is only a small change in the estimated ϕ. Save the estimated ϕ.

In the previous algorithm, saving the lasso solution entails saving the estimated shrinkage fraction s_i at each location, as well as an indicator vector $\boldsymbol{b}$ of which variable coefficients are shrunken to zero. The algorithm uses a binary search to find the ϕ that minimizes the RMSPE. The small change in ϕ is set exogenously. The square root of the weights are used to weight the data because this is how the weights are applied to the data in the estimation of GWR regression coefficients in equation (2).

The algorithm to estimate the final local scaling GWL solutions after CV estimation of the shrinkage and kernel bandwidth parameters is:

(1) Calculate the $n \times n$ weights matrix $\mathbf{W}$ using an $n \times n$ interpoint distance matrix $\mathbf{D}$ and ϕ, where $\mathbf{W}$ is as previously defined.
(2) For each location i, $i = 1, \ldots, n$:
  (a) Set $\mathbf{W}^{1/2}(i)$ to the square root of $\mathbf{W}(i)$.
  (b) Set $\boldsymbol{X}_w = \mathbf{W}^{1/2}(i)\boldsymbol{X}$ and $\mathbf{y}_w = \mathbf{W}^{1/2}(i)\boldsymbol{y}$ using the square root of the diagonal weights matrix $\mathbf{W}(i)$ at location i.
  (c) Call lars $(\boldsymbol{X}_w \mathbf{y}_w)$ and save the series of lasso solutions.
  (d) Select the lasso solution that matches the cross-validation solution according to the fraction s_i and the indicator vector $\boldsymbol{b}$.

The second GWL method, global scaling, calls the lars function only once, using specially structured input data matrices. This method fits all the local models at once, using global scaling of the x variables. It also estimates only one lasso parameter to control the amount of coefficient shrinkage. The weighted design matrix for the global version is an $(nn) \times (np)$ matrix and the weighted response vector is $(nn) \times 1$. This results in an $(np) \times 1$ vector of estimated regression coefficients. The weighted design matrix is such that the design matrix is repeated n times, shifting p columns in its starting position each time it is repeated. The kernel weights for the first location are applied to the first n rows of the matrix, the weights for the second location are applied to the next n rows of the matrix, and so forth. The weighted response vector has the response vector repeated n times, with the weights for the first location applied

to the first n elements of the vector, and so on. The algorithm to estimate the global scaling GWL parameters using cross-validation is:

For each attempted bandwidth ϕ in the binary search for the lowest RMSPE:
(1) Calculate the $n \times n$ weights matrix $\mathbf{W}$ from an $n \times n$ interpoint distance matrix $\mathbf{D}$ and ϕ.
(2) Set diagonal of $\mathbf{W} = \mathbf{0}$.
(3) Set $\mathbf{y}_w^G = \mathbf{W}^s \times (\boldsymbol{1} \cdot \boldsymbol{y}^T)$ with the matrix $\mathbf{W}^s$ set by taking the square root of each element of the weights matrix $\mathbf{W}$ and $\boldsymbol{1}$ defined as the column vector of length n with all elements set to 1. The operator $\times$ indicates element-by-element multiplication here. Set $k = 1$ and $m = 1$.
(4) For each location i, $i = 1, \ldots, n$:
  (a) Set $j = kn - (n - 1)$ and $l = mp - (p - 1)$.
  (b) Set $\boldsymbol{X}_w = \mathbf{W}^{1/2}(i)\boldsymbol{X}$ from the square root of the kernel weights $\mathbf{W}(i)$ at location i. Set $\boldsymbol{X}_w^G(j : nk, l : pm) = \boldsymbol{X}_w$.
  (c) Set $k = k + 1$, and $m = m + 1$.
(5) Call lars $[\boldsymbol{X}_w^G, \text{vec}(\mathbf{y}_w^G)^T]$ and save the series of lasso solutions, where the vec() operator turns a matrix into a vector by sequentially placing columns, starting with the first, into one row.

In the previous algorithm, saving the lasso solution entails saving the estimated overall shrinkage fraction s, as well as a vector $\boldsymbol{b}$ that indicates which of the variable coefficients are shrunken to zero. The algorithm uses a binary search to find the ϕ that minimizes the RMSPE. The small change in ϕ is set exogenously.

The algorithm to estimate the final global scaling GWL solutions after CV estimation of the shrinkage and kernel bandwidth parameter is:

(1) Calculate the $n \times n$ weights matrix $\mathbf{W}$ using an $n \times n$ interpoint distance matrix $\mathbf{D}$ and ϕ.
(2) Set $\mathbf{y}_w^G = \mathbf{W}^s \times (\boldsymbol{1} \cdot \boldsymbol{y}^T)$ with the matrix $\mathbf{W} \times s$ set by taking the square root of each element of the weights matrix $\mathbf{W}$ and $\boldsymbol{1}$ defined as the column vector of length n with all elements set to 1. Set $k = 1$, and $m = 1$.
(3) For each location i, $i = 1, \ldots, n$:
  (a) Set $j = kn - (n - 1)$, and $l = mp - (p - 1)$.
  (b) Set $\boldsymbol{X}_w = \mathbf{W}^{1/2}(i)\boldsymbol{X}$ using the square root of the kernel weights $\mathbf{W}(i)$ at location i. Set $\boldsymbol{X}_w^G(j : nk, l : pm) = \boldsymbol{X}_w$.
  (c) Set $k - k + 1$, and $m = m + 1$.
(4) Call lars $[\boldsymbol{X}_w^G, \text{vec}(\mathbf{y}_w^G)^T]$ and save the series of lasso solutions.
(5) Select the lasso solution that matches the cross-validation solution according to the fraction s and the indicator vector $\boldsymbol{b}$.

