A Geographically Weighed Regression-based Approach for the Change of Support Problem

Daisuke Murakami

Graduate School of Systems and Information Engineering, University of Tsukuba,

1-1-1 Tennodai, Tsukuba-shi, Ibaraki, 305-8573, Japan

muraka51@sk.tsukuba.ac.jp

Morito Tsutsumi

Faculty of Engineering, Information and Systems, University of Tsukuba,

1-1-1 Tennodai, Tsukuba-shi, Ibaraki, 305-8573, Japan

tsutsumi@sk.tsukuba.ac.jp

Differences in spatial supports (shape, size, etc.) among spatial data often complicate spatial data analyses, and spatial support conversions, such as aggregation, disaggregation, and interpolation, are often applied to cope with them. However, in general, spatial support conversions present the following problems: (i) how to convert data accurately? and (ii) how to reduce biases caused by the conversions? These problems have been discussed in the field of geography. On the other hand, recently, they have also been discussed in geostatistics under the framework of the change of support problem (COSP). Sophisticated methods for COSP, or for problems (i) and (ii), have been proposed in geostatistics. However, the discussion of COSP is still only within geostatistics. Thus, this study extends the geographically weighted regression (GWR), a geographical method for capturing spatial heterogeneity, to a method that addresses COSP. The constructed method can be used for both (i) spatial unit (support) conversion and (ii) reducing the biases due to aggregation (conversion). In addition, it gives local trend parameters in arbitrary spatial units. The effectiveness of the method is verified using a simulation study.

JEL codes: C43, C21, R12

Keywords: change of support problem, geographically weighted regression, modifiable areal unit problem, geostatistics

1. Introduction

In accordance with the recent development of geographical information systems (GIS), spatial or spatiotemporal data, including geostatistical data (data observed on a continuous space) and lattice data (data observed on a discrete space), have become more and more diverse. As a result, various datasets have often been applied in their analysis. For example, Dominici et al. (2000) and Lee et al. (2009) analyzed the effects of environmental pollutants on human health using both population data given in specific units (lattice data) and environmental data observed at monitoring stations (geostatistical data). In addition, Kim et al. (2003) and Anselin and Le Gallo (2006) conducted hedonic studies using datasets obtained at residences (geostatistical data) and environmental datasets obtained at monitoring stations (geostatistical data).

In such analyses, the spatial supports (shape, size, etc.) of the datasets were unified in advance. For example, aggregations and disaggregations have been used to unify spatial units and interpolations have been used to unify data locations. The problem is, in both cases, the spatial support conversions must be conducted while considering their influence on the subsequent analysis.

It is generally known that an aggregation or disaggregation introduces biases into parameter estimates. Such a problem is called a modifiable areal unit problem (MAUP; Openshaw and Taylor, 1979). The seriousness of MAUP depends on two factors (Wang, 2009). The first is the underlying spatial pattern of data. The influence of MAUP is large if the data are positively spatially dependent, whereas the influence is small if the data are negatively dependent (e.g., Reynolds, 1998). The second factor is the aggregation process. Because more variability is canceled out by the aggregation, the MAUP is prominent when variables are aggregated into larger aggregation units.

To cope with MAUP, at least five approaches have been proposed in the field of geography (Swift et al., 2008). The first is to apply geographically weighted regression (GWR: Fotheringham et al., 2002). GWR models the spatial pattern of data explicitly and is therefore believed to be robust to MAUP. However, GWR does not cope with MAUP explicitly, so it cannot really be regarded as a solution to the problem (Fotheringham et al., 2002; Wang, 2009). The second approach is to apply non-aggregated data, although non-aggregated data are usually unavailable (e.g., Tagashida and Okabe, 2002). The third approach is to estimate aggregated-level parameters while considering the non-aggregated level structure in a variance-covariance matrix (e.g., Tranmer and Steel, 1998). While this approach has the drawback that a hierarchical structure must be assumed in the variance-covariance matrix, it is a basis of the geostatistical approaches of the change of support problem (COSP) that is discussed below. The fourth approach is to optimize the zoning system by minimizing the intra-zone variances and maximizing the variances between zones (Openshaw, 1984). However, this is only applicable when the zonal system can be modified. Finally, the fifth approach is a sensitivity analysis that makes a decision by examining the behavior of parameters by altering the zoning systems (e.g., Odoi et al., 2003; Swift et al., 2008). However, in this case, a drawback is that arbitrariness might be introduced. Thus, there is no general solution to MAUP (Goodchild, 2001; Wang, 2009).

