Chapter 12
Geographically Weighted Regression

Yang Yang

Abstract The family of geographically weighted regression (GWR) methods has seen its wide applications in a variety of fields including ecology, agriculture, social science, and public health. The popularity of these methods stems from their ability to depict spatial heterogeneity, easy interpretation of outputs, and the availability of user-friendly software tools. These methods have evolved extensively in the recent decade to address the challenges of multicolinearity in predictors and variable selection in the era of big data, and a comprehensive review is needed to raise both awareness and practical validation of these progresses. Equally needed is an up-to-date introduction to the associated software packages, especially those developed on the popular statistical software platform R. This chapter provides a systematic overview of the foundation and recent development of the methodology of GWR, with a balance between rigidity and practicality. Via a case study, this chapter also offers step-by-step guideline to the use of three major GWR-dedicated R packages, including their facilities for multicolinearity diagnosis and variable selection. We hope a broadened user group of these methods will in turn motivate more methodological advances and improve the contribution of GWR methods to global health.

Keywords Geographically weighted regression · Spatial heterogeneity · Variable selection · Multicolinearity · GWR software

12.1 Introduction

Health outcomes such as chronic conditions and infectious diseases typically exhibit spatial and temporal variation, driven by both risk factors and random errors. While

Y. Yang (✉)
Department of Biostatistics, School of Public Health and Health Professions & Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA
e-mail: yangyang@ufl.edu

X. Chen, (Din) D.-G. Chen (eds.), *Statistical Methods for Global Health and Epidemiology*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-3-030-35260-8_12

changes in risk factors explain a significant amount of variations in health outcomes, it is quite often they do not explain all. Additional variation may be accounted for by spatial or temporal heterogeneity in the effects of risk factors. An example is shown in Fig. 12.1, where the effects of drug infection history (left) and drug rehabilitation history (right) on the risk of HCV infection in a Yi Ethnicity Autonomous Prefecture in Southwestern China were estimated using the geographically weighted regression (GWR) approach (Zhou et al. 2016). This figure clearly demonstrates spatial heterogeneity in the estimated effects. Such heterogeneity is more prominent at larger spatial scales, e.g., states, countries or continents.

Figure 12.2 shows the spatial distributions of local R-squares (upper left) and selected GWR-estimated regression coefficient estimates for teenage birth rate in rural counties of the United States during 2003 (Shoff & Yang, 2012). The effects of some factors such as clinic rate have opposite signs across different regions. If such heterogeneity is ignored and the covariate effects are assumed spatially homogeneous, the effects could be estimated as null and are thus very misleading.

A natural and widely used solution to spatial heterogeneity is to group data points by spatial regions, where the regions are usually defined as administrative units, e.g., counties or states, or as ecological zones. A categorical variable indicating the regions is then incorporated into the analysis. As a main effect, this variable can capture spatial variation in the intercept, i.e., the mean level of the dependent variable. Spatial variation in the effects of other predictors can be investigated by formulating interaction terms between these predictors and the region indicator variables. Assuming that n individuals are observed in a total of J spatial regions, a typical statistical presentation of the model is

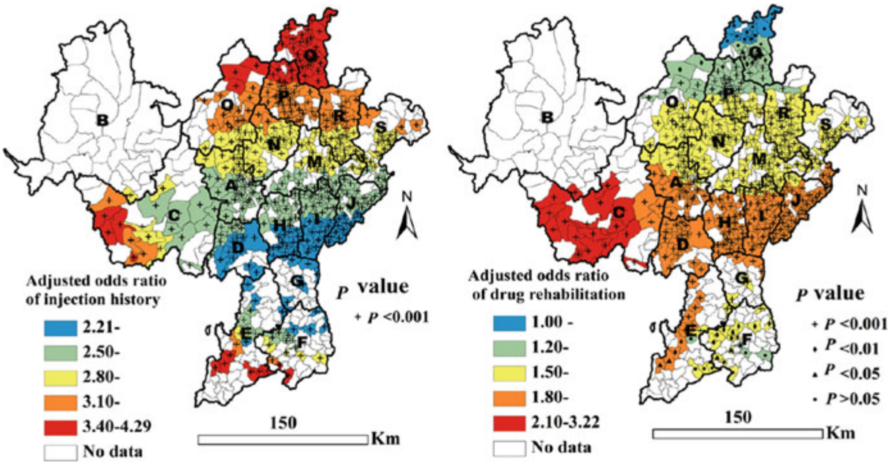

Fig. 12.1 Spatial distribution of GWR-fitted adjusted odds ratios for drug injection history (left) and drug rehabilitation history (right) with regard to their effects on HCV infection in Southwestern China (Zhou et al. 2016)

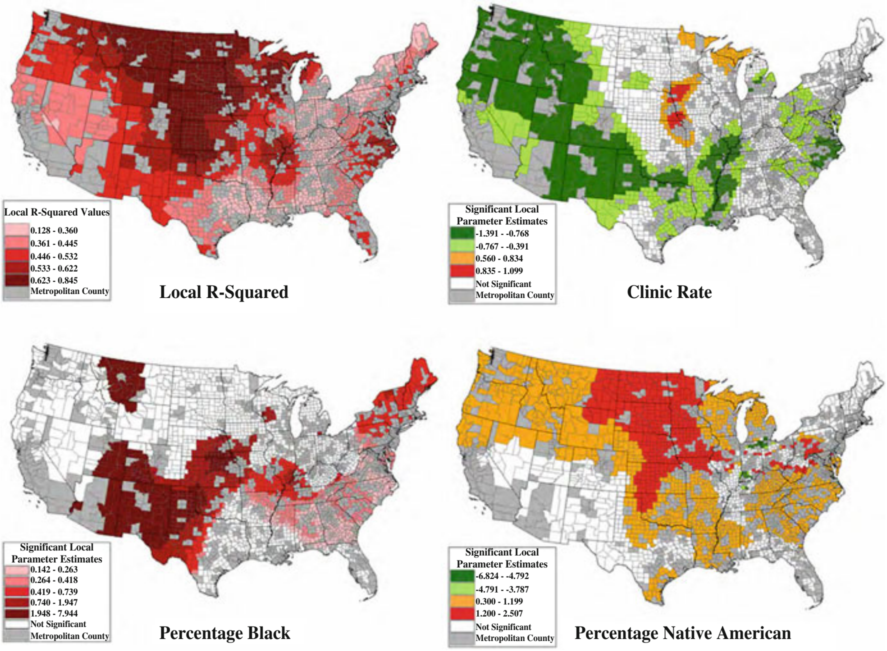

Fig. 12.2 Spatial distribution of GWR-fitted local R-squares and estimated coefficients of predictors for average teenage birth rates during 1999–2001 among non-metropolitan counties in the United States (Shoff & Yang, 2012)

$$y_i = \alpha + \beta x_i + \sum_{j=1}^{J-1} \eta_j z_{ij} + \sum_{j=1}^{J-1} \xi_j z_{ij} x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where, for individual i, y_i is the response, x_i is a risk factor of interest, z_{ij} indicates whether observation i is in region j (1=yes, 0=no), $j = 1, \ldots, J-1$, and ϵ_i is the error term that is often assumed identically and independently distributed (i.i.d.) as normal with mean 0 and an unknown variance σ^2. We assume only one risk factor for illustrative purpose. The coefficients α and β are the intercept and slope for the reference region defined by $z_{i1} = \ldots = z_{i(J-1)} = 0$. For the region associated with $z_{ij} = 1$, the intercept and slope are $\alpha + \eta_j$ and $\beta + \xi_j$, respectively. Several disadvantages of this approach are worth noting. First, the grouping of data points into regions is often a choice of convenience, not necessarily matching the true geographic pattern in data-generating mechanism. For example, the spatial variation of data-generating mechanism may be smooth over the whole study area rather than with abrupt changes at boundaries of regions as assumed by the grouping approach. Second, there is no widely accepted guideline on choosing the level and number of spatial regions. A few large regions may not be adequate to delineate the spatial heterogeneity, creating issues such as ecological fallacy, whereas too many small regions may encounter the identifiability issue because of sparse data in some regions. Ecological fallacy refers to the phenomenon that heterogeneous individual trends within a region are represented by a homogeneous but misleading trend for the whole region when individual data are aggregated by region for analysis (Wakefield 2008).

An extension of the grouping approach is the expansion method (Casetti 1972; Jones 1992), which models the regression coefficients as continuous functions of the geocoordinates. A general presentation of the expansion method is

$$y_i = \alpha(u_i, v_i) + \beta(u_i, v_i)x_i + \epsilon_i,$$

where (u_i, v_i) are the geocoordinates of individual i, e.g., longitude and latitude. A simple example is the linear mapping

$$\alpha_i = \alpha_0 + \alpha_1 u_i + \alpha_2 v_i,$$
$$\beta_i = \beta_0 + \beta_1 u_i + \beta_2 v_i.$$

Higher orders of u_i and v_i as well as their interactions can be added when needed. A major limitation of the expansion method is that spatial patterns in real life are often much more complex and cannot be satisfactorily captured by polynomials of geocoordinates. In addition, the interpretation of the estimated coefficients can be very difficult. Fotheringham, Charlton, and Brunsdon (1998) compared the expansion method and the geographically weight regression (GWR) using the data of limiting long-term illness (LLTI) in Northeastern England. Risk predictors under consideration were unemployment rate, household crowdedness, proportion of single-parent families among children <5 years, social class and population density. All predictors were modeled as linear effects in all models. Figure 12.3 shows the spatial distribution of (a) standardized LLTI in the study area; (b) intercept under the expansion method when linear expansion was applied to intercept and all slopes; (c) intercept under the expansion method when quadratic expansion (including interaction) was applied to intercept and all slopes; and (d) intercept under the GWR approach. The gradients of intercept under linear expansion (Fig. 12.3b) is largely consistent with the observed pattern of LLTI, although the direction of decrease appeared more from northeast to southwest for the model than from north to south as shown by the data. Under the quadratic expansion (Fig. 12.3c), however, the direction of gradients flipped, now increasing from northeast to southwest. The distribution of the intercept term of the GWR (Fig. 12.3d) reflects the spatial pattern observed in the data and reserves some non-directional spatial differences compared to the linear expansion method.

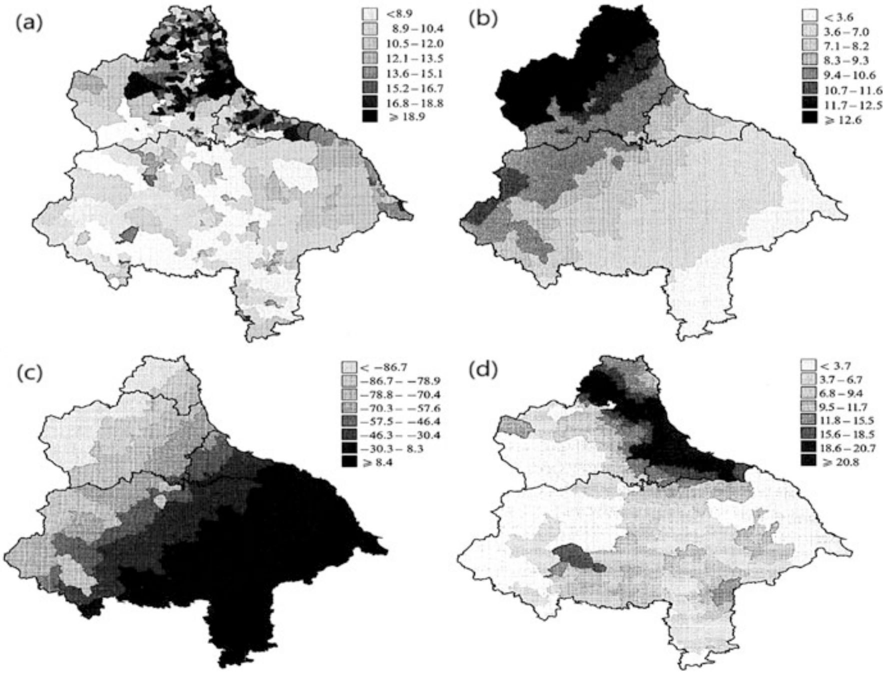

Fig. 12.3 (**a**) Spatial distribution of limiting long-term illness data in Northeastern England; (**b**) distribution of the intercept term under linear expansion; (**c**) distribution of the intercept term under quadratic expansion; (**d**) distribution of the intercept term fitted by the GWR method. The subfigures are adapted from Fotheringham et al. (1998)

12.2 Theory

12.2.1 Basic Model Structure and Inference

The model expression of GWR is very similar to that of the expansion method, except that no specific functional form is assumed for the dependence of coefficients on geocoordinates. Suppose we are regressing an dependent variable on p predictors. For notional simplicity, define $s = (u, v)$ for the geocoordinates. For a given individual i at location $s_i = (u_i, v_i)$, the model to be fitted is

$$y_i = \beta_0(s_i) + \sum_{k=1}^{p} \beta_k(s_i)x_{ik} + \epsilon_i = \boldsymbol{X}_i^{\tau}\boldsymbol{\beta}(s_i) + \epsilon_i, \tag{12.1}$$

where $\boldsymbol{X}_i = (1, x_{i1}, \ldots, x_{ip})^{\tau}$ and $\boldsymbol{\beta}(s_i) = (\beta_0(s_i), \ldots, \beta_p(s_i))^{\tau}$ are the vectors of covariates and coefficients, respectively, and τ denotes transpose of vectors or matrices. As usual, we assume $\epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} Normal(0, \sigma^2)$. The coefficients,

$\beta_0(s), \ldots, \beta_p(s)$ can be viewed as continuous spatial functions defined at any point s in the study area rather than only at the observed locations $s_1, \ldots, s_n$. The essence of GWR is the construction of the diagonal weight matrix $\boldsymbol{W}(s_i)_{n\times n} = \text{diag}\big(w_1(s_i), \ldots, w_n(s_i)\big)$, where each diagonal element $w_j(s_i)$ is determined by a predefined distance between s_i and s_j, $j = 1, \ldots, n$. Let $\boldsymbol{X}$ be a $n \times (p+1)$ matrix with $\boldsymbol{X}_i$ as its ith row, $i = 1, \ldots, n$, and let $\boldsymbol{Y} = (y_1, \ldots, y_n)^\tau$ be the column vector of the dependent variable. Following standard linear model theory, minimizing the sum of weighted squares of residuals chosen as the objective function

$$\underset{\boldsymbol{\beta}(s_i)}{\arg\min} \, ||\boldsymbol{W}(s_i)^{1/2}(Y - \boldsymbol{X}\boldsymbol{\beta}(s_i))||^2 = \underset{\boldsymbol{\beta}(s_i)}{\arg\min} \sum_{j=1}^{n} w_j(s_i)\big(y_j - \boldsymbol{X}_j^\tau \boldsymbol{\beta}(s_i)\big)^2 \quad (12.2)$$

yields the WLS estimates for the coefficients

$$\hat{\boldsymbol{\beta}}(s_i) = \boldsymbol{B}(s_i)\boldsymbol{Y}, \quad \text{where } \boldsymbol{B}(s_i) = \big(\boldsymbol{X}^\tau \boldsymbol{W}(s_i)\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\tau \boldsymbol{W}(s_i).$$

This is also the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}(s_i)$ under the normal assumption for ϵ_i's with a correlation matrix $\boldsymbol{W}(s_i)$. Model (12.1) is to be fitted n times, one at each individual location in the data. It can also be fitted at an arbitrary location s in the study area, but the model need to be restated in a more general form

$$y(s) = \beta_0(s) + \sum_{k=1}^{p} \beta_k(s)x_k(s) + \epsilon(s) = \boldsymbol{X}(s)^\tau \boldsymbol{\beta}(s) + \epsilon(s).$$

However, $y(s)$ and $\boldsymbol{X}(s)$ are observed only at the locations of the study-sampled individuals, i.e., s_i, $i = 1, \ldots, n$.