In comparing the local and global scaling GWL algorithms, the global GWL algorithm requires more computational time due to the matrix inversion of a much larger matrix. The global GWR algorithm must invert an $(np \times np)$ matrix, whereas the local GWR algorithm must invert a $(p \times p)$ matrix n times. Considering that calculating the inverse of a general $j \times j$ matrix takes between $O(j^2)$ and $O(j^3)$ time (Banerjee et al, 2004), there can be quite a difference in the computation time for the two versions of GWR. In fact, global GWL may not be possible for large datasets, where 'large' is defined relative to the computing environment, as the memory requirements of the method could exceed available computer system memory. In terms of expected model performance, the local GWL method should produce a lower prediction error of the response variable than the global GWL method, as adding more shrinkage parameters generally increases model stability and hence

lowers prediction error. The benefit of global GWL may be in the lower estimation error of the regression coefficients, as the one shrinkage parameter may control excessive coefficient variation in GWR without stabilizing the model to the degree of local GWL. In summary, the local GWL should be faster than the global GWL and should have a lower prediction error. The local and global versions of GWL are compared empirically with each other and with GWR in the data example and simulation study in the next two sections.

3 Houston and Columbus crime examples

In this section, I demonstrate the use of the GWL methodology with two existing datasets dealing with crime in Houston, TX, and Columbus, OH, and compare the GWL results with those both from GWR and from GWRR. Waller et al (2007) previously analyzed violent-crime incidence related to alcohol sales and drug-law violations in the Houston dataset using GWR and a Bayesian hierarchical model. The Columbus crime dataset has been analyzed in spatial analysis work (Anselin, 1988) and in GWR-related work (LeSage, 2004; Wheeler, 2007). Previously (Wheeler, 2007) I have demonstrated, with diagnostic tools, the presence of collinearity in a GWR model for Columbus neighborhood crime rates using median income and housing values. The Columbus crime dataset is used here as an illustrative example to compare model performance and it is selected because of its problem with collinearity in the GWR model. In analyzing the Columbus crime data, I used a nearest neighbor bi-square kernel function with cross-validation to estimate the GWR kernel bandwidth (Wheeler, 2007). In this work, I use an exponential kernel function with CV to demonstrate that the collinearity issue persists with a different kernel function. All subsequent GWR-related models presented here use this kernel function.

Previously (Wheeler, 2007) I introduced the collinearity diagnostics of variance-decomposition proportions, condition indexes, and variance-inflation factors for GWR and applied them to the Columbus crime data to illustrate collinearity issues with the GWR model. The details for the diagnostics are available in that paper and are omitted here for brevity. Instead, I briefly summarize the results of applying the variance-decomposition diagnostic tool to the Columbus crime data. The GWR model is

$$y(i) = \beta_0(i) + \beta_1(i)x_1(i) + \beta_2(i)x_2(i) + \varepsilon(i) \ , \tag{9}$$

where y is residential and vehicle thefts combined per thousand people for 1980, x_1 is mean income, x_2 is mean housing value, and i is the index for neighborhoods.

Found through cross-validation, the estimated GWR kernel bandwidth $\hat{\phi}$ is 1.26. This estimated bandwidth is used in the variance decomposition of the kernel weighted design matrix to assess the collinearity in the model. The variance decomposition is done through singular value decomposition, and it has an associated condition index which is the ratio of the largest singular value to the smallest singular value. The condition index gives a measure of the instability in a matrix, as a large index implies that the variance is largely explained by a few components in the matrix. In diagnosing collinearity, the larger the condition index, the stronger is the collinearity among the columns of the GWR weighted design matrix. Belsley (1991) recommends a conservation value of 30 for a condition index that indicates collinearity, but suggests that the threshold value could be as low as 10 when there are large variance proportions for the same component. The variance-decomposition proportion is the proportion of the variance of a regression coefficient that is affiliated with one component of its decomposition. In addition, the presence of two or more variance proportions greater than 0.5 in one component of the variance decomposition indicates that collinearity exists

between at least two regression terms, one of which may be the intercept. Of the 49 records in the data, 6 have a condition index above 30, 12 have a condition index above 20, and 45 have a condition index above 10 and have large shared variances for the same component. There are many observations with large variance proportions (>0.5) from the same component, with the shared component being between a covariate and the intercept for some records and between the two covariates for other records. Of the 47 records with a large shared variance component, 23 are with the intercept and income, 4 are with the intercept and housing value, and 20 are between income and housing value. Overall, the diagnostic values indicate local collinearity in the GWR model.

Because of the collinearity in the GWR model, it is beneficial to apply the GWL models to these data and compare their performance with the GWR and GWRR models in terms of prediction and estimation error of the response variable. The accuracy of the estimated and predicted responses is measured by calculating the root mean square error (RMSE) and RMSPE, respectively. The RMSE is the square root of the mean of the squared deviations of the estimates from the true values, and should be small for accurate estimators. The results of fitting all four models to the data provide the error values shown in table 1. The lowest prediction error and estimation error among the four models are shown in bold. In this case the constrained versions of GWR do substantially better than GWR at predicting the dependent variable, and GWL – local performs better than GWRR and GWL – global. The RMSPE for the GWL – local model is 32% lower than for GWR and 24% lower than for GWRR. For estimating the dependent variable, GWL – global performs best, and substantially better than the other models. The RMSE for the GWL – global model is 17% lower than for the GWR model. Overall, GWL performs better than either GWR or GWRR. Figure 1 shows the estimated GWR coefficients and the GWL – local coefficients for income (β_1) and housing value (β_2). The figure shows the nature of the shrinkage in the estimated GWL coefficients, and how GWL enforces local model selection by shrinking some estimated coefficients to zero. In some neighborhoods, either the income value or the housing value has effectively been removed from the model. The estimated shrinkage parameter $\hat{s}$ is 0.75 for the GWL – global model, and the mean of the estimated shrinkage parameters is 0.76 for the GWL – local model.