In geostatistics, the problem of changing spatial supports of data is called COSP (e.g., Gotway and Young, 2002; Schabenberger and Gotway, 2005; Fuentes and Raftery, 2005; Lee et al., 2009; Young et al., 2009a, b; Berrocal et al., 2010; Sahu et al., 2010; Gelfand, 2010, 2012; Nagle et al., 2011; Berrocal et al., 2012). COSP refers to a general framework that contains MAUP, aggregation/disaggregation, and interpolation (Gotway and Young, 2002), and all these problems have been discussed using models describing non-aggregate level underlying spatial processes. Whereas the majority of COSP studies have focused on the changes of supports themselves, some studies have demonstrated that geostatistical models effectively reduce the biases caused by these changes of support. For example, Nagle et al. (2011) discusses how the parameters in a non-aggregate level process are estimated consistently using a dataset obtained by aggregating the process.

Although COSP studies focus on geostatistics, they are still only within this field. This study introduces certain ideas from COSP into GWR, a geographical method, and then proposes an extension of GWR that copes explicitly with COSP and MAUP.

The paper is organized as follows. GWR and a geostatistical approach for COSP are briefly discussed in sections 2 and 3, respectively. A GWR-based method for COSP is constructed in section 4. Finally, the effectiveness of the constructed method and standard GWR is discussed using a simulation study in section 5.

2. Geographically weighted regression

GWR is a spatial extension of the standard linear regression model that allows parameter values to vary continuously in geographical space. The basic model is as follows (see Fotheringham et al., 2002):

$$y(s_i) = \sum_p x_p(s_i)\beta_p(s_i) + \varepsilon(s_i)\,, \tag{1}$$

where i is an index of sites and s_i denotes a site on $D \subset \Re^2$, $y(s_i)$ are explained variables, $x_p(s_i)$ are the p-th explanatory variables, $\varepsilon(s_i)$ is an independent normally distributed disturbance, and $\beta_{s,p}$ is a spatially varying parameter.

$\beta_{s,p}$ is estimated by assigning more weights to nearby observations than to more distant observations. More precisely, let $\mathbf{\beta}_i$ is a P×1 vector whose p-th element is $\beta_p(s_i)$, the vector is given as follows:

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}'\mathbf{W}_i\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_i\mathbf{y}\,, \tag{2}$$

where $\mathbf{X}$ is a matrix of the explanatory variables, and $\mathbf{y}$ is a vector of the explained variables. $\mathbf{W}_i$ is the diagonal matrix whose j-th element is $w(s_i,s_j)$, the strength of the connectivity between s_i and s_j. $w(s_i,s_j)$ is given using a distance-decay function. For instance, the Gaussian function is one of the most standard functions:

$$w(s_i, s_j) = \exp\left(-\frac{d(s_i, s_j)^2}{h^2}\right), \tag{3}$$

where $d(s_i,s_j)$ is the distance between s_i and s_j, and h is a bandwidth parameter. The bandwidth parameter is calibrated using a cross-validation.

3. Geostatistics for COSP

This section briefly explains approaches for COSP in geostatistics. Consider a non-aggregate level spatial process $z(s_i)$, where $z(s_i)$ has mean μ and covariance function Cov[$z(s_i)$, $z(s_j)$] = $c(s_i, s_j)$. Suppose a spatial unit B_m on $D \subset \Re^2$, then, an aggregate level spatial process is expressed as:

$$z(B_m) = \frac{1}{|B_m|}\int_{B_m} z(s_i)ds_i \, . \tag{4}$$

Using Eq.(4), the covariance between two spatial units B_m and B_n, Cov[$z(B_m)$, $z(B_n)$]= $c(B_m, B_n)$, is defined as:

$$c(B_m, B_n) = \frac{1}{|B_m||B_n|}\int_{B_m}\int_{B_n} c(s_i, s_j)ds_i ds_j \, . \tag{5}$$

Similarly, the covariance between the site s_i and the unit B_m, Cov[$z(s_i)$, $z(B_n)$]= $c(s_i, B_n)$, is defined as:

$$c(s_i, B_m) = \frac{1}{|B_m|}\int_{B_m} c(s_i, s_i)ds_i \, . \tag{6}$$

Note that Eqs.(4), (5), (6) are described using matrix notations as

$$\bar{\mathbf{z}} = \mathbf{Nz} \, . \tag{7}$$

$$\overline{\mathbf{C}} = \mathbf{NCN}' \, . \tag{8}$$

$$\bar{\mathbf{c}} = \mathbf{NC} \, . \tag{9}$$

where $\mathbf{z}$ and $\bar{\mathbf{z}}$ are vectors whose elements are $z(s_i)$ and $z(B_m)$ respectively, $\mathbf{C}$ and $\overline{\mathbf{C}}$ are covariance matrixes whose elements are $c(s_i, s_j)$ and $c(B_m, B_n)$ respectively, and $\bar{\mathbf{c}}$ is a vector of covariances whose elements are $c(s_i, B_n)$.