To obtain a variance estimate for $\hat{\boldsymbol{\beta}}(s_i)$ so that we can construct confidence intervals for all the local coefficients, we need to estimate σ^2, the variance of the error term. Let $\hat{y}_i = \boldsymbol{X}_i^\tau \hat{\boldsymbol{\beta}}(s_i)$ be the model-fitted value at s_i, and let $\hat{\boldsymbol{Y}} = (\hat{y}_1, \ldots, \hat{y}_n)^\tau$. We can write $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$, where

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{X}_1^\tau \boldsymbol{B}(s_1) \\ \boldsymbol{X}_2^\tau \boldsymbol{B}(s_2) \\ \vdots \\ \boldsymbol{X}_n^\tau \boldsymbol{B}(s_n) \end{pmatrix}$$

is analogous to the *hat* matrix in the ordinary linear regression setting. Let $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ be the vector of residuals. An important statistic is the residual sum of square

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} \big(y_i - \hat{y}_i\big)^2 = \hat{\boldsymbol{\epsilon}}^\tau \hat{\boldsymbol{\epsilon}} \\ &= (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\tau (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{Y}^\tau (\boldsymbol{I} - \boldsymbol{H})^\tau (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}, \end{aligned}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix. Following Leung, Mei, and Zhang (2000), under the assumption that $\mathbb{E}(\hat{\boldsymbol{Y}}) = \mathbb{E}(\boldsymbol{Y})$, RSS can further written as

$$\begin{aligned} \text{RSS} &= \big[\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y})\big]^{\tau}(\boldsymbol{I} - \boldsymbol{H})^{\tau}(\boldsymbol{I} - \boldsymbol{H})\big[\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y})\big] \\ &= \boldsymbol{\epsilon}^{\tau}(\boldsymbol{I} - \boldsymbol{H})^{\tau}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon} \end{aligned}$$

Note that

$$\begin{aligned} \mathbb{E}(\text{RSS}) &= \mathbb{E}\big[\boldsymbol{\epsilon}^{\tau}(\boldsymbol{I} - \boldsymbol{H})^{\tau}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}\big] \\ &= \mathbb{E}\Big[trace\big(\boldsymbol{\epsilon}^{\tau}(\boldsymbol{I} - \boldsymbol{H})^{\tau}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}\big)\Big] \\ &= \mathbb{E}\Big[trace\big((\boldsymbol{I} - \boldsymbol{H})^{\tau}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\tau}\big)\Big] \\ &= trace\Big[(\boldsymbol{I} - \boldsymbol{H})^{\tau}(\boldsymbol{I} - \boldsymbol{H})\mathbb{E}\big(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\tau}\big)\Big] \\ &= \sigma^2 \nu_1 \end{aligned}$$

where $\nu_1 = trace\big((\boldsymbol{I} - \boldsymbol{H})^{\tau}(\boldsymbol{I} - \boldsymbol{H})\big) = n - \big[2tr(\boldsymbol{H}) - tr(\boldsymbol{H}^{\tau}\boldsymbol{H})\big]$ is the degree of freedom of the RSS, and $2tr(\boldsymbol{H}) - tr(\boldsymbol{H}^{\tau}\boldsymbol{H})$ represents the effective number of parameters (Fotheringham, Brunsdon, & Charlton, 2002). The trace function of a matrix is simply the sum of diagonal elements of that matrix. An unbiased estimate for σ^2 is then $\hat{\sigma}^2 = \text{RSS}/\nu_1$, and the variance-covariance matrix of $\hat{\boldsymbol{\beta}}(s_i)$ can then be estimated as

$$\widehat{\text{Cov}}\big[\hat{\boldsymbol{\beta}}(s_i)\big] = \boldsymbol{B}(s_i)\boldsymbol{B}(s_i)^{\tau}\hat{\sigma}^2. \tag{12.3}$$

The Wald-type 95% confidence interval for each coefficient can be established as

$$\hat{\beta}_k(s_i) \pm 1.96 \times \sqrt{\widehat{\text{Var}}\big[\hat{\beta}_k(s_i)\big]},$$

where $\widehat{\text{Var}}\big[\hat{\beta}_k(s_i)\big]$ is the kth diagonal element of $\widehat{\text{Cov}}\big[\hat{\boldsymbol{\beta}}(s_i)\big]$, $k = 0, 1, \ldots, p$.

12.2.2 Constructing Weights

The choices of the weights $w_j(s_i)$ depends on the nature of the data. When the individuals under observation are relatively large spatial units such as zip codes, counties or states, the neighbor indicator is a reasonable choice, i.e., $w_j(s_i) = 1$ if units i and j share a common border and 0 otherwise. The concept of adjacency-based weighting can be generalized to k-order neighbors (Zhang & Murayama, 2000). The neighbor structure can be viewed as an undirected graph with directly neighbored units connected by edges. If the shortest path between two units has k edges, then the two units are k-order neighbors. Consequently, a more general adjacency-based weighting scheme is to let $w_j(s_i) = 1$ if units i and j are k-order neighbors for any $k \leq h$, where h is called the bandwidth. When it is appropriate to consider predefined geographic distances between individuals, a kernel function of the distance subject to a bandwidth constraint is often used, e.g., the bisquare function

$$w_j(s_i) = \begin{cases} \left(1 - d_{ij}^2/h^2\right)^2, & \text{if } d_{ij} < h, \\ 0, & \text{otherwise} \end{cases},$$

where h is the bandwidth. Two other popular choices are the Gaussian density kernel, $w_j(s_i) = \exp(-d_{ij}^2/h^2)$, and the exponential kernel, $w_j(s_i) = \exp(-d_{ij}/h)$. Any kernel function, $K(d)$, of distance d that satisfies (1) $K(0) = 1$, (2) $K(\infty) = 0$, (3) $K(d) > 0$ for $d > 0$ and (4) non-increasing can be considered. A monotone decreasing kernel function ensure that higher weights are put on observations that are closer to the current location s_i. The bandwidth, h, further controls how fast the weight should decay according to the distance, and its choice is crucial for the inferential performance of the method. If h is large so that the decay is slow, then distant observations contributed almost equally as the nearby ones. On one hand, the effective sample size increases and thus the estimates will be more stable and less variant; on the other hand, however, if the coefficients vary substantially over the space, severe bias is likely to occur. This is a large-bias-small-variance situation. Conversely, if h is too small, only close-by observations will contribute, leading to small bias but large variance. Consequently, the choice of h is actually an issue of balance between bias and variance, and cross-validation (CV) procedure is recommended to choose the bandwidth (Brunsdon, Fotheringham, & Charlton, 1996). Let $\hat{y}_{(i)}(h) = \boldsymbol{X}_i^\tau \hat{\boldsymbol{\beta}}(s_i, h)$ be the fitted value at the location of individual i with a given value of h, where the model is fitted with individual i excluded. We use the notation $\hat{\boldsymbol{\beta}}(s_i, h)$ instead of $\hat{\boldsymbol{\beta}}(s_i)$ to reflect its dependence on h. Then, h is chosen by minimizing the sum of squares of residuals:

$$h^\star = \arg\min_h \sum_{i=1}^n \left[y_i - \hat{y}_{(i)}(h)\right]^2.$$

Another popular objective function for bandwidth calibration is related to the Akeike Information Criterion (AIC) that balances between goodness-of-fit and model parsimony,

$$AIC_c(h) = \log\left[RSS/n\right] + \frac{n + trace(H)}{n - 2 - trace(H)},$$

where both RSS and H depend on h (Fotheringham et al. 2002). The optimal bandwidth is then $h^\star = \arg\min_h AIC_c(h)$. In practice, empirical data or expert opinions may also inform the choice of h.

12.2.3 Testing Spatial Nonstationarity

Two questions naturally arise with the use of GWR:

1. Is the GWR fitting the data better than the ordinary least squares (OLS) regression that assumes spatial stationarity in covariate coefficients?
2. Which covariate coefficients have significant spatial heterogeneity?

The first question is about testing the global presence of spatial nonstationarity, while the second is about testing spatial variation for each individual predictor. Statistically, the null hypotheses to be tested are

1. H_{0a}: $\boldsymbol{\beta}(s_1) = \boldsymbol{\beta}(s_2) = \cdots = \boldsymbol{\beta}(s_n)$.
2. H_{0b}: $\beta_k(s_1) = \beta_k(s_2) = \cdots = \beta_k(s_n)$ for a given k.

For the global testing, Brunsdon et al. (1996) suggested the CV-derived bandwidth parameter $h^\star$ as a test statistic, as the smaller the bandwidth the larger the spatial heterogeneity. A GWR with $h^\star = \infty$ is equivalent to the OLS. For the testing of specific coefficients, Brunsdon et al. (1996) recommended the sample variance (or standard deviation) of each location-specific coefficient, i.e., $S_k = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_k(s_i) - \frac{1}{n}\sum_{j=1}^{n}\hat{\beta}_k(s_j)\right)^2$. As the theoretical distribution under each null hypothesis is not easy to derive, a permutation-based approach was proposed (Brunsdon et al. 1996). Under the global null hypothesis H_{0a}, the data $(y_i, \boldsymbol{X}_i)$ can be randomly permuted across all locations. Suppose M permuted sample data sets are generated, and let the bandwidth derived based on the mth sample dataset by h_m. The p-value of the observed value $h^\star$ from the original data for testing H_{0a} is given by

$$\Pr(h \le h^\star) = \frac{1}{M}\sum_{j=1}^{M}\mathcal{I}(h_m \le h^\star),$$

where $\mathcal{I}(c)$ is the indicator function taking 1 if condition c is true and 0 otherwise. Although $h = \infty$ corresponds to spatial stationarity theoretically, the meaningful null distribution of h in reality is bounded by the size of the study area. Brunsdon et al. (1996) suggested using the sample permutation approach to find the null distribution of S_k; however, the obtained distribution is under H_{0a} instead of H_{0b}. As the parameter space under H_{0a} is a subspace of that under H_{0b}, using the distribution of S_k under the global null will not necessarily yield a valid type I error, and will likely lack sufficient statistical power.

The computational burden of the permutation approach is heavy for large n, especially for the calculation of h_m which itself involves a search for the optimal bandwidth for each sample dataset. Leung et al. (2000) explored the possibility of asymptotical tests for the two hypotheses. The test statistic they proposed for H_{0a} is $F_a = \frac{\mathrm{RSS}_{gwr/\nu_1}}{\mathrm{RSS}_{ols/(n-p-1)}}$, where $\mathrm{RSS}_{gwr} = \boldsymbol{Y}^\tau(\boldsymbol{I} - \boldsymbol{H})^\tau(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ and $\mathrm{RSS}_{ols} = \boldsymbol{Y}^\tau(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\tau\boldsymbol{X})^{-1}\boldsymbol{X}^\tau)\boldsymbol{Y}$ are the residual sums of squares based on GWR and OLS, respectively. Leung et al. (2000) suggested the null distribution of F_1 be approximated by the $F(\nu_1/\nu_2, n-p-1)$ distribution, where $\nu_2 = trace\Big[\big((\boldsymbol{I} - \boldsymbol{H})^\tau(\boldsymbol{I} - \boldsymbol{H})\big)^2\Big]$. H_{0a} is rejected if the observed value of F_1 is less than the $\alpha \times 100\%$ percentile for a given type I error α. To test H_{0b}, the suggested test statistic is $F_b = \frac{S_k^2/\gamma_1}{\hat{\sigma}^2}$, where $\gamma_1 = \frac{1}{n}trace\Big[\boldsymbol{B}^\tau(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{B}\Big]$, $\boldsymbol{J}$ is a $n \times n$ matrix with all elements equal to one,

$$\boldsymbol{B}_k = \begin{pmatrix} \boldsymbol{e}_k^\tau \boldsymbol{B}(s_1) \\ \boldsymbol{e}_k^\tau \boldsymbol{B}(s_2) \\ \vdots \\ \boldsymbol{e}_k^\tau \boldsymbol{B}(s_n) \end{pmatrix},$$

and $\boldsymbol{e}_k$ is a vector of zeros of length $p+1$ except for the $(k+1)$th element being one. The null distribution of F_b is approximated by $F(\gamma_1^2/\gamma_2, \nu_1^2/\nu_2)$, where $\gamma_2 = trace\Big[\big(\frac{1}{n}\boldsymbol{B}^\tau(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{B}\big)^2\Big]$. H_{0b} is rejected if the observed F_b exceeds the $(1-\alpha) \times 100\%$ percentile of the null distribution.

These approximate asymptotic null distributions, however, remain to be rigorously justified in the sense that the numerators and denominators are not necessarily independent. In addition, Fotheringham et al. (2002) noted that the computation load of these asymptotic tests is as heavy as the permutation tests in practice. Finally, the F_1 statistic could be used as an alternative to h in the permutation test for the global null hypothesis H_{0a}, in particular when h is fixed rather than calibrated in cross-validation, e.g., the binary neighbor indicator matrix.

Mei, Wang, and Zhang (2006) proposed a resampling based approach to test the hypotheses in a more general form

- H_0: $\beta_k(s_1) = \beta_k(s_2) = \cdots = \beta_k(s_n)$ for $k \in \Delta$, where Δ is a given subset of $\{1, \ldots, p\}$.
- H_1: All coefficients vary over space.

The set Δ could be a single coefficient, a subset of or all of the coefficients. This bootstrap procedure goes with the following steps:

1. Fit the unrestricted model to obtain the residuals $\hat{\boldsymbol{\epsilon}} = (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^\tau = (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}$ and the residual sum of squares $RSS_1 = \boldsymbol{Y}^\tau(\boldsymbol{I} - \boldsymbol{H}_1)^\tau(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}$, where H_1 is the hat matrix. Let $\hat{\boldsymbol{\epsilon}}_c = (\hat{\epsilon}_{c1}, \ldots, \hat{\epsilon}_{cn})^\tau$ be the centered residuals, i.e., $\hat{\epsilon}_{ci} = \hat{\epsilon}_{ci} - \frac{1}{n}\sum_{j=1}^n \hat{\epsilon}_j$.

2. Fit the null model using the two-step method of Fotheringham et al. (2002) to obtain the residual sum of squares $RSS_0 = \boldsymbol{Y}^\tau(\boldsymbol{I} - \boldsymbol{H}_0)^\tau(\boldsymbol{I} - \boldsymbol{H}_0)\boldsymbol{Y}$, where H_0 is the hat matrix under the null.
3. Compute the observed statistic $T = \frac{RSS_0 - RSS_1}{RSS_1}$.
4. For $m = 1, , \dots, M$, draw with replacement random samples $\tilde{\boldsymbol{\epsilon}}^{(m)} = (\tilde{\epsilon}_1^{(m)}, \dots, \tilde{\epsilon}_n^{(m)})^\tau$ from the centered residuals $\hat{\boldsymbol{\epsilon}}_c$, formulate new responses as $\tilde{\boldsymbol{Y}}^{(m)} = (\tilde{y}_1^{(m)}, \dots, \tilde{y}_n)^{(m)})^\tau = \boldsymbol{H}_0\boldsymbol{Y} + \tilde{\boldsymbol{\epsilon}}^{(m)}$, and calculate the statistics $\tilde{T}^{(m)} = \frac{\widetilde{RSS}_0^{(m)} - \widetilde{RSS}_1^{(m)}}{\widetilde{RSS}_1^{(m)}}$, where $\widetilde{RSS}_k^{(m)} = (\tilde{\boldsymbol{Y}}^{(m)})^\tau(\boldsymbol{I} - \boldsymbol{H}_k)^\tau(\boldsymbol{I} - \boldsymbol{H}_k)\tilde{\boldsymbol{Y}}^{(m)}$, $k = 0, 1$.
5. The p-value is calculated as $p = \frac{1}{M}\sum_{m=1}^{M}\mathcal{I}(\tilde{T}^{(m)} \geq T)$.