The Houston crime data consist of 439 census tracts in the City of Houston, with attributes from year 2000. The number of violent crimes per person in each census tract is displayed in figure 2. There are a few census tracts with a total number of violent crimes that exceeds the population size. For the Houston crime data, the GWR model notation is the same as in equation (9), but where y is the number of violent crimes (murder, robbery, rape, and aggregated assault) per person, x_1 is the number of drug law violations per person, x_2 is the number of alcohol outlets per person, and i is the index for census tracts. Since the distribution of the response variable is

Table 1. RMSPE (root mean square prediction error) and RMSE (root mean square error) of the response variable for the GWR (geographically weighted regression), GWRR (geographically weighted ridge regression), GWL (geographically weighted lasso) – global, and GWL – local models for the Columbus crime data.

Method	RMSPE (y)	RMSE (y)
GWR	11.074	2.640
GWRR	9.808	2.800
GWL – global	9.946	**2.197**
GWL – local	**7.483**	2.687

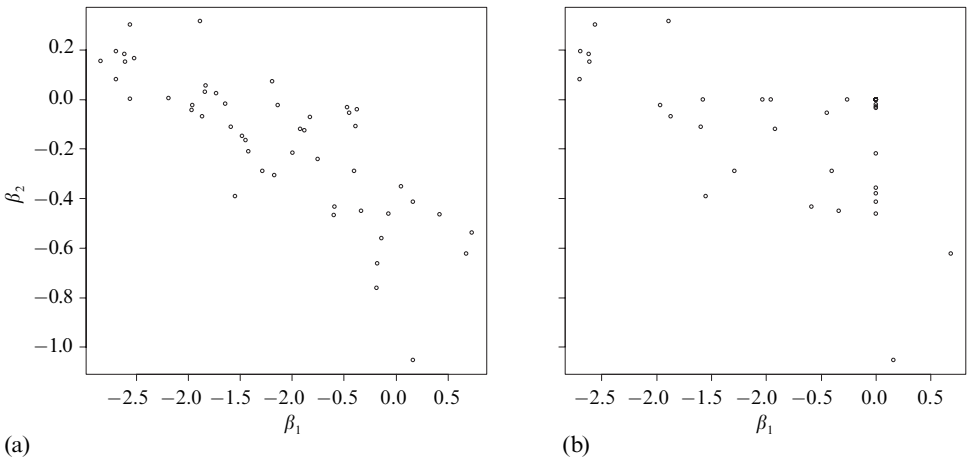


Figure 1. GWR (geographically weighted regression) estimated coefficients (a) and GWL (geographically weighted lasso) – local estimated coefficients (b) for the income (β_1) and housing value (β_2) covariates in the Columbus crime data.

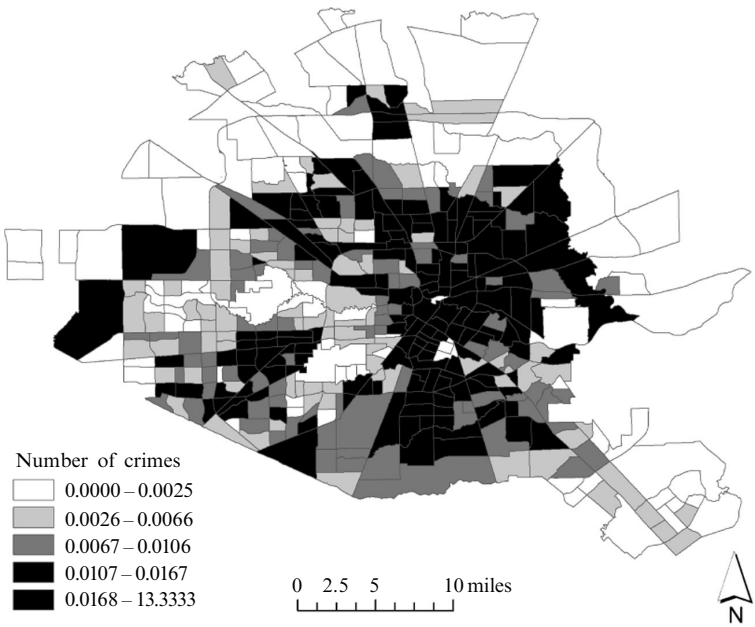


Figure 2. Number of violent crimes per person in Houston in 2000.

positively skewed, the natural logarithm of violent crimes was used in the model and also the natural logarithm of both covariates was used to maintain linear relationships with violence rates. The GWR estimated kernel bandwidth $\hat{\phi}$ is 0.89, found through CV. To assess collinearity in the GWR model, I use the variance-decomposition diagnostic. The variance-decomposition proportions and condition indexes are listed in table 2 for records with the largest condition indexes. These 10 records are labeled in the plot of estimated GWR coefficients for the drug and alcohol covariates in figure 3(a). These labeled records comprise many of the more extreme points in the plot. Observation 153 is clearly the most extreme of the points, as it has the largest value for the drug-rate effect and the smallest value for the alcohol-rate effect. In table 2,

Table 2. Record number, condition index (k), and variance-decomposition proportions (p_1—intercept, p_2—drug, p_3—alcohol) for the Houston crime data.

Record number	k	p_1	p_2	p_3
1	27.60	0.996	0.995	0.136
2	87.66	0.992	0.992	0.001
5	21.29	0.995	0.993	0.188
27	24.25	0.997	0.690	0.947
33	35.45	0.865	0.949	0.045
67	29.49	0.994	0.982	0.371
114	40.58	0.739	0.988	0.283
116	39.45	0.579	0.996	0.922
153	38.38	0.737	0.999	0.609
158	21.94	0.955	0.942	0.006

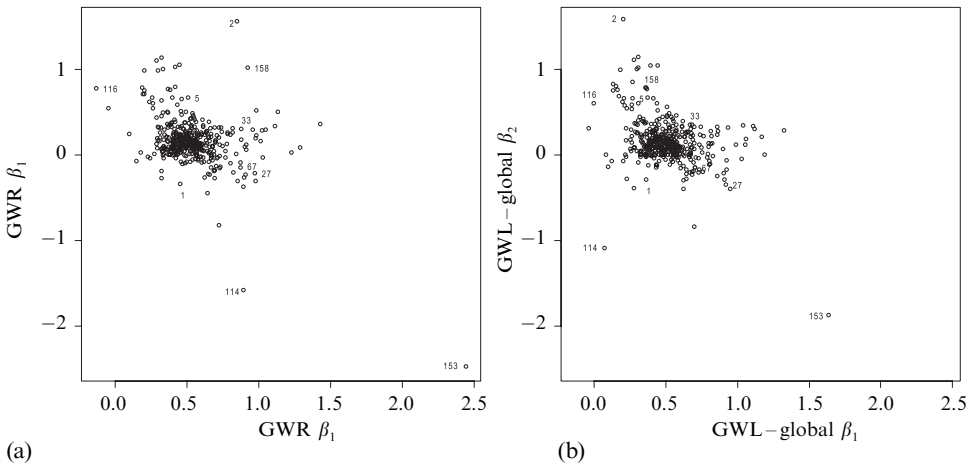


Figure 3. Geographically weighted regression (GWR) estimated coefficients (a) and geographically weighted lasso (GWL) – global estimated coefficients (b) for the drug (β_1) and alcohol (β_2) covariates in the Houston crime data.

this record has large variance proportions for the same component for all three regression terms. Of the 439 records in the dataset, 5 have a condition index above 30, 10 have a condition index above 20, and 41 have a condition index above 10. There are 411 records in the data with large variance proportions (>0.5) from the same component, with the shared component being between a covariate and the intercept for some records and between the two covariates for other records. Overall, the variance-decomposition proportions and condition index values indicate the presence of local collinearity in the GWR model.

Given the presence of local collinearity in the GWR model for violent crime in Houston, the constrained versions of GWR were also fitted and compared with GWR in terms of model performance. The RMSE and RMSPE values for the response variable are listed in table 3 for the GWR, GWRR, GWL – global, and GWL – local models. As with the Columbus crime data, the constrained versions of GWR improve on GWR in prediction of the response variable. The GWL – local model again produces the lowest RMSPE—18% lower than the GWR model. In estimating violent crime, the GWL models improve upon the GWR model. The GWL – global model produces the lowest RMSE and its RMSE is 14% lower than in the GWR model. The estimated

Table 3. RMSPE (root mean square prediction error) and RMSE (root mean square error) of the response variable for the GWR (geographically weighted regression), GWRR (geographically weighted ridge regression), GWL (geographically weighted lasso) – global, and GWL – local models for the Houston crime data.

Method	RMSPE (y)	RMSE (y)
GWR	0.720	0.342
GWRR	0.713	0.349
GWL – global	0.714	**0.300**
GWL – local	**0.590**	0.311

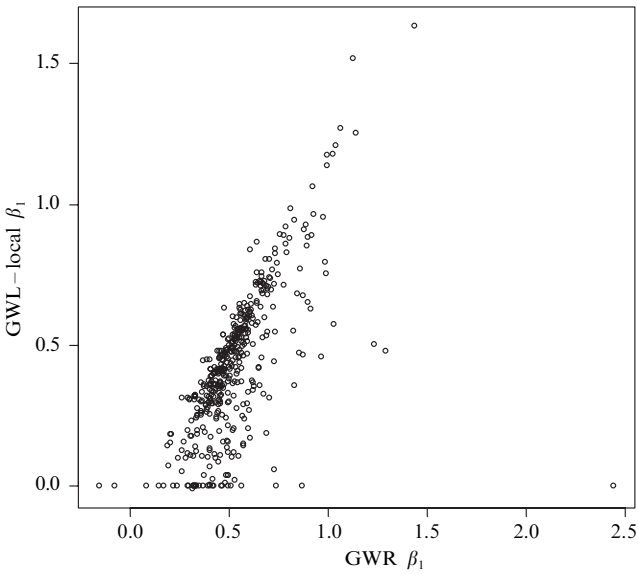


Figure 4. GWR (geographically weighted regression) estimated coefficients and GWL (geographically weighted lasso) – local estimated coefficients for the drug (β_1) covariate in the Houston crime data.

regression coefficients for the GWL – global model in figure 3(b) show that the GWL model has penalized some of the most extreme coefficients in the GWR model in figure 3(a), particularly record 153. Figure 4 displays the estimated regression coefficients for the drug covariate from the GWL – local model plotted against the estimated coefficients from the GWR model. This figure shows the effective shrinkage of the GWL – local model, where the GWL – local model shrinks certain larger GWR coefficients—some to zero. The large estimated regression coefficient for record 153 is greatly reduced in the GWL – local model. The estimated shrinkage parameter $\hat{s}$ is 0.92 for the GWL – global model, and the mean of the estimated shrinkage parameters is 0.65 for the GWL – local model. The correlation in the estimated regression coefficients for the drug and alcohol covariates is -0.41 with GWR, -0.39 with GWRR, -0.37 with GWL – global, and 0.03 with GWL – local. The results with the crime-data examples consistently show that the constrained versions of GWR improve on the performance of GWR, and that the GWL – local model produces the lowest prediction error and the GWL – global model produces the lowest estimation error.