12.2.4 Geographically Weighted Generalized Linear Models

Analogous to the generalized linear global models (GLM) for fitting binary and count data, geographically weighted generalized linear models (GWGLM) have also been developed. The GLM theory is based on the exponential family of statistical distributions in the form

$$f(y|\theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} - c(y, \phi)\right\},$$

to which many commonly seen distributions such as normal, binomial, Poisson and negative binomial (when the overdispersion parameter is assumed known) belong. Parameters θ and ϕ are called the canonical parameter and the dispersion parameter, respectively, and the functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are assumed known. For example, the Poisson distribution with parameter λ can be written as

$$f(y|\lambda) = \exp\left\{\frac{y\log(\lambda) - \lambda}{1} - \log\Gamma(y + 1)\right\},$$

with $\theta = \log(\lambda)$, $b(\theta) = \exp(\theta)$, $a(\phi) = 1$, and $c(y, \phi) = \log\Gamma(y + 1)$. The mean and variance of y are related to this parameterization via $\mathbb{E}(Y) = \mu = b'(\theta)$ and $\text{Var}(Y) = a(\phi)b''(\theta)$, where $b'(\theta)$ and $b''(\theta)$ are the first and second derivatives of $b(\theta)$. The mean μ is related to linear predictors $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \boldsymbol{x}^\tau\boldsymbol{\beta}$ via the link function $\eta = g(\mu)$. Table 12.1 lists the model components for several distributions in the exponential family.

A general Iteratively Reweighted Least Squares (IRLS) algorithm is available to fit these models (Nelder & Wedderburn, 1972), which can be adapted to the GWGLM setting (da Silva & Rodrigues, 2014; Fotheringham et al. 2002; Nakaya, Fotheringham, Brunsdon, & Charlton, 2005). The following algorithm is modified from Fotheringham et al. (2002). The algorithm is applied to the local fitting at

Table 12.1 GLM components for selected distributions

Distribution	θ	ϕ	$a(\phi)$	$b(\theta)$	$c(y,\phi)$	$g(\mu)$
$Normal(\mu,\sigma^2)$	μ	σ^2	ϕ	$\theta^2/2$	$-\frac{1}{2}\big[\frac{y^2}{\phi}+\log(2\pi\phi)\big]$	μ
$Binomial(n,p)^{\text{a}}$	$\log\big(\frac{p}{1-p}\big)$		$1/n$	$\log(1+e^{\theta})$	$\log\big[\binom{n}{ny}\big]$	$\log\big(\frac{\mu}{1-\mu}\big)$
$Poisson(\mu)$	$\log(\mu)$		1	e^{θ}	$-\log(y!)$	$\log(\mu)$
$NB(\mu,\alpha)^{\text{b}}$	$\log\big(\frac{\mu}{r+\mu}\big)$		1	$-r\log\big(1-e^{\theta}\big)$	$\log\frac{\Gamma(r+y)}{y!\Gamma(r)}$	$\log(\mu)$

[a] y/n is viewed as the random variable

[b] For negative binomial to be in exponential family, r is assumed known

each location s_i separately, i.e., all parameters $(\boldsymbol{\beta},\eta,\mu,\theta,\phi)$ depend on s_i, but we suppress such dependence in notation for simplicity. η, μ and θ further depend on $\boldsymbol{x}_j$ at all observation locations when we fit the local model centered around s_i, and thus we use η_j, μ_j and θ_j, $j=1,\ldots,n$ to reflect such dependence.

1. Choose initial estimate $\boldsymbol{\beta}^{(0)}$ and $\phi^{(0)}$, and obtain $\eta_j^{(0)}=\boldsymbol{x}_j^{\tau}\boldsymbol{\beta}^{(0)}$, $\mu_j^{(0)}=g^{-1}(\eta_j^{(0)})$, and $\theta_j^{(0)}=b'^{-1}(\mu_j^{(0)})$, where $g^{-1}(\cdot)$ and $b'^{-1}(\cdot)$ are inverse functions of $g(\cdot)$ and $b'(\cdot)$ respectively. For iterations $k=0,1,\ldots$, do the following steps:
2. Derive the adjusted dependent variable $z_j^{(k)}=\eta_j^{(k)}+g'(\mu_j^{(k)})(y_j-\mu_j^{(k)})$.
3. Construct a diagonal matrix $\boldsymbol{A}^{(k)}$ with its jth diagonal element being
$$a_{jj}^{(k)}=\Big\{\big[g'(\mu_j^{(k)})\big]^2a(\phi^{(k)})b''(\theta_j^{(k)})\Big\}^{-1}.$$
4. Update the coefficients as
$$\boldsymbol{\beta}^{(k+1)}=(\boldsymbol{X}^{\tau}\boldsymbol{W}\boldsymbol{A}^{(k)}\boldsymbol{X})^{-1}\boldsymbol{X}^{\tau}\boldsymbol{W}\boldsymbol{A}^{(k)}\boldsymbol{Z}^{(k)},$$
where $\boldsymbol{X}$ and $\boldsymbol{W}$ are the covariate matrix (including first column of 1s) and weight matrix as defined before, and $\boldsymbol{Z}^{(k)}$ is the column vector $(z_1^{(k)},\ldots,z_n^{(k)})^{\tau}$.
5. Estimate the dispersion parameter using the Newton Raphson approach, i.e.,
$$\phi^{(k+1)}=\phi^{(k)}-\Big\{\Big[\frac{\partial^2l(\phi,\boldsymbol{\beta}^{(k+1)})}{\partial\phi^2}\Big]^{-1}\frac{\partial l(\phi,\boldsymbol{\beta}^{(k+1)})}{\partial\phi}\Big\}_{\phi=\phi^{(k)}},$$
where $l(\phi,\boldsymbol{\beta}^{(k+1)})$ is the log-likelihood function with $\boldsymbol{\beta}$ fixed at $\boldsymbol{\beta}^{(k+1)}$.
6. Repeat steps 2–5 until convergence of the parameter estimates.
7. Let $\hat{\boldsymbol{\beta}}$, $\hat{\phi}$ and $\boldsymbol{A}$ be the final estimates after convergence. The variances can be estimated by
$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})=\widetilde{\boldsymbol{B}}\boldsymbol{A}^{-1}\widetilde{\boldsymbol{B}},\qquad\widehat{\text{Var}}(\hat{\phi})=\Big[-\frac{\partial^2l(\phi,\widehat{\boldsymbol{\beta}})}{\partial\phi^2}\Big]^{-1}\Big|_{\phi=\hat{\phi}},$$
where $\widetilde{\boldsymbol{B}}=(\boldsymbol{X}^{\tau}\boldsymbol{W}\boldsymbol{A}\boldsymbol{X})^{-1}\boldsymbol{X}^{\tau}\boldsymbol{W}\boldsymbol{A}$.

Table 12.2 Expression of z_j and a_{jj} in the IRLS algorithm for selected distributions

Distribution	z_j	a_{jj}
$Normal(\mu, \sigma^2)$	y_j	1
$Binomial(n, p)$	$\eta_j + \frac{y_j/n_j - p_j}{p_j(1-p_j)}$	$n_j p_j (1 - p_j)$
$Poisson(\mu)$	$\eta_j + \frac{y_j - \mu_j}{\mu_j}$	μ_j
$NB(\mu, \alpha)$	$\eta_j + \frac{y_j - \mu_j}{\mu_j}$	$\frac{\mu_j r}{\mu_j + r}$

If the model does not involve an overdispersion parameter such as binomial and Poisson, step 5 is skipped. Table 12.2 gives z_j and a_{jj} (iteration index suppressed) for Gaussian, logistic, Poisson and negative binomial geographically weighted regression models. Note that the expressions of z_j and a_{jj} are based on Fisher information, i.e., the negative of the expectation of the second derivative of the log-likelihood with regard to η_j. They can be based on the observed Fisher information (without taking expectation) as well, as da Silva and Rodrigues (2014) did for the negative binomial model.

12.2.5 Colinearity and Remedies

Hereinafter, we assume each model has a total of p rather than $p + 1$ covariates, which may or may not include an intercept. This is because some of the regularized models require or recommend the response variable $\boldsymbol{Y}$ to be standardized which will eliminate the intercept. Typically, GWR fits models with $n \times p$ parameters, using only n observations. This overfitting is constrained by local weight matrices to yield valid inference, but probably at certain price. Wheeler and Tiefelsdorf (2005) took a close look at how the multicolinearity among predictors (covariates) might affect the correlation among GWR-estimated coefficients and their interpretability. The correlation between estimated coefficients is twofold. First, at any given location, the estimated local coefficients are correlated in terms that the estimated covariance between them are nontrivial. Second, and more importantly, the estimated coefficients for a given pair of covariates are correlated across all observation locations.

Wheeler and Tiefelsdorf (2005) first showed the possibility of local coefficient estimates contradicting the global regression results and scientific evidence by analyzing the bladder cancer mortality data from the Atlas of Cancer Mortality from the National Cancer Institute. Bladder cancer mortality was expected to be positively associated with population density (a proxy for urban vs. rural environment) and lung cancer mortality rate (a proxy for smoking prevalence), consistent with the global regression results. However, the GWR showed vast geographic heterogeneity, and the local coefficient estimates were negatively correlated. Counter-intuitive negative association with the outcome variable was found for population density in the West and Northeast and for lung cancer in the Midwest.

They then performed a series of simulation studies and found that the local coefficient estimates could be substantially correlated with each other, even when the corresponding covariates are not correlated globally. (Figs. 12.8and 12.9), adapted from Wheeler and Tiefelsdorf (2005), illustrates the correlation in the second sense. The two covariates, called exogenous variables, are generated for two scenarios according to

$$\text{Scenario A (dashed curve)}: \ x_1 = Evac_3, \quad x_2 = \sin(\theta)Evac_3 + \cos(\theta)Evac_1,$$

$$\text{Scenario B (solid curve)}: \ x_1 = Evac_4, \quad x_2 = \sin(\theta)Evac_4 + \cos(\theta)Evac1,$$

where $Evac_k$, $k = 1, \ldots, n$ are eigenvalues of the spatial adjacency matrix based on all the counties of the Georgia State in the U.S., after certain transformation and re-scaling. These eigenvalues capture orthogonal spatial characteristics. The parameter θ induces correlation between x_1 and x_2 via $\text{corr}(x_1, x_2) = \sin(\theta)$. Both scenarios indicate strong negative associations: the higher the correlation between the covariates, the lower the correlation between the coefficient estimates. Most surprisingly, zero correlation between the covariates is associated with a high level of negative correlation, -0.8, between the coefficient estimates for scenario B.

Wheeler and Tiefelsdorf (2005) speculated that when two covariates are highly positively correlated, the GWR model tends to explain the variation in the outcome by one coefficient, while pushing the other coefficient to the opposite direction. However, this speculation does not explain the strong positive correlation in coefficient estimates when the covariates are highly negatively correlated as seen in Fig. 12.4. In addition, although the negative correlation between coefficient estimates is obvious in their simulations even when the covariates are not correlated (Figs. 12.8 and 12.9 in Wheeler and Tiefelsdorf (2005)), it is not clear whether such negative correlation matters in practice because the coefficient estimates are mostly close to their global true values (statistical significance level was not given). When nontrivial multicolinearity among covariates is present, however, caution and diagnostic efforts need to be taken, like in global regressions, e.g., examining correlation in the coefficient estimates and the sensitivity of coefficient estimates to addition or deletion of other covariates.

Local Linear Estimation

The local linear estimation approach was proposed by Wang, Mei, and Yan (2008) to target reduction of bias in coefficient estimates of GWR, not to directly address the issue of multicolinearity. In particular, they were concerned about the boundary-effect problem, that is, the GWR estimates tend to be more biased at the boundaries than in the interior part of the study area. Nevertheless, multicolinearity could contribute to bias as evidenced by above discussion. Therefore, the local linear estimation approach can potentially alleviate the problem caused by multicolinearity, while being able to reduce bias from other sources. The idea of this approach is based on first-order Taylor expansion of local coefficients with regard to spatial points:

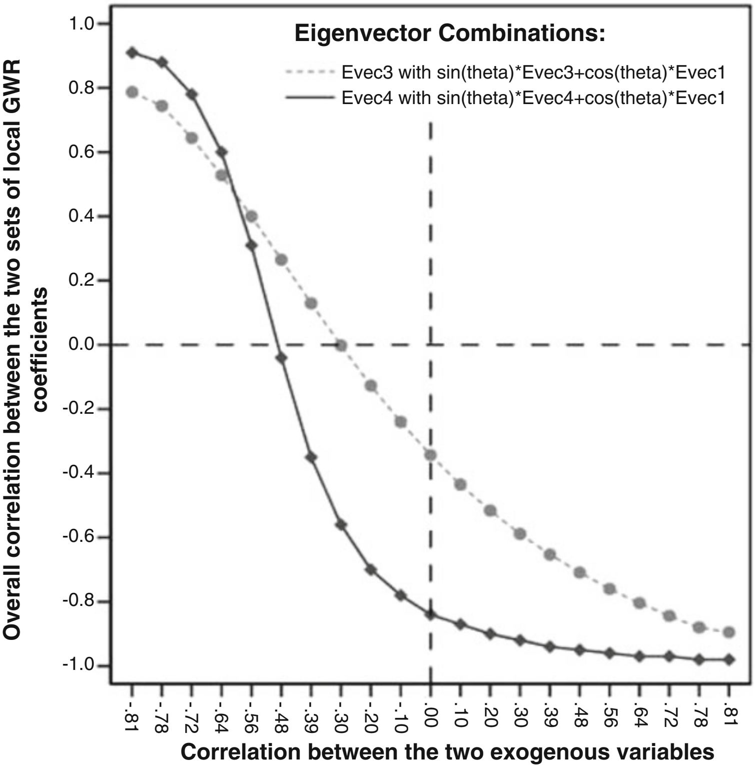

Fig. 12.4 Overall correlation between two coefficient estimates as a function of correlation between the two corresponding covariates, adapted from Wheeler and Tiefelsdorf (2005)

$$\beta_k(u, v) \approx \beta_k(u_i, v_i) + \beta_k^{(u)}(u_i, v_i)(u - u_i) + \beta_k^{(v)}(u_i, v_i)(v - v_i),$$

where $\beta_k^{(u)}(u_i, v_i)$ and $\beta_k^{(u)}(u_i, v_i)$ are partial derivatives of $\beta_k(u, v)$ with regard to u and v, respectively, evaluated at $s_j = (u_i, v_i)$. For each local regression centered at (u_i, v_i), it is no longer necessary to assume the same coefficients as at (u_i, v_i) for all other observation locations. Instead, the objective function becomes

$$\begin{aligned} L_i\big(\boldsymbol{\beta}(s_i)\big) &= \sum_{j=1}^{n} w_j(s_j)\Big\{y_j - \sum_{k=1}^{p} x_{jk}\big[\beta_k(u_i, v_i) + \beta_k^{(u)}(u_i, v_i)(u_j - u_i) \\ &\quad + \beta_k^{(v)}(u_i, v_i)(v_j - v_i)\big]\Big\}^2 \\ &= ||\boldsymbol{W}(s_i)^{1/2}(Y - \boldsymbol{X}\boldsymbol{\beta}(s_i)||^2 \end{aligned} \tag{12.4}$$

The fitting of the model is as usual

$$\hat{\boldsymbol{\beta}}(s_i) = \left(\boldsymbol{X}(s_i)^\tau \boldsymbol{W}(s_i)\boldsymbol{X}(s_i)\right)^{-1}\boldsymbol{X}(s_i)^\tau \boldsymbol{W}(s_i)\boldsymbol{Y},$$

except that the coefficient vector and design matrix are of an extended form:

$$\boldsymbol{\beta}(s_i) = \left(\beta_1(s_i), \beta_1^{(u)}(s_i), \beta_1^{(v)}(s_i), \ldots, \beta_p(s_i), \beta_p^{(u)}(s_i), \beta_p^{(v)}(s_i)\right)^\tau$$

and

$$\boldsymbol{X}(s_i) = \begin{bmatrix} x_{11} & x_{11}(u_1-u_i) & x_{11}(v_1-v_i) & \cdots, & x_{1p} & x_{1p}(u_1-u_i) & x_{1p}(v_1-v_i) \\ x_{21} & x_{21}(u_2-u_i) & x_{21}(v_2-v_i) & \cdots, & x_{2p} & x_{2p}(u_2-u_i) & x_{2p}(v_2-v_i) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n1}(u_n-u_i) & x_{n1}(v_n-v_i) & \cdots, & x_{np} & x_{np}(u_n-u_i) & x_{np}(v_n-v_i) \end{bmatrix}.$$

Note that $\boldsymbol{X}(s_i)$ depends on location s_i, different from the traditional GWR. In their simulation studies, Wang et al. (2008) generated two coefficients as nonlinear continuous functions of the spatial coordinates (Fig.12.5), and the local linear fitting approach (Fig. 12.6b, d) clearly demonstrated its bias reduction utility in comparison to the traditional GWR (Fig. 12.6a, c).