4 Simulation study

In this section I use a simulation study to evaluate and compare the accuracy of the predicted and estimated responses and the estimated regression coefficients from the GWR, GWRR, and GWL models. The accuracy of the models is estimated both when there is no collinearity in the explanatory variables and when there is collinearity—expressed at various levels. The expectation is that the GWL model will improve on GWR for regression-coefficient estimation when there is collinearity in the model. Another expectation is that the GWL model will improve on GWR for prediction and estimation of the response variable. While it has been conventional for researchers to apply a newly introduced method to an existing dataset as a demonstration of the utility of the method, use is made here of simulated data to learn about the performance of the method in a comparative setting. It is necessary to use simulation in order to set the 'true' values of the regression coefficients, which are unknown with real data, so that it is possible to measure the deviation from the truth of the estimates from competing models. The simulation study presented here is not intended to be exhaustive but, rather, is an appealing alternative to the use of existing data for demonstrating the performance of the introduced method in a certain situation.

The data-generating model in the simulation study has four explanatory variables, with the true coefficients used to generate the dataset equal to nearly zero for one explanatory variable. The model to generate the data for this simulation study is

$$y(i) \;=\; \beta_1^*(i)x_1(i) + \beta_2^*(i)x_2(i) + \beta_3^*(i)x_3(i) + \beta_4^*(i)x_4(i) + \varepsilon(i) \;, \tag{10}$$

where x_1, x_2, x_3, x_4 are the first four principal components from a random sample drawn from a multivariate normal distribution of dimension 10 with a mean vector of zeros and an identity covariance matrix, the errors ε are sampled independently from a normal distribution with mean 0 and variance of τ^{2*}, and i denotes the location. The star notation denotes the true values of the parameters used to generate the data. Note that there is no true intercept in the model used to generate the data and an intercept is not fitted in the simulation study. The data points are equally spaced on a 14×14 grid, for a total of 196 observations. The goal of the simulation study is to use the model in equation (10) to generate the data and see if the regression coefficient estimates match $\boldsymbol{\beta}^*$ and if the estimated and predicted responses approximate $\boldsymbol{y}$ for the GWR, GWRR, and GWL models. To produce comparable summary measures of deviance of the estimates and responses from the true values, we generate 100 realizations of the coefficient process, estimate the model parameters and responses for each data realization, measure the error in the estimates, and then produce average errors over the many realizations of the data. The use of 100 realizations of the data-generating process is advantageous compared with the use of one dataset because it allows us to assess model performance over 100 datasets.

Each realization of the true regression coefficients, $\boldsymbol{\beta}^*$, is sampled through the distribution

$$[\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta] \;=\; \mathrm{N}(\boldsymbol{I}_{n\times 1} \otimes \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \;, \tag{11}$$

where the vector $\boldsymbol{\mu}_\beta = (\mu_{\beta_0}, \ldots, \mu_{\beta_p})^{\mathrm{T}}$ contains the means of the regression coefficients corresponding to each of the p explanatory variables, and spatial dependence in the coefficients is specified through the covariance, $\boldsymbol{\Sigma}_\beta$. We assume a separable covariance matrix (Gelfand et al, 2003) for $\boldsymbol{\beta}$ of the form

$$\boldsymbol{\Sigma}_\beta \;=\; \mathbf{H}(\gamma) \otimes \mathbf{T} \;, \tag{12}$$

where $\mathbf{H}(\gamma)$ is the $n \times n$ correlation matrix that captures the spatial association between the n locations, γ is the spatial dependence parameter, $\mathbf{T}$ is a positive-definite $p \times p$

matrix for the covariance of the regression coefficients at any spatial location, and $\otimes$ denotes the Kronecker product operator, which is the multiplication of every element in $\mathbf{H}(\gamma)$ by $\mathbf{T}$. In the specification of the variance in the distribution for $\boldsymbol{\beta}$ [equation (12)], the Kronecker product results in an $np \times np$ positive-definite covariance matrix, since $\mathbf{H}(\gamma)$ and $\mathbf{T}$ are both positive definite. The elements of the correlation matrix $\mathbf{H}(\gamma)$, $H(\gamma)_{jk} = \rho(i_j - i_k; \gamma)$, are calculated from the exponential function $\rho(d; y) = \exp(-d/\gamma)$.

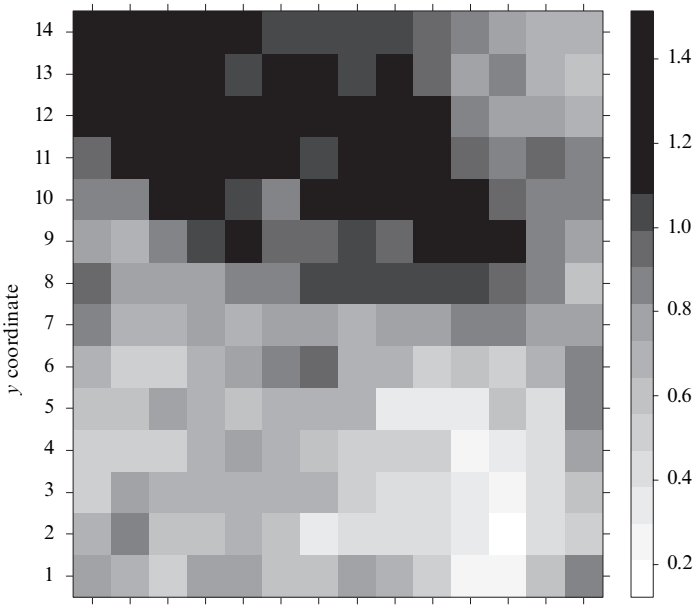
For this simulation study, the true values used to generate the data are $\boldsymbol{\mu}^*_\beta = (1, 5, 5, 0)$, $\tau^{2*} = 1$, $\gamma^* = 10$, and $\mathbf{T}^* = \text{diag}(0.1, 0.5, 0.5, 0.0000001)$, where diag() makes a diagonal matrix with the input numbers on the diagonal. The mean of 0 and the small variance for the fourth type of regression coefficient produce a variable effect that is effectively zero across the study area. More information regarding the drawing of samples from the coefficient distribution utilized here is available in Wheeler and Calder (2007). In general, as γ^* increases there are more consistent and clearer patterns in the true regression coefficients. The range is the distance beyond which the spatial association becomes insignificant and is approximately $3\gamma^*$ with the covariance function parameterization used here, so there is some dependence in the coefficients for each covariate throughout the study area. Figure 5 illustrates the pattern in the true coefficients for two covariates for one realization of the coefficient process, and shows that there is some smoothness and spatial variation in the true coefficients when $\gamma^* = 10$. This pattern reflects a situation where there is spatial parametric nonstationarity: in other words, one in which GWR is intended to be applied.