Regularized Fitting

Ridge Regression To further constrain the coefficient estimates which serves both purposes of reducing correlation in coefficient estimates and variable selection, Wheeler (2007) suggested the coupling of GWR with ridge regression (GWRR), i.e., adding a penalty term for the L_2 norm of the coefficients:

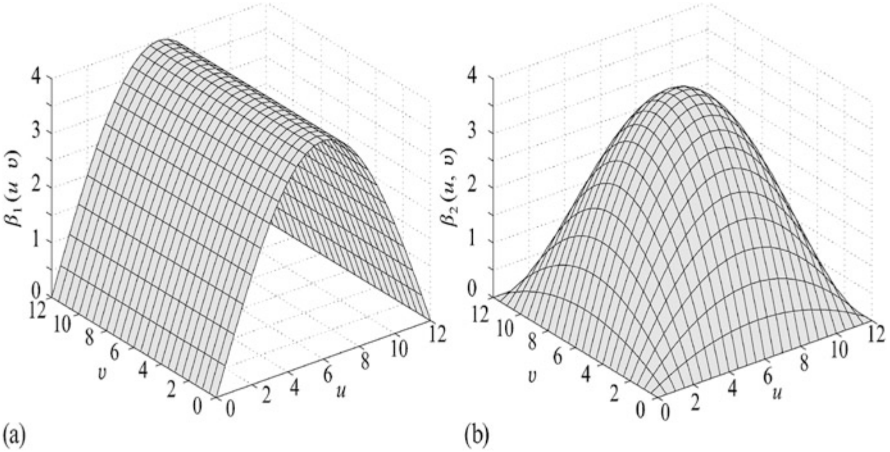

Fig. 12.5 True coefficient functions of geocoordinates in a simulation study in Wang et al. (2008): (**a**) coeficient $\beta_1(u, v)$ for predictor 1; (**b**) coeficient $\beta_2(u, v)$ for predictor 2

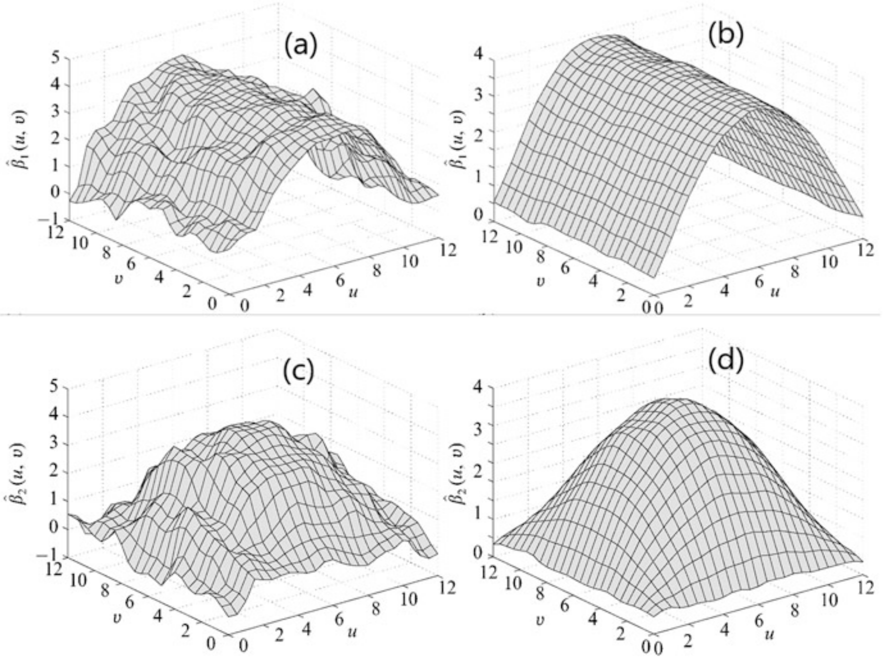

Fig. 12.6 Coefficient functions fitted by GWR (left) and local linear fitting (right) in a simulation study in Wang et al. (2008)

$$\hat{\boldsymbol{\beta}}(s_i) = \arg\min_{\boldsymbol{\beta}} \Big\{ \sum_{j=1}^{n} w_j(s_i) \big[\tilde{y}_j - \tilde{\boldsymbol{X}}_j^{\tau} \boldsymbol{\beta}(s_i) \big]^2 + \lambda \sum_{k=1}^{p} \beta_k^2(s_i) \Big\} \tag{12.5}$$

$$= \big[\tilde{\boldsymbol{X}}^{\tau} \boldsymbol{W}(s_i) \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I} \big]^{-1} \tilde{\boldsymbol{X}}^{\tau} \boldsymbol{W}(s_i) \tilde{\boldsymbol{Y}}, \tag{12.6}$$

where $\tilde{\boldsymbol{X}}_j$ and $\tilde{y}_j$ are standardized covariates and response, $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ are the standardized version of the covariate matrix and response vector, and λ is the global shrinkage parameter. Several standardization schemes were discussed by Wheeler (2007), but here we only introduce the most straightforward scheme:

$$\tilde{x}_{jk} = \frac{x_{jk} - \bar{x}_k\big(\boldsymbol{W}(s_i)\big)}{\hat{\sigma}_k^{(x)}}, \quad \tilde{y}_j = \frac{y_j - \bar{y}\big(\boldsymbol{W}(s_i)\big)}{\hat{\sigma}^{(y)}},$$

where

$$\bar{x}_k\big(\boldsymbol{W}(s_i)\big) = \sum_{i=1}^{n} \sqrt{w_j(s_i)} x_{jk} \Big/ \sum_{i=1}^{n} \sqrt{w_j(s_i)},$$

$$\bar{y}_k\big(\boldsymbol{W}(s_i)\big) = \sum_{i=1}^{n} \sqrt{w_j(s_i)} y_{jk} \Big/ \sum_{i=1}^{n} \sqrt{w_j(s_i)}$$

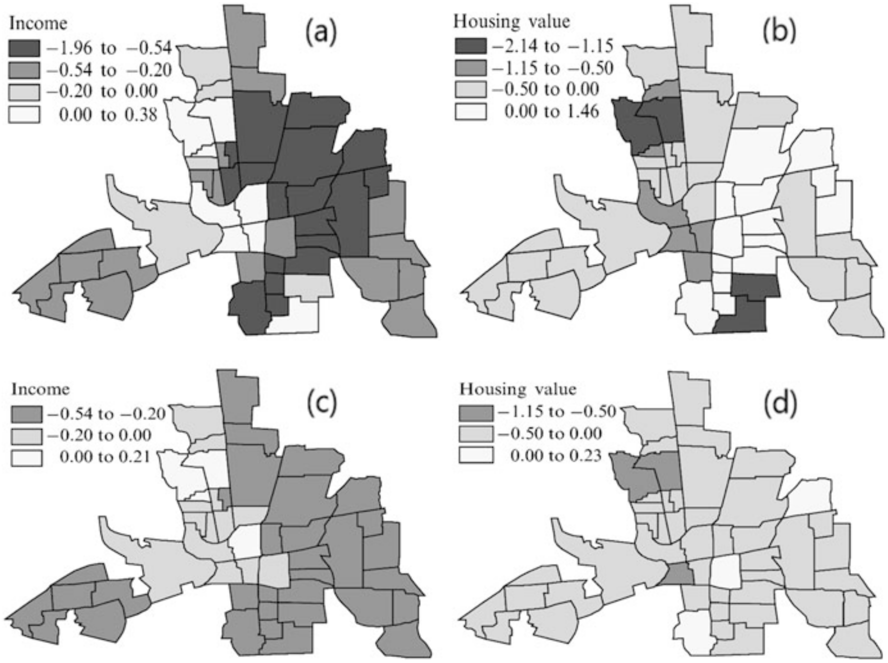

Fig. 12.7 Spatial distributions of coefficient estimates based on GWR and GWRR of crime rate on household income and housing value in Columbus, Ohio: (**a**) Household income, GWR; (**b**) Housing value, GWR; (**c**) Household income, GWRR; and (**d**) Housing value, GWRR (adapted from Wheeler (2007))

are the weighted means, and $\hat{\sigma}_k^{(x)}$ and $\hat{\sigma}^{(y)}$ are the unweighted sample standard deviations, of the kth covariate and the response. Standardization matters for ridge regression, because centering removes the intercept from the model, as ridge regression does not regulate intercept, and rescaling by sample standard deviation ensures fair shrinkage of the regression coefficients. Similar to the bandwidth h, the shrinkage parameter λ can be chosen by cross-validation, and Wheeler (2007) recommended tuning h and λ simultaneously. In his example of fitting crime rate on household income and housing value in Columbus, Ohio, Wheeler (2007) showed appreciable reduction in the correlation between coefficient estimates and more reasonable local coefficient estimates (Fig. 12.7) provided by the GWRR as compared to traditional GWR. Indeed, in Fig. 12.7, the negative association between the coefficient estimates by contrasting (c) with (d) for GWRR is much less obvious than that by contrasting (a) with (b). In addition, the counter-intuitive positive association between crime rate and housing value in East Columbus with GWR (b) become largely negative with GWRR (d).

GWGlasso Analogous to the ridge regression, Wang and Li (2017) proposed a method that couples GWR with adaptive group LASSO to identify model structure and select variables, and named it GWGlasso. The shrinkage of classic lasso is reached by using the L_1-norm penalty, $\lambda \sum_{k=1}^{p} |\beta_k(s_i)|$, instead of square of the L_2 norm, $\lambda \sum_{k=1}^{p} \beta_k^2(s_i)$. One major distinction between ridge regression and lasso is that the former tends to allow for small coefficients close to 0 whereas the latter shrinks small coefficients to 0 exactly (Wheeler 2007). Group lasso applies L_2 norm to the each group of coefficients, $\lambda \sum_{g=1}^{G} \sqrt{\sum_{k=1}^{p_g} \beta_k^2(s_i)}$, where G is the number of groups and p_g is the number of coefficients in group g. It reduces to classic lasso when $p_g = 1$ for all groups. Rather than shrinking each coefficient separately, group lasso shrinks the whole group of coefficients to 0, if shrinkage does occur. As a result, group lasso serves better as a model selector. Adaptive group lasso attaches a different tuning parameter to each group, $\sum_{g=1}^{G} \lambda_g \sqrt{\sum_{k=1}^{p_g} \beta_k^2(s_i)}$. Due to the adaptive feature, it is not necessary to remove the intercept by centering the responses and covariates.

To facilitate the description of GWGlasso, the following notation is defined. Let $\boldsymbol{a}$ be a $n \times p$ matrix with its element at ith row and kth column being $a_{ik} = \beta_k(s_i)$. We denote the columns and rows of $\boldsymbol{a}$ by $\boldsymbol{a}_{,k} = (\beta_k(s_1), \ldots, \beta_k(s_n))^\tau$, $k = 1, \ldots, p$, and $\boldsymbol{a}_{i,} = (\beta_1(s_i), \ldots, \beta_p(s_i))^\tau$, $i = 1, \ldots, n$, respectively. Note that both $\boldsymbol{a}_{,k}$ and $\boldsymbol{a}_{i,}$ are column vectors. Let $\boldsymbol{b}$ be a $2n \times p$ matrix with its element at ith row and kth column being

$$b_{ik} = \begin{cases} \beta_k^{(u)}(s_i), & 1 \le i \le n, \\ \beta_k^{(v)}(s_{i-n}), & n < i \le 2n \end{cases}.$$

Similarly, the columns of $\boldsymbol{b}$ are represented by $\boldsymbol{b}_{,k} = \big(\beta_k^{(u)}(s_1), \ldots, \beta_k^{(u)}(s_n), \beta_k^{(v)}(s_1), \ldots, \beta_k^{(v)}(s_n)\big)^\tau$, $k = 1, \ldots, p$. The rows of $\boldsymbol{b}$ are represented by $\boldsymbol{b}_{i,} = \big(\beta_1^{(u)}(s_i), \ldots, \beta_p^{(u)}(s_i)\big)^\tau$ for $1 \le i \le n$, and $\boldsymbol{b}_{i,} = \big(\beta_1^{(v)}(s_1), \ldots, \beta_p^{(v)}(s_i)\big)^\tau$ for $n < i \le 2n$. The objective function to be minimized is

$$L(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{n} L_i\big(\boldsymbol{\beta}(s_i)\big) + 2\lambda \sum_{k=1}^{p} \Big(\omega_{1k}||\boldsymbol{a}_{,k}||_2 + \omega_{2k}||\boldsymbol{b}_{,k}||_2\Big), \tag{12.7}$$

where $L_i\big((\boldsymbol{\beta}(s_i)$ is given in (12.4), $||\boldsymbol{a}_{,k}||_2 = \sqrt{\sum_{i=1}^{n}[\beta_k(s_i)]^2}$, and $||\boldsymbol{b}_{,k}||_2 = \sqrt{\sum_{i=1}^{n}[\beta_k^{(u)}(s_i)]^2 + [\beta_k^{(v)}(s_i)]^2}$. $||\cdot||_2$ is the L_2 norm (also called Euclidean norm). The weights, $\omega_{1k} = \sqrt{n}/||\hat{\boldsymbol{a}}_{,k}^{(0)}||_2$ and $\omega_{2k} = \sqrt{2n}/||\hat{\boldsymbol{b}}_{,k}^{(0)}||_2$, are used to account for the scales of the coefficient groups and turn multiple tuning parameters to a single one, where $\hat{\boldsymbol{a}}_{,k}^{(0)}$ and $\hat{\boldsymbol{b}}_{,k}^{(0)}$ are estimates of $\boldsymbol{a}_{,k}$ and $\boldsymbol{b}_{,k}$ without the penalty (Wang & Li, 2017). The objective function is not differentiable at the origin for the same reason that $f(x) = |x|$ is not. To circumvent this difficulty, a local quadratic approximation can be used (Wang & Li, 2017). For example, suppose the current estimate of $\boldsymbol{a}_{,k}$ is $\hat{\boldsymbol{a}}_{,k}^{(m)}$ at the mth iteration in an iterative evaluation procedure. Based on the first-order Taylor expansion of $f(y) = \sqrt{y}$, the approximation is

$$||\boldsymbol{a}_{,k}||_2 \approx ||\hat{\boldsymbol{a}}_{,k}^{(m)}||_2 + \frac{||\boldsymbol{a}_{,k}||_2{}^2 - ||\hat{\boldsymbol{a}}_{,k}^{(m)}||_2{}^2}{2||\hat{\boldsymbol{a}}_{,k}^{(m)}||_2}.$$

$||\boldsymbol{a}_k||_2{}^2$ is differentiable everywhere. The same approximation applies to $||\boldsymbol{b}_k||$ as well. With such approximations plugged in, the objective function becomes

$$L(\boldsymbol{a}, \boldsymbol{b}) \propto \sum_{i=1}^{n} \Big[L_i\big(\boldsymbol{\beta}(s_i)\big) + \lambda \big(\boldsymbol{a}_{i,}^{\tau} \boldsymbol{D}_1^{(m)} \boldsymbol{a}_{i,} + \boldsymbol{b}_{i,}^{\tau} \boldsymbol{D}_2^{(m)} \boldsymbol{b}_{i,} + \boldsymbol{b}_{n+i,}^{\tau} \boldsymbol{D}_2^{(m)} \boldsymbol{b}_{n+i,}\big) \Big],$$

where $\boldsymbol{D}_1^{(m)} = \text{diag}\big(\omega_{11}/||\boldsymbol{a}_{,1}^{(m)}||_2, \ldots, \omega_{1p}/||\boldsymbol{a}_{,p}^{(m)}||_2\big)$ and $\boldsymbol{D}_2^{(m)} = \text{diag}$ $\big(\omega_{21p}/||\boldsymbol{b}_{,1}^{(m)}||_2, \ldots, \omega_{2p}/||\boldsymbol{b}_{,p}^{(m)}||_2\big)$.

Depending on whether the coefficient groups are shrunk to 0 or not, the model is naturally structured:

- When $\hat{\boldsymbol{a}}_{,k} = \hat{\boldsymbol{b}}_{,k} = \mathbf{0}$, then we conclude $\beta_k(s_i) = 0$ for all i, i.e., the kth covariate is not influential.
- When $\hat{\boldsymbol{b}}_{,k} = \mathbf{0}$ but $\hat{\boldsymbol{a}}_{,k} \neq \mathbf{0}$, then $\beta_k(s_i) = \beta_k$ for all i, i.e., there is no spatial heterogeneity in the effect of the kth covariate, and β_k is estimated by $\frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_k(s_i)$, where $\hat{\beta}_k(s_i)$'s are elements of $\hat{\boldsymbol{a}}, k$.
- When $\hat{\boldsymbol{a}}_{,k} \neq \mathbf{0}$ and $\hat{\boldsymbol{b}}_{,k} \neq \mathbf{0}$, then there is spatial heterogeneity in the effect of the kth covariate, and $\beta_k(s_i)$ is estimated by $\hat{\beta}_k(s_i)$.

To alleviate computational overhead, Wang and Li (2017) suggested the optimal bandwidth, $h^\star$, be chosen using cross-validation based on the unpenalized local linear estimation approach. After fixing the bandwidth, the shrinkage parameter λ can be chosen to minimize the Bayesian information criteria,

$$BIC = \log\Big[\frac{1}{n^2}\sum_{i=1}^{n} L_i\big(\hat{\boldsymbol{\beta}}_\lambda(s_i)\big)\Big] + df_\lambda \frac{\log(nh)}{nh} + (p - df_\lambda)\frac{\log(n)}{n},$$

where df_λ is the number of spatially-varying coefficients, $L_i\big(\hat{\boldsymbol{\beta}}_\lambda(s_i)\big)$ is given in (12.4), and $\hat{\boldsymbol{\beta}}_\lambda(s_i)$, $i = 1, \ldots, n$, are the estimated coefficients in $\boldsymbol{a}$ and $\boldsymbol{b}$ under a given value of λ.

While GWGlasso is able to identify model structures, it is often desired to find sparse local coefficients within groups. Such desire entails the need for a method between lasso and group lasso, where the sparse group lasso fits (Simon et al. 2001). A possible extension of (12.7) is to incorporate sparse group lasso into GWGlasso is

$$\begin{aligned} L(\boldsymbol{a}, \boldsymbol{b}) = &\sum_{i=1}^{n} L_i\big(\boldsymbol{\beta}(s_i)\big) + 2\lambda_1 \sum_{k=1}^{p}\Big(\omega_{1k}||\boldsymbol{a}_{,k}||_2 + \omega_{2k}||\boldsymbol{b}_{,k}||_2\Big) \\ &+ \lambda_2 \sum_{k=1}^{p}\Big(\omega_{1k}^{\star}||\boldsymbol{a}_{,k}||_1 + \omega_{2k}^{\star}||\boldsymbol{b}_{,k}||_1\Big), \end{aligned}$$

where $||\boldsymbol{a}_{,k}||_1 = \sum_{i=1}^{n} |a_{ik}|$ is the L_1 norm. Whether this is a valid extension, as well as what form should $\omega_{1k}^{\star}$ and $\omega_{2k}^{\star}$ take, are open questions.

GW Elastic Net lasso is able to identify influential covariates to form a parsimonious model, sacrificing the predictive power of the model to some level. On the other hand, ridge regression is able to reduce the impact of multicolinearity without compromising predictive performance, but it may retain non-predictive covariates. To find a balance between parsimony and predictive performance, a geographically weighted elastic net (GWEN) blends lasso and ridge penalties (Li & Lam, 2018):

$$L_i\big(\boldsymbol{\beta}(s_i)\big) = \sum_{j=1}^{n} w_j(s_i)\big[y_j - \beta_0(s_i) - \sum_{k=1}^{p} \beta_k(s_i)x_{jk}\big]^2 + \lambda \sum_{k=1}^{p} \big[(1-\alpha)\beta_k^2(s_i) + \alpha|\beta_k(s_i)|\big]$$

Similar to (12.5), one can center and rescale covariates and responses to remove the intercept:

$$L_i\big(\boldsymbol{\beta}(s_i)\big) = \sum_{j=1}^{n} w_j(s_i)\big[\tilde{y}_j - \tilde{\boldsymbol{X}}_j^{\tau}\boldsymbol{\beta}(s_i)\big]^2 + \lambda \sum_{k=1}^{p} \big[(1-\alpha)\beta_k^2(s_i) + \alpha|\beta_k(s_i)|\big]$$

where $\tilde{\boldsymbol{X}}_j$ and $\tilde{y}_j$, $j = 1, \ldots, n$, are standardized covariates and responses.

In an analysis of population size change from 2000 to 2010 regressed on thirty-five socio-environmental variables in the Lower Mississippi River Basin, Li and Lam (2018) compared the results between several GWR models as shown in Table 12.3. In this table, root MSE measures goodness of fit, mean VIF measures multicolinearity among covariates, and global Moran's I measures spatial assesses spatial autocorrelation among residuals. For all these quantities, the lower the better. GWEN resembles GWR-lasso in model parsimony and explaining spatial correlation, and is comparable to GWR-Ridge in terms of goodness-of-fit and multicolinearity. As expected, GWEN offers a reasonable trade-off between parsimony and goodness-of-fit.

Table 12.3 Comparison between various GWR models

Metrics	Classic GWR	GWR-Ridge	GWR-lasso	GWEN
Root MSE	0.42	0.51	0.59	0.55
Mean VIF	252.64	4.92	2.78	4.04
Average # of selected covariates	35.00	35.00	10.18	10.18
Global Moran's I	0.009	0.019	0.046	0.045

This table is adapted from Li and Lam (2018)

12.3 Software and Case Study

Currently, there are multiple choices of software tools implementing various versions of GWR. GWR4 is a standalone software package dedicated to GWR, implementing GWR for three distribution families (Gaussian, Poisson and Logistic), with useful features such as variable selection and simultaneously considering global and local regression coefficients (Nakaya, Fotheringham, Charlton, & Brunsdon, 2009). GWR has also been integrated into several commonly used GIS or spatial statistics software tools such as ArcGIS (Environmental Systems Resource Institute 2013), SpaceStat (BioMedware 2011), and SAM (Rangel, Diniz-Filho, & Bini, 2010). Both ArcGIS and SpaceStat are commercial software packages that are not free. We introduce three GWR-related R packages, spgwr, gwrr and GWmodel using a case study with a real data set, mainly for the free availability and versatility of R as a software platform for data manipulation, data presentation, and statistical analysis. To facilitate our description, we refer to the three regularized models (GW ridge regression, GWR with lasso and GWR with locally compensated ridge) available in packages gwrr and GWmodel as GWR-Ridge, GWR-LASSO and GWR-LCR. A brief comparison of features of the three packages is shown in Table 12.4. Clearly, these packages have some non-overlapping features, and it could be fruitful to use them in combination.

12.3.1 Data

For the case study we use the hand, foot and mouth disease (HFMD) surveillance data during 2009 in China, provided by the courtesy of Chinese Center for Disease Control and Prevention (CCDC). Epidemiological description and statistical analyses of these data can be found elsewhere (Tang, Yang, Yu, Liao, & Bliznyuk, 2019; Wang et al. 2011). Briefly, HFMD is a disease mainly among children under 6 years of age caused by a spectrum of enteroviruses. In China, mandated reporting of this disease was initiated by a large outbreak in 2008. The reporting became well established and relatively complete since 2009. As in Wang et al. (2011), we will analyze the data at the prefecture level, which is an administrative level between province and county. The dependent variable (outcome) of interest is disease incidence, i.e., number of cases per year and 100,000 people, after log-transformation. The independent variables (predictors) are log-transformed population density, (log-popden) temperature (temp), relative humidity (rh), and wind speed (ws), and all three climatic predictors are taken as annual averages. All predictors have been standardized to have 0 for sample means and 1 for sample variance. The spatial distributions of case numbers and incidences are shown in Fig. 12.8. High disease incidences are found in Guangxi, Guang and northern Hunan provinces. Moron's I statistic based on Euclidean distances as the weights is -0.056 with a p-value <0.001, indicating a pattern of more spatially dispersed than expected.

Table 12.4 Comparison of features between various GWR-related R packages, partially adapted from Gollini et al. (2015)

Model	Function	Option	Packages			
			spgwr	gwrr	GWmodel	McSpatial
All	Kernel[a]	Gaussian	Y	Y	Y	Y
		Bi-square	Y		Y	Y
		Tri-cube			Y	Y
		Exponential		Y	Y	
Classic	Bandwidth	CV	Y	Y	Y	Y
GWR	Selection	AICc	Y		Y	
	Adaptive					
	Bandwidth		Y		Y	
	Colinearity	VIF			Y	
	Diagnosis	VDP		Y	Y	
		Condition number		Y	Y	
	Colinearity	Global ridge		Y	Y[b]	
	Solution	Local ridge			Y[b]	
		LASSO		Y		
	Test global vs. local coefficients		Y		Y	
Generalized	Distribution[c]	Poisson	Y		Y	
GWR		Binomial	Y		Y	Y
		Multinomial				Y
		Quasi-Poisson	Y			
	Bandwidth	CV	Y		Y	Y
	Selection	AICc			Y	
	Adaptive					
	Bandwidth		Y		Y	

[a] GWmodel and McSpatial provide additional kernel functions
[b] GWmodel requires users to provide ridge parameters, i.e., no optimization
[c] spgwr offers additional distributions as specified in glm() function

12.3.2 *Data Analysis with R Packages*

We use R 3.3.0 for all the analyses (R Core Team 2013). The following packages are to be loaded.

```
Packages <- c(”maptools”, ”shapefiles”, ”RColorBrewer”, ”rgdal”, ”spdep”,
            ”sp”, ”spgwr”, ”gwrr”, ”GWmodel”, ”lattice”, ”ape”)
invisible(lapply(Packages, library, character.only = TRUE))
```

To load an individual package, e.g., gwrr, simply use “library(gwrr)”. To load the data, use either of the following options:

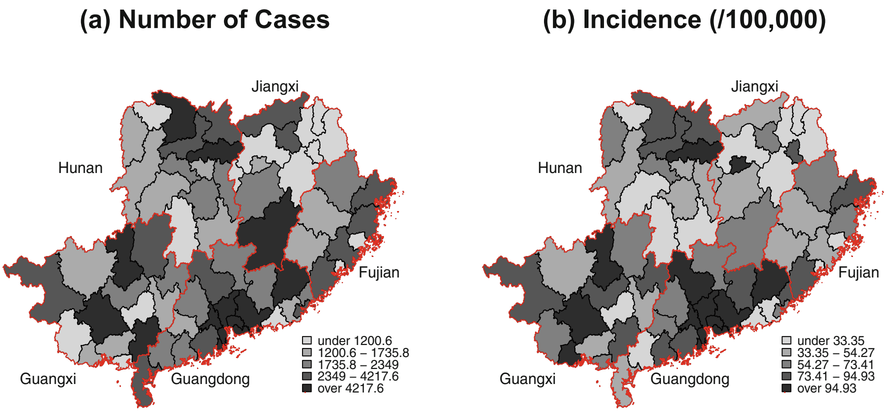

Fig. 12.8 Spatial distributions of (**a**) numbers of cases and (**b**) annual incidences at the prefecture level in five southern provinces of China during the 2009 epidemic of the hand, foot and mouth disease

```
# Option 1
setwd(path); load(case_study_gwr.RData)
# Option 2
load(paste(path, ”case_study_gwr.RData”, sep=’’))
```

where *path* is the directory where your put the data, e.g., “C:/gwr/data/”. Three data sets will be loaded, two spatial data objects (of type SpatialPolygonsDataFrame in R spatial statistics packages), named “south5.shp” and “`south5_province_line`”, and one usual data frame, named “my.data”, which contains the same data as in south5.shp but no polygon structures. “`south5_province_line`” is only for drawing provincial boundaries.

To assess Moran’s I for disease incidence, use the code

```
centroids <- coordinates(south5.shp)
my.dMat <- gw.dist(centroids)
Moran.I(my.data$log_incidence, my.dMat)
```

We first examine the relationship between the outcome and the predictors to see if nonlinear trend exists. There appears to moderate levels of nonlinear trends for population density, temperature and wind speed (Fig. 12.9). However, for illustrative purpose, we move ahead with only linear terms. To select a bandwidth via cross-validation using the Gaussian kernel and to fit a classic GWR with the chosen bandwidth, one can use the following functions from the spgwr package:

```
bw.gwr<-gwr.sel(log_incidence~log_popden+rh+temp+ws, data=south5.shp,
                        adapt=FALSE, gweight=gwr.Gauss, verbose=TRUE)
fit.gwr<-gwr(log_incidence~log_popden+rh+temp+ws, data=south5.shp,
                 bandwidth=bw.gwr, gweight=gwr.Gauss, hatmatrix=TRUE)
```

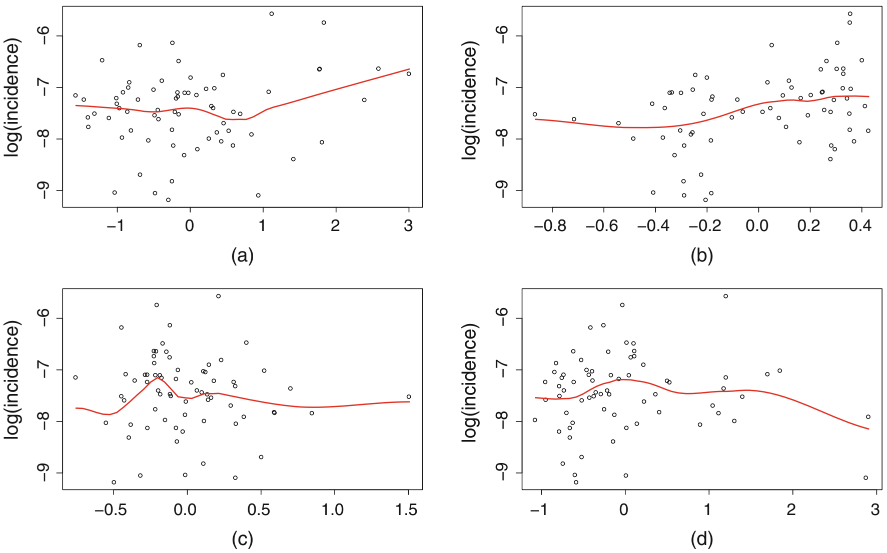

Fig. 12.9 Scatter plots of the outcome with (**a**) log(population density), (**b**) temperature, (**c**) relative humidity and (**d**) wind speed. The red solid curves represent loess smoothing