In the simulation study we start with no substantial collinearity in the model and systematically increase collinearity until the explanatory variables are nearly perfectly collinear. This is done by replacing one of the original explanatory variables with one created from a weighted linear combination of the original explanatory variables, where the weight determines the amount of correlation of the variables. The formula for the new weighted variable is

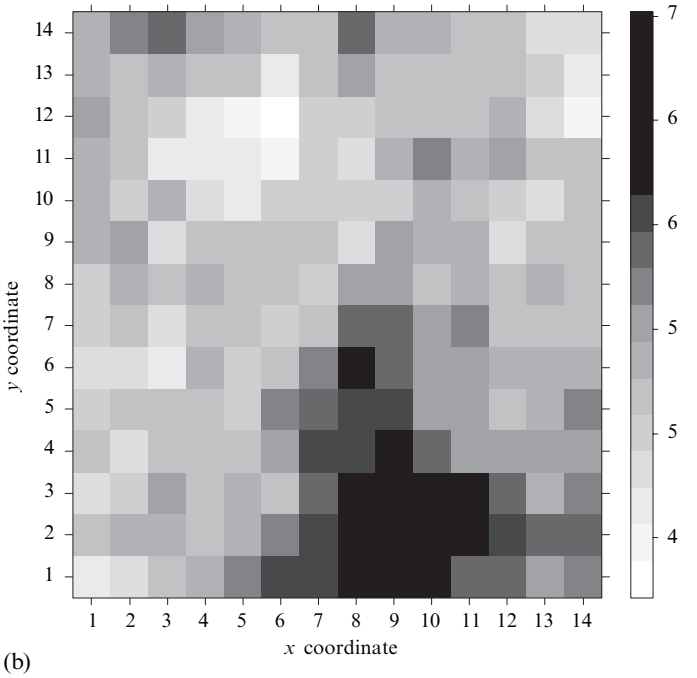
$$x_2^c = cx_1 + (1 - c)x_2 \ , \tag{13}$$

where x_2^c replaces x_2 in the model in equation (10) and c is a weight between 0 and 1. The simulation study is carried out with four levels of explanatory variable collinearity. The weights used in equation (13) to create the collinearity are $c = (0.0, 0.5, 0.7, 0.9)$, which coincide with explanatory variable correlation of $r = (0.00, 0.74, 0.93, 0.99)$. These levels of correlation correspond to no collinearity as a baseline, and then moderate, strong, and nearly perfect collinearity. In this study, the model parameters and responses are estimated for each realization for each of the following models: GWR, GWRR, GWL – global, and GWL – local. The kernel bandwidth is estimated for each data realization by means of CV and is thus potentially different for each realization. To measure the accuracy of the estimated regression coefficients and estimated responses, the RMSPE and RMSE are calculated for the responses $\hat{\boldsymbol{y}}$ and the RMSE is calculated for the coefficients $\hat{\boldsymbol{\beta}}$ for each data realization. The average RMSE for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{y}}$ and the average RMSPE for $\hat{\boldsymbol{y}}$ are then calculated from averaging the individual RMSEs and RMSPEs from the 100 realizations of the coefficients.

The average RMSPE and RMSE for $\hat{\boldsymbol{y}}$ and the average RMSE for $\hat{\boldsymbol{\beta}}$ for each model are listed in table 4. The lowest value for each error measure for each level of variable correlation (column) is in bold. The results in the table show that the GWL – local model produces the lowest prediction error of the response. The GWL – local model prediction error is approximately 20% lower on average than the GWR error. This is not an unexpected result, as the GWL – local model adds the most local penalization parameters to the GWR model—which should lower the prediction error by stabilizing the model.



(a)



(b)

Figure 5. Coefficient patterns for the first two β^* parameters for one realization of the coefficient process in the simulation study: (a) β_1^* and (b) β_2^*.

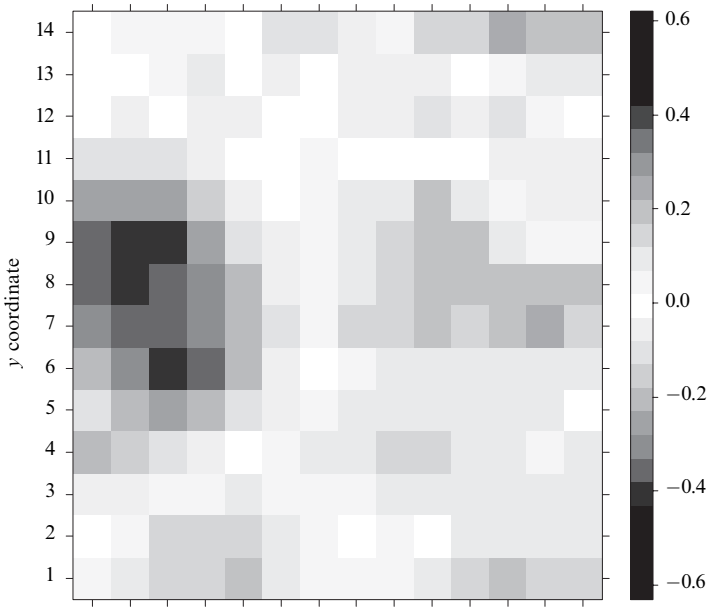
Table 4. RMSPE (root mean square prediction error) and RMSE (root mean square error) of the response variable and RMSE of the regression coefficients for each model used in the simulation study at four levels of explanatory variable correlation.