The option adapt=FALSE tells gwr.sel() that a fixed rather than adaptive bandwidth is desired. If adapt=TRUE, then a fraction is returned. For example, if the optimal fraction found by gwr.sel() is 0.8, then the nearest 80% of all data points will be included for analysis at each location. The option hatmatrix=TRUE specifies that the hat matrix is to be included in the returned object fit.gwr, in addition to other model outputs. The spatial distributions of the estimated coefficients, as shown in Fig. 12.9, are produced by

```
par(mfrow=c(2,2))
do.map(fit.gwr$SDF, fit.gwr$SDF$log_popden, '(a) log(Pop Density)')
do.map(fit.gwr$SDF, fit.gwr$SDF$temp, '(b) Temperature')
do.map(fit.gwr$SDF, fit.gwr$SDF$rh, '(c) Rel. Humidity')
do.map(fit.gwr$SDF, fit.gwr$SDF$ws, '(d) Wind Speed')
```

and the function *do.map()* can be found in the accompanying online code with this book. Figure 12.10 appears to indicate geographical clustering of different levels of regression coefficients for each predictor, particularly relative humidity and wind speed; nevertheless, such geographic heterogeneity may not be of practical importance. For example, the absolute value of the coefficient of variation is 0.23 for the estimated coefficients for temperature, much lower than 3.33 and 3.35 for relative humidity and wind speed, suggesting a lesser degree of spatial heterogeneity for temperature. The coefficient of variation is simply the ratio of standard deviation to mean and can be computed by `sd(fit.gwr$SDF$temp)/mean(fit.gwr$SDF$temp)`. Formal statistical tests can be used to test for spatial nonstationary in all or specific predictors:

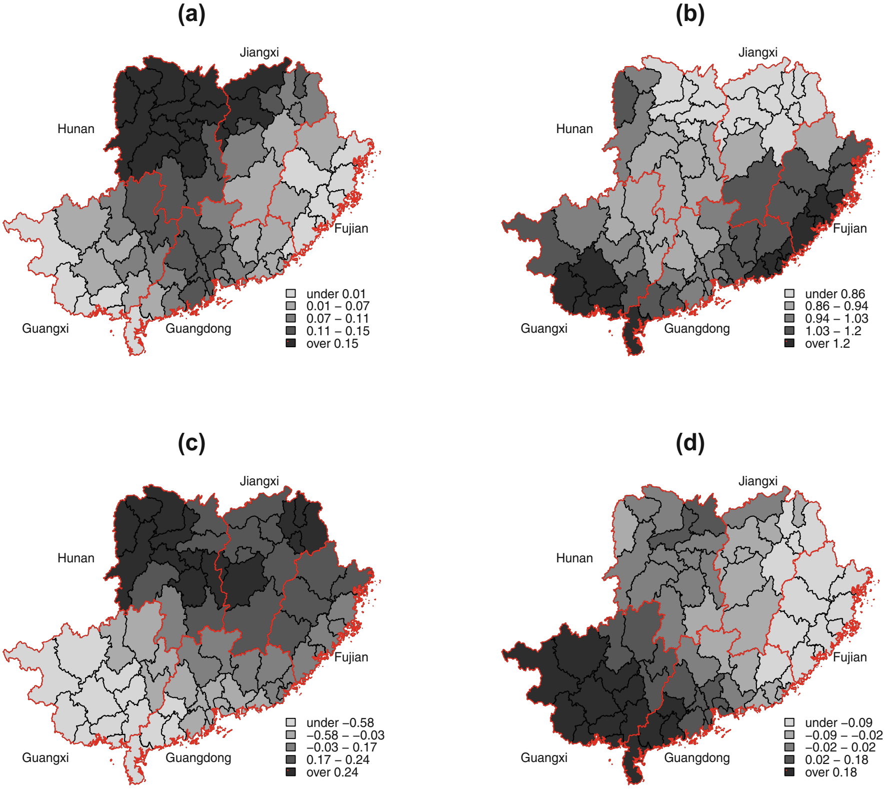

Fig. 12.10 Spatial distributions of local coefficients estimated by classic GWR with a fixed Gaussian kernel for (**a**) log(population density), (**b**) temperature, (**c**) relative humidity, and (**d**) wind speed

```
> BFC02.gwr.test(fit.gwr)$p.value
0.1932179
> LMZ.F1GWR.test(fit.gwr)$p.value
0.4162748
> LMZ.F2GWR.test(fit.gwr)$p.value
0.2315591
> LMZ.F3GWR.test(fit.gwr)
Leung et al. (2000) F(3) test

                F statistic Numerator d.f. Denominator d.f.     Pr(>)
(Intercept)       0.87779       34.29761           58.356   0.65434
log_popden        0.96944       23.37794           58.356   0.51567
rh                4.21961       31.52168           58.356 1.008e-06 ***
temp              0.57866       22.33473           58.356   0.92288
ws                1.71297       23.17358           58.356   0.05019 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

BFC02.gwr refers to the resampling-based test (Brunsdon et al. 1996; Fotheringham et al. 2002), and LMZ.F1GWR and LMZ.F2GWR refer to the asymptotic test (Leung et al. 2000), for the global null (spatial stationarity in all coefficients). LMZ.F3GWR refers to the asymptotic tests for specific predictors (Leung et al. 2000). Spatial nonstationarity seems to be significant for relative humidity and marginally significant for wind speed. However, none of the global tests are significant. Statistically speaking, one should take the tests for specific predictors seriously only if the global test is significant.

To assess colinearity in predictors and its impact on GWR, we look at local weighted correlations among predictors' values (weights determined by the same distance matrix as in GWR) and correlations among GWR-estimated local coefficients of these predictors.

```
> var.list <- c('log_popden', 'temp', 'rh', 'ws')
> cov.gwr<-gw.cov(fit.gwr$SDF, vars=var.list,
+                 bw=bw.gwr, gweight=gwr.Gauss)
> mean.cor.gwr<-with(cov.gwr$SDF@data, apply(cbind(cor.log_popden.temp.,
+                    cor.log_popden.rh., cor.log_popden.ws., cor.temp.rh.,
+                     cor.temp.ws., cor.rh.ws.), 2, mean))
> b<-matrix(0, nrow=4, ncol=4, dimnames=list(var.list, var.list))
> b[lower.tri(b, diag=FALSE)] <- mean.cor.gwr
> t(b)
           log_popden       temp         rh          ws
log_popden          0 -0.6352849  0.3803019  0.01462292
temp                0  0.0000000 -0.4499036  0.16462085
rh                  0  0.0000000  0.0000000 -0.81590363
ws                  0  0.0000000  0.0000000  0.00000000
> with(fit.gwr$SDF@data, cor(cbind(log_popden, temp, rh, ws)))
           log_popden       temp         rh         ws
log_popden  1.0000000 -0.5994083  0.4320501 -0.1735663
temp       -0.5994083  1.0000000 -0.5898919  0.4249700
rh          0.4320501 -0.5898919  1.0000000 -0.9482771
ws         -0.1735663  0.4249700 -0.9482771  1.0000000
```

The local weighted correlations flagged alarming colinearity between population density and temperature, −0.63, as well as between relative humidity and wind speed, −0.82. The impact of colinearity of the latter pair on GWR is confirmed by the high negative correlation, −0.95, in local coefficients between relative humidity and wind speed. The colinearity associated with relative humidity and wind speed are visualized as the scatter plot of the estimated local coefficients (Fig. 12.11a) and the mapping of estimated local correlation based on Fisher information in the estimates of local coefficients (Fig. 12.11b). Unfortunately, the package spgwr does not provide means of addressing colinearity. It is probably worth mentioning that, for colinearity in the GWR setting, it could be misleading to only check the global correlations among the predictors, as the following code shows.

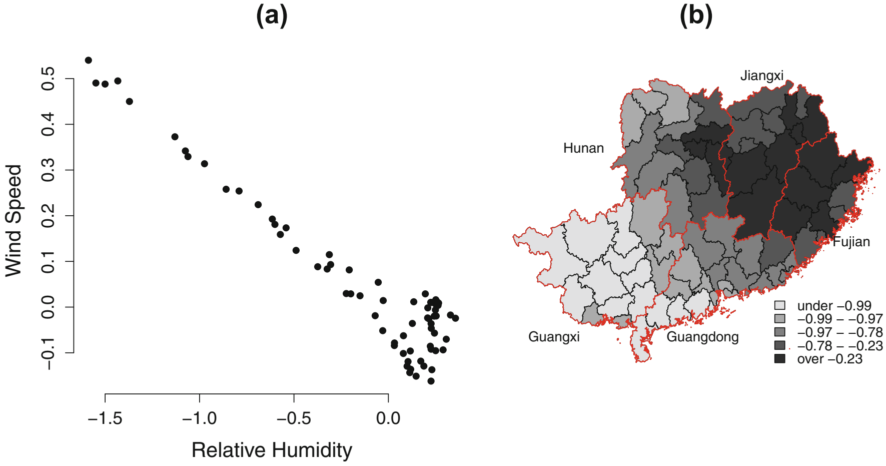

Fig. 12.11 Colinearity between relative humidity and wind speed is examined by (**a**) scatter plot of estimated local coefficients and (**b**) spatial distribution of estimated local correlation (using Fisher information) in the estimates of local coefficients

```
> with(my.data, cor(cbind(log_popden, temp, rh, ws)))
            log_popden        temp          rh         ws
log_popden  1.00000000  0.27318920 -0.05270545 0.44078059
temp        0.27318920  1.00000000 -0.11509446 0.04691729
rh         -0.05270545 -0.11509446  1.00000000 0.34480512
ws          0.44078059  0.04691729  0.34480512 1.00000000
```

The package spgwr does provide generalized GWR models, and the distribution families are as many wide as specified by glm(). The Poisson family is appropriate if we model the number of cases directly with population size as an offset.

```
bw.ggwr<-ggwr.sel(n_cases~log_popden+rh+temp+ws+offset(log(pop)),
                  data = south5.shp, family=poisson, adapt = FALSE,
                  gweight = gwr.Gauss, verbose = TRUE)
fit.ggwr<-ggwr(n_cases~log_popden+rh+temp+ws+offset(log(pop)),
               data = south5.shp, family=poisson,
               bandwidth=bw.ggwr, gweight=gwr.Gauss)
```

This Poisson GWR yields largely similar spatial patterns of local coefficient estimates (Fig. 12.12) as compared to the classic GWR (Fig. 12.10), except that large coefficients for temperature became more clustered in the southwest corner of the study region. However, spgwr does not offer predicted values of generalized models (mean rate in the Poisson case) or any measure for goodness-of-fit to the data.

Before exploring the ridge regression and LASSO facilities in the gwrr package, we first introduce the function vdp.gwr() in this package for diagnosing colinearity. Unlike spgwr, functions in gwrr do not handle spatial data structure (such as

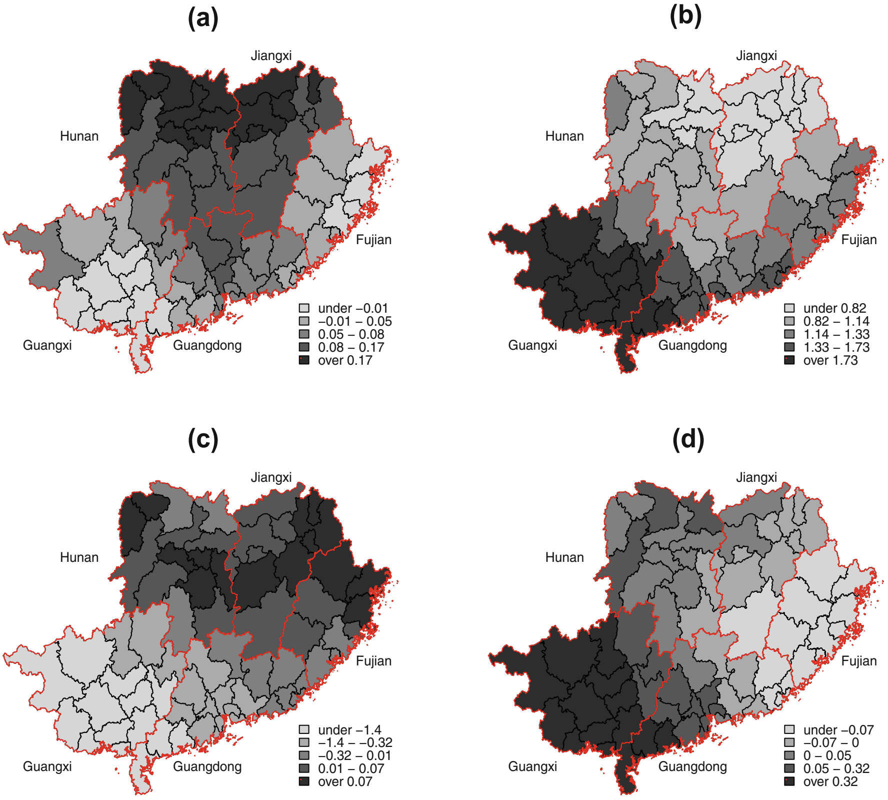

Fig. 12.12 Spatial distributions of local coefficients estimated by Poisson GWR with a fixed Gaussian kernel for (**a**) log(population density), (**b**) temperature, (**c**) relative humidity, and (**d**) wind speed

south5.shp used above) directly; instead, they require usual data frames with geo-coordinates as a separate argument.

```
bw.gwr.exp<-gwr.bw.est(log_incidence~log_popden+rh+temp+ws, data=my.data,
                       locs=centroids, kernel="exp")
vdp.gwr<-gwr.vdp(log_incidence~log_popden+rh+temp+ws, data=my.data,
                 locs=centroids, kernel="exp", phi=bw.gwr.exp$phi,
                 sel.ci = 30, sel.vdp = 0.5)
```

The first command optimizes the bandwidth via cross-validation, using the exponential kernel. Only two kernels are available in gwrr, Gaussian ("gauss") or exponential ("exp"). We got an numeric error by using the Gaussian kernel on the HFMD data, which leaves exponential as the only feasible option. Function gwr.bw.est() returns a structure rather than a value, and the selected bandwidth can be retrieved with `bw.gwr.exp$phi`. Function gwr.vdp() outputs both variance decomposition proportions (VDP) and condition numbers (CN), both derived from the weighted covariance matrix $\sqrt{\boldsymbol{W}(s_i)}\boldsymbol{X}$. According to Wheeler (2007), let the singular value decomposition of $\sqrt{\boldsymbol{W}(s_i)}\boldsymbol{X}$ be $\sqrt{\boldsymbol{W}(s_i)}\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\tau}$, where $\boldsymbol{U}$ is a $n \times (p+1)$ matrix with orthogonal columns, $\boldsymbol{D}$ is a $(p+1) \times (p+1)$ diagonal matrix, and $\boldsymbol{V}$ is a $(p+1) \times (p+1)$ matrix with orthogonal columns. The diagonal elements $(d_1, \ldots, d_{p+1})$ of $\boldsymbol{D}$ are called singular values, and the columns of $\boldsymbol{V}$ are called (right) singular vectors. CNs are defined as $CN_j = \max(d_1, \ldots, d_{p+1})/d_j$ for $j = 1, \ldots, p+1$, and VDPs are defined as a matrix $\boldsymbol{\Phi} = \{\phi_{ij} = \frac{v_{ij}^2}{d_j^2}\}_{(p+1)\times(p+1)}$ rescaled by its row sums, i.e., $VDP_{ij} = \phi_{ij}/\phi_i$, where $\phi_i = \sum_{k=1}^{p+1} \phi_{ik}$. The rationale behind VDP is $\mathrm{Var}(\hat{\beta}_k(s_i)) = \sigma^2\phi_i$. However, function gwr.vdp() returns only a single value rather than a vector for CN and a vector rather than a matrix for VDP. After some investigation (see the online R code associated with this book), we found that the returned CN is the ratio of the largest to the smallest singular value, and the returned VDP vector is the column of the matrix $\{VDP_{ij}\}$ associated with the smallest singular value. This CN definition is analogous to (not exactly the same as) the one introduced in Gollini, Lu, Charlton, Brunsdon, and Harris (2015) for the package GWmodel. CN values > 30 or VDPs > 0.5 are thought to flag potential issues of colinearity, and that is why 30 and 0.5 are used for threshold options sel.ci and sel.vdp in the above code to indicate which locations have alarming RNs and VDPs. CN is also called condition index in Wheeler (2007), which explains the naming of the option sel.ci.