Method	Correlation			
	$r = 0.00$	$r = 0.74$	$r = 0.93$	$r = 0.99$
RMSPE (y)				
GWR	1.187	1.154	1.158	1.174
GWRR	1.187	1.153	1.158	1.168
GWL – global	1.181	1.144	1.148	1.158
GWL – local	**0.928**	**0.932**	**0.954**	**0.959**
RMSE (y)				
GWR	0.856	0.873	0.869	0.860
GWRR	0.856	0.873	0.871	0.877
GWL – global	0.849	0.862	0.858	0.821
GWL – local	**0.662**	**0.675**	**0.669**	**0.700**
RMSE (β)				
GWR	0.503	0.553	0.689	1.586
GWRR	0.504	0.554	0.691	1.515
GWL – global	**0.499**	**0.549**	**0.686**	**1.513**
GWL – local	1.815	2.147	2.101	1.991

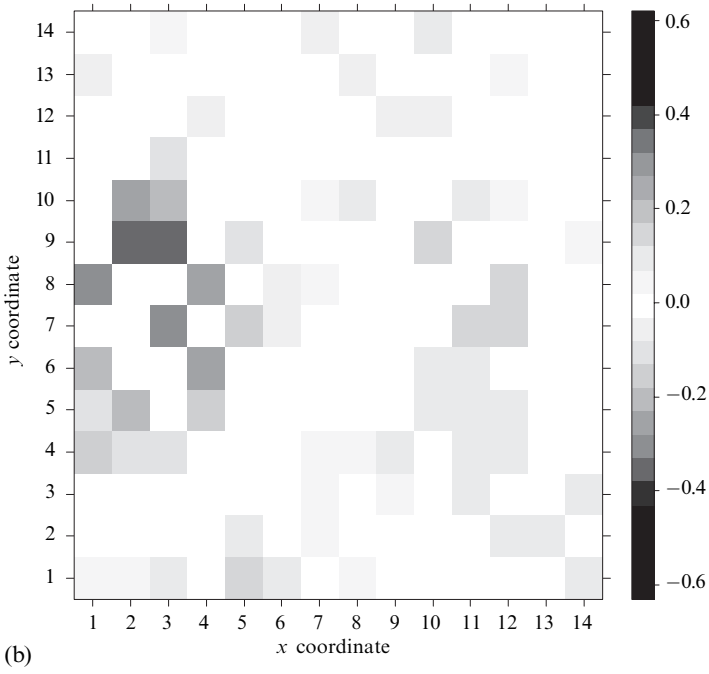
The next-best performer in terms of RMSPE of the response is the GWL – global model. GWR has the highest average prediction error of the response at each level of collinearity. These results demonstrate that adding penalization terms for the regression coefficients in GWR results in a lower prediction error of the response than with GWR.

The RMSE results in the table show that the GWL – local model produces the lowest estimation error of the response at all levels of collinearity: the GWL – local estimation error is approximately 20% lower on average than the GWR error. Overall, the simulation study shows that the GWL models perform better than GWR in explaining the response variable. The better performance of the two versions of GWL relative to the GWR is not unexpected, given that the GWL methods can shrink the regression coefficients to zero to match the true values for one of the variables in an effort to estimate the response variable. Taken together, the results in table 4 indicate that the GWL – local model is best for predicting and estimating the response variable in the presence of an insignificant explanatory variable.

The RMSE results for $\hat{\boldsymbol{\beta}}$ in the table show that the GWL – global model produces the lowest average estimation error of the regression coefficients. The GWR model performs next best, except when there is nearly perfect collinearity, and the GWRR model outperforms GWR. An explanation for the leading performance of the GWL – global model is that it applies moderate shrinkage to the coefficients towards zero for the variable with true coefficients set to zero to in effect remove its effect from the model. It strikes a balance between the stronger shrinkage of the GWL – local and the weaker shrinkage of GWRR. The GWL – local model overshrinks the estimated regression coefficients that are not truly near-zero and, as a result, the GWL – local clearly performs the worst in estimating the regression coefficients. The simulation-study results here suggest that the GWL – local should not be used for inference on regression coefficients, but should instead be used for prediction and estimation of the response variable—where its level of shrinkage is more suitable for model stabilization. The RMSE results for $\hat{\boldsymbol{\beta}}$ also suggest that an improvement in marginal inference on the regression coefficients in the presence of collinearity or insignificant explanatory variables is possible with the GWL – global model used in place of GWR.



(a)



(b)

Figure 6. Coefficient estimates for β_4 from GWR (geographically weighted regression) (a) and GWL (geographically weighted lasso) – local (b) for one realization of the coefficient process in the simulation study.

An example of the difference in the estimated coefficients from GWR and the GWL – local is illustrated in figure 6, which displays the estimated coefficients for β_4 from the GWR and GWL – local models for one realization of the coefficient process when there is no collinearity in the model. The true coefficients for this variable are all approximately zero, so a plot of them would be an even white surface. Figure 6 shows that the GWL – local model estimates more of the coefficients near zero for this variable through coefficient shrinkage than does GWR. This results in lower prediction and estimation error of the response variable.