```
> summary(vdp.gwr$condition)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.831   2.344   2.581   2.616   2.787   3.978
> apply(vdp.gwr$vdp, 2, summary)
                [,1]    [,2]      [,3]      [,4]    [,5]
Min.       6.206e-05 0.01967 2.764e-05 0.0001593 0.02339
1st Qu.    1.260e-02 0.23340 4.909e-02 0.3273000 0.16460
Median     2.410e-01 0.35770 1.298e-01 0.6547000 0.31700
Mean       2.994e-01 0.39300 2.208e-01 0.5513000 0.39120
3rd Qu.    5.281e-01 0.52170 4.188e-01 0.8028000 0.66820
Max.       8.384e-01 0.84120 6.724e-01 0.8949000 0.86160
```

For the HFMD data, none of the CNs are over the threshold, but a substantial amount of VDPs are exceeding 0.5, suggesting that a certain level of colinearity. The following code fits classic GWR and GWRs with ridge and LASSO penalties using functions in the gwrr package, where both bandwidth and shrinkage parameter are automatically chosen via cross-validation.

```
fit2.gwr<-gwr.est(log_incidence~log_popden+rh+temp+ws, data=my.data,
                    locs=centroids, kernel=”exp”, bw=bw.gwr.exp$phi)
fit.gwrr<-gwrr.est(log_incidence~log_popden+rh+temp+ws, data=my.data,
                    locs=centroids, kernel=”exp”, bw=TRUE, rd=TRUE)
fit.gwl<-gwl.est(log_incidence~log_popden+rh+temp+ws,
                 data=my.data, locs=centroids, kernel=”exp”)
```

We found that the ridge shrinkage parameter chosen by gwrr.est() for the HFMD data was 0, i.e., no shrinkage at all. For illustration, we fitted a GWR-Ridge model with a bandwidth equal to `bw.gwr.exp$phi` and an rather arbitrary shrinkage parameter of 0.01, and presented all results about GWR-Ridge based on this model.

```
> fit.gwrr<-gwrr.est(log_incidence~log_popden+rh+temp+ws,
+                    data=my.data, locs=centroids, kernel="exp",
+                    bw=bw.gwr.exp$phi, rd=0.01)
>  gof <-rbind(c(fit2.gwr$rsquare, fit.gwrr$rsquare, fit.gwl$rsquare),
+              c(fit2.gwr$RMSE, fit.gwrr$RMSE, fit.gwl$RMSE))
> dimnames(gof)<-list(c('R-square', 'RMSE'),
+                     c('Classic GWR', 'GWR-Ridge', 'GWR-LASSO'))
> gof
          Classic GWR GWR-Ridge GWR-LASSO
R-square    0.4584649 0.5085547 0.6956734
RMSE        0.5431182 0.5173908 0.4071468
```

The approximate R-squares of the regularized GWRs are larger, whereas the root mean square errors (RMSE) are smaller, than those of the classic GWR, indicating better goodness-of-fit to the data for the regularized regressions. Nonetheless, this does not necessarily imply better predictive power on new data. The scales and spatial distributions of local coefficients are shown in Fig. 12.13 for GWR-Ridge and in Fig. 12.14 for GWR-LASSO. Compared to classic GWR (Fig. 12.10), GWR-Ridge yielded more or less similar spatial patterns of coefficients except for temperature for which large coefficients in southern Guangxi province shifted to northern Hunan province. Shrinkage of coefficients towards zero is only noticeable for temperature. GWR-LASSO clearly shrank local coefficients towards 0 more aggressively, and led to more scattered spatial patterns of coefficients than the other two models.

Gollini et al. (2015) proposed a GWR approach with locally compensated ridge (GWR-LCR) parameters which is implemented in the R package GWmodel. However, the statistical presentation in the paper was a little loose, and we were not able to find further technical details elsewhere. As a result, we briefly summarize the rationale here, per our understanding of the paper. In a global regression setting, the standard solution to ridge regression is expressed as $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\tau \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\tau \boldsymbol{Y}$, where λ is the ridge penalty tuning parameter and $\boldsymbol{I}$ is the identity matrix. Let $\theta_1, \ldots, \theta_p$ be the eigenvalues of $\boldsymbol{X}^\tau \boldsymbol{X}$ in decreasing order, i.e., θ_1 and θ_p are the largest and smallest eigenvalues. Eigenvalues returned by R functions usually are also ordered decreasingly. Gollini et al. (2015) defined the condition number as θ_1/θ_p. The eigenvalues of the ridge-adjusted cross-product matrix $\boldsymbol{X}^\tau \boldsymbol{X} + \lambda \boldsymbol{I}$ is $\theta_1 + \lambda, \ldots, \theta_p + \lambda$, and the associated CN is $(\theta_1 + \lambda)/(\theta_p + \lambda)$. To have $CN \leq \kappa$ for some threshold κ, one can choose $\lambda \geq (\theta_1 - \kappa\theta_p)/(\kappa - 1)$. In a GWR-Ridge setting where a local ridge penalty is applied to each location, the local solution

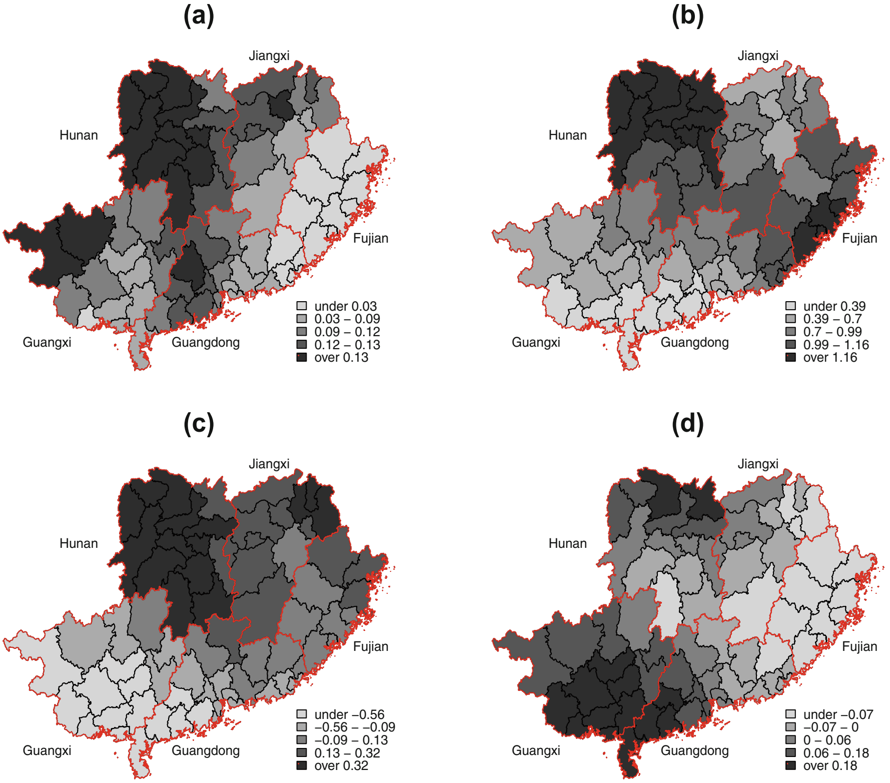

Fig. 12.13 Spatial distributions of local coefficients estimated by GWR-Ridge with an exponential kernel and a fixed bandwidth for (**a**) log(population density), (**b**) temperature, (**c**) relative humidity, and (**d**) wind speed

is given by $\hat{\boldsymbol{\beta}}(s_i) = (\boldsymbol{X}^\tau \boldsymbol{W}(s_i)\boldsymbol{X} + \lambda_i \boldsymbol{I})^{-1}\boldsymbol{X}^\tau \boldsymbol{W}(s_i)\boldsymbol{Y}$. The local ridge parameters λ_i's can be tuned separately to reach acceptable local CNs. This is in contrast to the usual GWR-Ridge where a global λ is tuned to minimize prediction error via cross-validation. Tuning of either fixed or adaptive bandwidth proceeds as usual, using either cross validation or AICc.

We want to point out that, the definition of CN differs from that in Wheeler (2007). The eigenvalues of $\boldsymbol{X}^\tau \boldsymbol{W}(s_i)\boldsymbol{X}$ and the singular values of $\sqrt{\boldsymbol{W}(s_i)}\boldsymbol{X}$ do not yield exactly the same CN because of their relationship $\theta_j = d_j^2$. If the definitions of CN are accurate in the two papers, we suspect that CNs produced by GWmodel are squares of those produced by gwrr. We compared the CNs produced by the two packages in Fig. 12.15, which was generated by the following code:

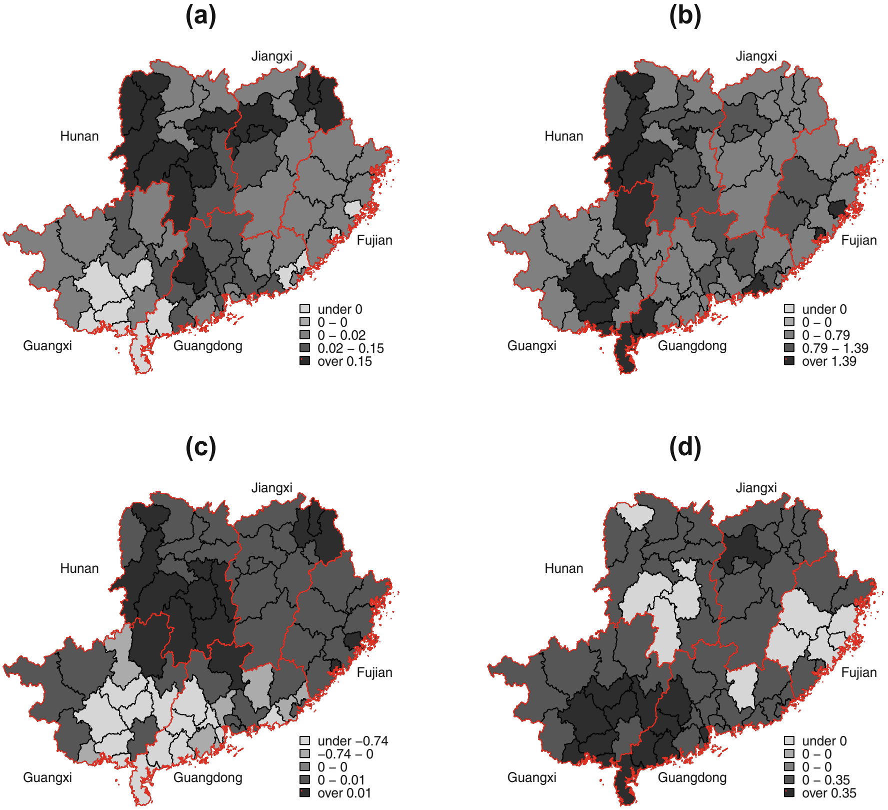

Fig. 12.14 Spatial distributions of local coefficients estimated by GWR-LASSO with an exponential kernel and a fixed bandwidth for (**a**) log(population density), (**b**) temperature, (**c**) relative humidity, and (**d**) wind speed

```
fit0<−gwr.lcr(log_incidence~log_popden+rh+temp+ws,
              data=south5.shp, kernel='exponential',
              bw=bw.gwr$phi, lambda=0, adaptive=FALSE, dMat=my.dMat)
bound<−max(c(fit0$SDF$Local_CN, vdp.gwr$condition))
plot(fit0$SDF$Local_CN, vdp.gwr$condition, xlim=c(0, bound),
     ylim=c(0,bound), xlab='GW-LCR', ylab='GWR', type='p',
     pch=19, main='Local condition indices')
points(sqrt(fit0$SDF$Local_CN), vdp.gwr$condition, pch=1, col='blue')
abline(coef = c(0,1))
```

The original CN outputs of the two packages are very different, shown by the black dots. The square roots of the CNs produced by GWmodel are close to the CNs produced by gwrr, but not exactly. We are not sure about the source of the subtle differences even after taking square root. Both packages recommended the same threshold (30) for flagging potentially problematic CNs.

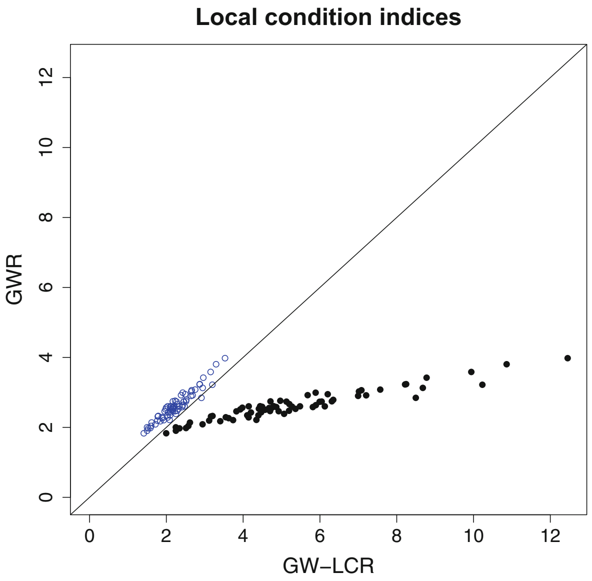

Fig. 12.15 Comparison between condition numbers produced by function gwr.vdp() in gwrr and those by function gwr.lcr() in GWmodel. Black solid dots are scatter plot of the CNs from the two functions. Blue circles are scatter plot after applying square root to the CNs produced by gwr.lcr()

The following code selects bandwidth for and fits the GWR-LCR model, using the exponential kernel and an adaptive bandwidth.

```
bw.lcr <- bw.gwr.lcr(log_incidence~log_popden + rh + temp + ws,
                     data=south5.shp, kernel='exponential',
                     adaptive=TRUE, lambda.adjust=TRUE, cn.thresh=30)
fit.lcr <- gwr.lcr(log_incidence~log_popden + rh + temp + ws,
                   data=south5.shp, kernel='exponential', bw=bw.lcr,
                   adaptive=TRUE, lambda.adjust=TRUE, cn. thresh=30)
```

The AIC, AICc and residual sum of squares obtained from GWR-LCR are 178, 191 and 24.4 respectively, compared to 137, 157 and 24.9 for the classic GWR. The less satisfactory AIC and AICc of GWR-LCR compared to the classic GWR could be due to the weak association of the predictors with the outcome and the substantially more parameters in the GWR-LCR model. The spatial patterns of local predictor coefficients fitted by GWR-LCR are shown in Fig. 12.16. The shrinkage effect of GWR-LCR is clear for relative humidity and wind speed. Interestingly, the spatial patterns are more comparable to the results of the classic GWR (Fig. 12.10) than to those of GWR-Ridge (Fig. 12.13), especially for temperature and population density.

The following code uses Moran's I to test spatial randomness of residuals for all four models. No special spatial patterns were found as all p-values are relatively large. This result suggests that, while the predictors did not show high predictive power overall, they indeed explain the spatial auto-correlation in the HFMD incidences.

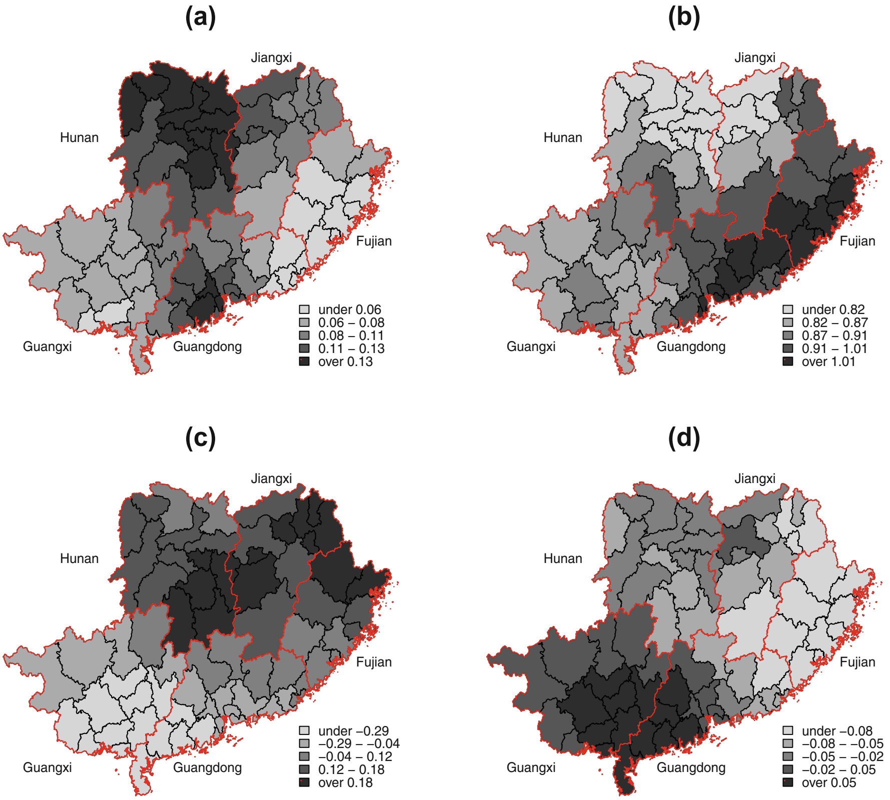

Fig. 12.16 Spatial distributions of local coefficients estimated by GWR-CLR with an exponential kernel and a adaptive bandwidth for (**a**) log(population density), (**b**) temperature, (**c**) relative humidity, and (**d**) wind speed