Often in traditional regression analyses, researchers consider using only penalization methods, such as the lasso and ridge regression, when there are many explanatory variables to include in the model. However, the results from this simulation study show that one can improve on GWR in terms of prediction and estimation of the response and estimation of the regression coefficients for even relatively small models. There may be situations, however, where it is beneficial to use GWR without penalization when prediction is not of primary interest, particularly for quick descriptive analyses of spatially varying relationships in data where collinearity is not present. However, I anticipate that the benefits of penalization in GWR for prediction will increase with an increasing number of potentially correlated explanatory variables.

5 Conclusions

There has been increasing interest in spatially varying relationships between variables in recent years in the spatial analysis literature. Recent attempts at modeling these relationships have resulted in numerous forms of geographically weighted regression, which has its technical origins in locally weighted regression. While GWR offers the promise of an understanding of the spatially varying relationships between variables, local collinearity in the weighted explanatory variables used in GWR can produce unstable models and dependence in the local regression coefficients, which can interfere with conclusions about these relationships. While GWR has been applied to numerous real-world datasets in the literature, there has been inadequate consideration of the accuracy of inferences derived from this model and an unclear distinction as to its use for prediction and estimation of the response variable versus its role in inference on the relationships between variables. The work in this paper uses real and simulated data to evaluate the accuracy of the response-variable estimates and predictions provided from GWR and constrained versions of GWR, namely, geographically weighted ridge regression and the newly introduced geographically weighted lasso models. I have also evaluated the accuracy of regression coefficient from GWR, GWRR, and the GWL models using simulated data, while considering the presence of collinearity and an insignificant variable.

The work presented here shows that it is possible to implement the lasso in the GWR framework to perform regression-coefficient shrinkage while simultaneously performing local model selection and reducing prediction and estimation error of the response variable. The data example and simulation-study results show that the penalized versions of GWR can outperform GWR in terms of response variable prediction and estimation, both when there is no collinearity and where there are various levels of collinearity in the model. In both the real and the simulated data, the GWL – local model produces the lowest prediction error of the response variable among the methods considered. For the actual data, the GWL – global model produced the lowest response variable estimation error. For estimating regression coefficients, the simulation study results suggest that some gain in accuracy over GWR is possible with the GWL – global model. Other related preliminary work (Wheeler, 2006) suggests that the geographically weighted lasso may perform better

in dependent-variable estimation than a Bayesian spatially varying coefficient process (SVCP) model (Gelfand et al, 2003) which may be viewed as an alternative to GWR. It has been demonstrated (Wheeler and Calder, 2007) that the SVCP model can offer more accurate coefficient inference and lower response variable estimation error than GWR, although at a greater computational cost. A theoretical comparison of the performance of GWR, all penalized versions of GWR, and the SVCP model is planned for future work. In summary, the penalized versions of GWR introduced in this paper extend the method of GWR to improve prediction and estimation of the response variable, which is in agreement with its statistical theoretical origins.

References

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

Banerjee S, Carlin B P, Gelfand A E, 2004 *Hierarchical Modeling and Analysis for Spatial Data* (Chapman and Hall, Boca Raton, FL)

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

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

Burnham K, Anderson D, 2004 *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach* (Springer, Berlin)

Casetti E, 1992, "Generating models by the expansion method: applications to geographic research" *Geographical Analysis* **4** 81 – 91

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

Congdon P, 2003, "Modelling spatially varying impacts of socioeconomic predictors on mortality outcomes" *Journal of Geographical Systems* **5** 161 – 184

Efron B, Hastie T, Johnstone I, Tibshirani R, 2004a, "Least angle regression" *The Annals of Statistics* **32** 407 – 451

Efron B, Hastie T, Johnstone I, Tibshirani R, 2004b, "Rejoinder to least angle regression" *The Annals of Statistics* **32** 494 – 499

Farber S, Páez A, 2007, "A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations" *Journal of Geographical Systems* **9** 371 – 396

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

Frank I E, Friedman J H, 1993, "A statistical view of some chemometrics regression tools" *Technometrics* **35** 109 – 148

Gelfand A E, Kim H, Sirmans C F, Banerjee S, 2003, "Spatial modeling with spatially varying coefficient processes" *Journal of the American Statistical Association* **98** 387 – 396

Grandvalet Y, 1998, "Least absolute shrinkage is equivalent to quadratic penalization", in *ICANN'98, Volume I of Perspectives in Neural Computing* Eds L Niklasson, M Boden, T Ziemske (Springer, Berlin) pp 201 – 206

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

LeSage J P, 2004, "A family of geographically weighted regression models", in *Advances in Spatial Econometrics. Methodology, Tools and Applications* Eds L Anselin, R J G M Florax, S J Rey (Springer, Berlin) pp 241 – 264

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

Martinez W L, Martinez A R, 2002 *Computational Statistics Handbook with Matlab* (Chapman and Hall, Boca Raton, FL)

Neter J, Kitner 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

Tibshirani R, 1996, "Regression shrinkage and selection via the lasso" *Journal of the Royal Statistical Society B* **58** 267 – 288

Waller L, Zhu L, Gotway C, Gorman D, Gruenewald P, 2007, "Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models" *Stochastic Environmental Research and Risk Assessment* **21** 573 – 588

Wheeler D, 2006 *Diagnostic Tools and Remedial Methods for Collinearity in Linear Regression Models with Spatially Varying Coefficients* PhD dissertation, Department of Geography, The Ohio State University, Columbus, OH

Wheeler D, 2007, "Diagnostic tools and a remedial method for collinearity in geographically weighted regression" *Environment and Planning A* **39** 2464 – 2481

Wheeler D, Calder C, 2007, "An assessment of coefficient accuracy in linear regression models with spatially varying coefficients" *Journal of Geographical Systems* **9** 145 – 166

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