```
> res.gwr <- fit.gwr$SDF$pred - my.data$log_incidence
> res.gwrr <- fit.gwrr$yhat - my.data$log_incidence
> res.gwl <- fit.gwl$yhat - my.data$log_incidence
> res.lcr <- fit.lcr$SDF$yhat - my.data$log_incidence
> out<-rbind(Moran.I(my.data$log_incidence, my.dMat),
+       Moran.I(res.gwr, my.dMat), Moran.I(res.gwrr, my.dMat),
+       Moran.I(res.gwl, my.dMat), Moran.I(res.lcr, my.dMat))
> rownames(out) <- c('Raw Response', 'Residual:GWR',
+   'Residual:GWR-Ridge', 'Residual:GWR-LASSO', 'Residual:GWR-LCR')
> out
                   observed     expected    sd          p.value
Raw Response       -0.05610592  -0.01470588 0.008402828 8.353727e-07
 Residual:GWR      -0.007730267 -0.01470588 0.008427097 0.4078063
 Residual:GWR-Ridge -0.008066695 -0.01470588 0.00844284  0.4316514
 Residual:GWR-LASSO -0.009189744 -0.01470588 0.008329576 0.5078205
 Residual:GWR-LCR  -0.007598853 -0.01470588 0.008431426 0.3992724
```

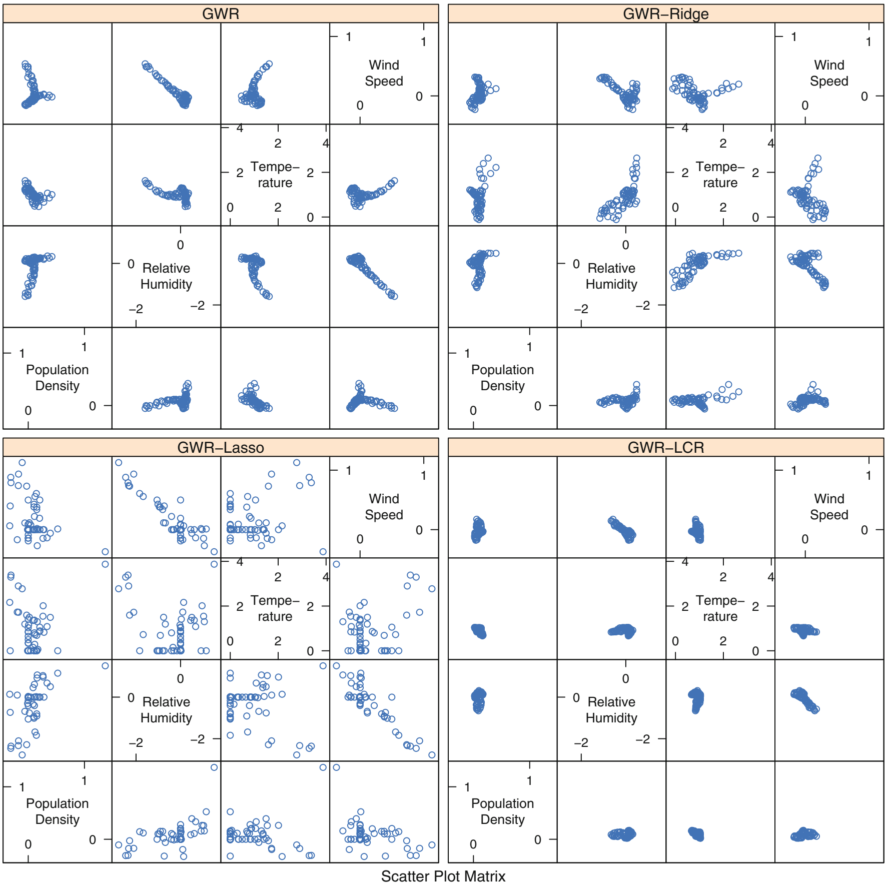

Fig. 12.17 Comparison of correlations among local regression coefficients for classic GWR (upper left), GWR-Ridge (upper right), GWR-LASSO (lower left) and GWR-LCR (lower right)

We further compared at colinearity in local coefficients of all four models, classic GWR, GWR-Ridge, GWR-LASSO and GWR-LCR in Fig. 12.17. The regularized models all reduced colinearity to some extent, and GWR-LASSO did a more satisfactory job. GWR-LCR shrank the ranges of coefficients for all parameters, particularly so for temperature and population density, but such shrinkage is not necessarily towards 0. Although the linear correlation in coefficients between relative humidity and wind speed still remains clear in the regularized models, it is generally not wise to increase the shrinkage level too much, as the price paid for less colinearity is bias in the estimated coefficients (Gollini et al. 2015).

12.3.3 Conclusion

We have introduced the theoretical background of the classic GWR and several regularized versions that impose different constraints on the magnitude of regression coefficients. We also touched the base of some diagnostic statistics such as VDP and condition number. In the case study, we reviewed the use of three R packages dedicated to GWR. We conclude this chapter with comments on future development of GWR and a brief introduction of a Bayesian GWR framework. Thus far, most applications of GWR are limited to cross-sectional data, despite valuable information such as seasonality contained in longitudinal data. There seems to be no theoretical obstacle to the application of GWR to longitudinal data, as long as a reasonable distance can be defined in space-time. However, there is no consensus or guideline on such definitions. One possibility is to introduce an extra tuning parameter for the importance of time relative to space in the construction of distance, and let data guide the choice of this parameter, e.g., via cross-validation. Through the case study, we have seen the difficulties in real data analysis. The existing software packages may not offer the desired tools or output for diagnosis or inference. Documentation of most packages is far from sufficient. For instance, we had to do our own investigation to find out how VDP and condition number are calculated in some packages, and yet did not succeed in such investigation for other packages. Better documentation and literature support will greatly increase the popularity of GWR methods. Finally, there appears to be a gap in the GWR literature about how to handle missing data.

All the methods formally introduced in this chapter are likelihood- or frequentist-oriented. The Bayesian approach is known to be able to incorporate prior knowledge about parameters, which is useful when data are inadequate to support inference about some parameters. Another advantage of the Bayesian approach is its convenience in handling complex likelihood with latent or missing data, when the inference is performed by Markov chain Monte Carlo (MCMC), e.g., using either Gibb's sampler or the Metropolis-Hasting's algorithm (Gelman, Carlin, Stern, & Rubin, 1995; Gilks, Richardson, & Spiegelhalter, 1996). A promising direction that has been researched in the past two decades is the Bayesian spatially varying coefficient (SVC) models (Banerjee, Carlin, & Gelfand, 2004; Finley 2011; Gelfand, Kim, Sirmans, & Banerjee, 2003; Wheeler & Calder, 2007; Wheeler & Waller, 2009). The basic hierarchical structure of the Bayesian SVC model is

$$\begin{aligned}
\boldsymbol{Y} &= \boldsymbol{X}^{\star}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \sigma_{\epsilon}^{2}\boldsymbol{I}) \\
\boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{1}_{n\times 1} \otimes \boldsymbol{\mu}, \boldsymbol{H}(\phi) \otimes \boldsymbol{T}) \\
\boldsymbol{\mu} &\sim \mathcal{N}(\boldsymbol{\eta}, \sigma_{\mu}^{2}\boldsymbol{I}) \\
\boldsymbol{T} &\sim InverseWishart_{\nu}(\boldsymbol{\Omega}^{-1}) \\
\phi &\sim gamma(a_{\phi}, b_{\phi}) \\
\sigma_{\epsilon}^{2} &\sim InverseGamma(a_{\tau}, b_{\tau})
\end{aligned} \tag{12.8}$$

where $\boldsymbol{Y}$ is the outcome vector of length n, $\boldsymbol{\beta}$ is a vector of length $n \times p$ stacking regression coefficients at all locations together, $\boldsymbol{X}^{\star}$ is a $n \times np$ block diagonal matrix with $\boldsymbol{X}_i$ (ith row of $\boldsymbol{X}$) as each diagonal block ad 0 elsewhere, and $\boldsymbol{\epsilon}$ is the vector of *i.i.d.* normal errors. The prior distribution of $\boldsymbol{\beta}$ is the essence of the Bayesian SVC model. The notation $\boldsymbol{A} \otimes \boldsymbol{B}$ denotes the Kronecker product which multiplies every element of matrix $\boldsymbol{A}$ with matrix B and yields a block matrix of dimensions as the products of corresponding dimensions of $\boldsymbol{A}$ and $\boldsymbol{B}$. For example, the mean of $\boldsymbol{\beta}$, $\mathbf{1}_{n\times 1} \otimes \boldsymbol{\mu}$, gives a column vector with n $\boldsymbol{\mu}$'s stacked together, where $\mathbf{1}_{n\times 1}$ is a column vector of n 1's. The covariance matrix of $\boldsymbol{\beta}$ is the Kronecker product of a $n \times n$ correlation matrix $\boldsymbol{H}(\phi)$ and a $p \times p$ covariance matrix $\boldsymbol{T}$. $\boldsymbol{H}(\phi)$ captures spatial correlation between study locations and is assumed to depend only on distance and a decay parameter ϕ, e.g., the (i, j)th element being $h_{ij} = \exp(-d_{ij}/\phi)$, where d_{ij} is the distance between locations i and j. $\boldsymbol{T}$ is the covariance among regression coefficient at any location. This Kronecker product structure ensures $\boldsymbol{\Sigma}$ is positive definite and hence a valid covariance matrix. The last four expressions in (12.8) specify prior distributions of $\boldsymbol{\mu}$, $\boldsymbol{T}$, ϕ and σ_{ϵ} with known hyper-parameters. The inference of model was implemented via MCMC in Wheeler and Calder (2007). Due to the high-dimension nature of GWR (number of coefficients increases with locations), the computational burden of this model can be heavy. There are two R packages, spBayes (Finley 2011; Finley & Banerjee, 2019) and spTDyn (Bakar, Kokic, & Jin, 2016), implementing Bayesian SVC models. In particular, parallel computing via openMP is available in spBayes to expedite computation. As computer engineering continues to advance at a fast pace, we can see that Bayesian approach will become more computationally affordable and popular.

References

Bakar, K., Kokic, P., & Jin, H. (2016). Hierarchical spatially varying coefficient and temporal dynamic process models using spTDyn. *Journal of Statistical Computation and Simulation, 86*, 820–840.

Banerjee, S., Carlin, B., & Gelfand, A. (2004). *Hierarchical modeling and analysis for spatial data*. Boca Raton, FL: Chapman & Hall/CRC.

BioMedware. (2011). *Spacestat user manual, version 2.2*. http://www.biomedware.com/

Brunsdon, C., Fotheringham, A., & Charlton, M. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis, 28*(4), 281–298.

Casetti, E. (1972). Generating models by the expansion method: Applications to geographic research. *Geographical Analysis, 4*, 81–91.

da Silva, A., & Rodrigues, T. (2014). Geographically weighted negative binomial regression–incorporating overdispersion. *Statistics and Computing, 24*, 769–783.

Environmental Systems Resource Institute. (2013). ArcGIS Resource Center, version 10.2. http://www.arcgis.com/

Finley, A. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution, 2*, 143–154.

Finley, A., & Banerjee, S. (2019). Bayesian spatially varying coefficient models in the spBayes R package. arXiv:1903.03028v1. https://arxiv.org/pdf/1903.03028

Fotheringham, A., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. West Sussex, UK: Wiley.

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

Gelfand, A., Kim, H., Sirmans, C., & Banerjee, S. (2003). Spatial modelling with spatially varying coefficient processes. *Journal of American Statistical Association, 98*, 387–396.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (1995). *Bayesian data analysis*. London, UK: Chapman & Hall.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2015). Gwmodel: An r package for exploring spatial heterogeneity using geographically weighted models. *Journal of Statistical Software, 63*(17), 1–50.

Jones, K. (1992). Specifying and estimating multilevel models for geographical research. *Transactions of the Institute of British Geographers, 16*, 148–159.

Leung, Y., Mei, C., & Zhang, W. (2000). Statistical tests for spatial nonstationarity based on the geographically weighted regression model. *Environment and Planning A, 32*, 9–32.

Li, K., & Lam, N. (2018). Geographically weighted elastic net: A variable-selection and modeling method under the spatially nonstationary condition. *Annals of the American Association of Geographers, 108*, 1582–1600.

Mei, C.-L., Wang, N., & Zhang, W.-X. (2006). Testing the importance of the explanatory variables in a mixed geographically weighted regression model. *Environment and Planning A, 38*(14), 587–598.

Nakaya, T., Fotheringham, A., Brunsdon, C., & Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine, 24*, 2695–2717.

Nakaya, T., Fotheringham, A., Charlton, M., & Brunsdon, C. (2009). Semiparametric geographically weighted generalised linear modelling in GWR 4.0. In *10th International Conference on GeoComputation*, Sydney, Australia. Available at http://mural.maynoothuniversity.ie/4846/

Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society, 135*, 370–384.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Rangel, T., Diniz-Filho, J. A. F., & Bini, L. M. (2010). Sam: A comprehensive application for spatial analysis in macroecology. *Ecography, 33*, 46–50.

Shoff, C., & Yang, T.-C. (2012). Spatially varying predictors of teenage birth rates among counties in the United States. *Demographic Research, 27*(14), 377–418.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2001). A sparse-group lasso. *Journal of Computational and Graphical Statistics, 22*(2), 231–245.

Tang, X.-Y., Yang, Y., Yu, H.-J., Liao, Q.-H., & Bliznyuk, N. (2019). A spatio-temporal modeling framework for surveillance data of multiple infectious pathogens with small laboratory validation sets. *Journal of the American Statistical Association, 114*(528), 1561–1573.

Wakefield, J. (2008). Ecologic studies revisited. *Annu Review of Public Health, 29*, 75–90.

Wang, N., Mei, C.-L., & Yan, X.-D. (2008). Local linear estimation of spatially varying coefficient models: An improvement on the geographically weighted regression technique. *Environment and Planning A, 40*, 986–1005.

Wang, W., & Li, D. (2017). Structure identification and variable selection in geographically weighted regression models. *Journal of Statistical Computation and Simulation, 87*(10), 2050–2068.

Wang, Y., Feng, Z., Yang, Y., Self, S., Gao, Y., Wakefield, J., . . . , Yang, W. (2011). Hand, foot and mouth disease in China: Patterns of spread during 2008–2009. *Epidemiology, 22*, 781–792.

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). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems, 7*, 161–187.

Wheeler, D., & Waller, L. (2009). Comparing spatially varying coefficient models: A case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. *Journal of Geographical Systems, 11*, 1–22.

Zhang, C., & Murayama, Y. (2000). Testing local spatial autocorrelation using k-order neighbors. *International Journal of Geographical Information Science, 14*, 681–692.

Zhou, Y., Wang, Q., Yang, M., Gong, Y., Yang, Y., Nie, S. J., . . . , Jiang, Q. (2016). Geographical variations of risk factors associated with HCV infection in drug users in southwestern China. *Epidemiology and Infection, 144*, 1291–1300.