Nagle et al. (2011) constructed an aggregate level model considering underlying spatial process, which is defined as follows:

$$\bar{\mathbf{y}} = \bar{\mathbf{X}}\boldsymbol{\beta} + \vec{\boldsymbol{\varepsilon}} \qquad \bar{\boldsymbol{\varepsilon}} \sim N(\mathbf{0}, \bar{\mathbf{C}}), \tag{10}$$

where $\bar{\mathbf{y}}$ =**Ny**, $\bar{\mathbf{X}}$ = **NX**, and $\vec{\boldsymbol{\varepsilon}}$ = **Nε**. COSP studies generally assumes that **y** in $\bar{\mathbf{y}}$ is unknown.

Nagle et al. (2011) showed that the maximum likelihood (ML) estimates of **β** and parameters in $\bar{\mathbf{C}}$ are consistent, asymptotically normally, and asymptotically efficient. The ML estimates of **β** is given as follows:

$$\hat{\boldsymbol{\beta}} = (\bar{\mathbf{X}}'\bar{\mathbf{C}}^{-1}\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}'\bar{\mathbf{C}}^{-1}\bar{\mathbf{y}}. \tag{11}$$

Best linear unbiased predictor (BLUP) of the non-aggregate level explained variables **y** is given as follows:

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \bar{\mathbf{c}}'\bar{\mathbf{C}}^{-1}(\bar{\mathbf{y}} - \bar{\mathbf{X}}\boldsymbol{\beta}), \tag{12}$$

Note that the equation is expanded using the aggregation matrix **N** as follows:

$$\begin{aligned} \mathbf{N}\hat{\mathbf{y}} &= \mathbf{N}\mathbf{X}\boldsymbol{\beta} + \mathbf{N}\bar{\mathbf{c}}'\bar{\mathbf{C}}^{-1}(\bar{\mathbf{y}} - \bar{\mathbf{X}}\boldsymbol{\beta}), \\ &= \mathbf{N}\mathbf{X}\boldsymbol{\beta} + \mathbf{N}\mathbf{C}\mathbf{N}'(\mathbf{N}\mathbf{C}\mathbf{N}')^{-1}(\bar{\mathbf{y}} - \mathbf{N}\mathbf{X}\boldsymbol{\beta}), \\ &= \mathbf{N}\mathbf{X}\boldsymbol{\beta} + \bar{\mathbf{y}} - \mathbf{N}\mathbf{X}\boldsymbol{\beta}, \\ &= \bar{\mathbf{y}}, \end{aligned} \tag{13}$$

Eq.(13) suggests that aggregations of the predictors equal to the aggregate level known explained variables.

Using a similar discussion, we would extend GWR.

4. GWR for COSP

4.1. Model

Suppose i.i.d. disturbances, as in standard GWR (Fotheringham et al., 2002; Wheeler and Tiefelsdorf, 2005); then, non-aggregate level GWR for site s_i is written using matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}). \tag{14}$$

Again, let us assume that **y** is unknown but its aggregation $\bar{\mathbf{y}}$ is known. By aggregating Eq.(14) using **N**, the aggregate level GWR can be given as

$$\bar{\mathbf{y}} = \bar{\mathbf{X}}\boldsymbol{\beta}_i + \bar{\boldsymbol{\varepsilon}} \quad \bar{\boldsymbol{\varepsilon}} \sim N(\mathbf{0}, \sigma^2\mathbf{N}\mathbf{N}'). \tag{15}$$

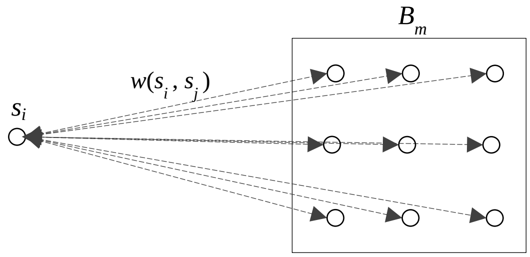

Figure 1: Image of connectivity between s_i and B_m: m-th diagonal of $\overline{\mathbf{W}}_i$ is given by the average of the spatial weights described using arrows (◄-------►)

Note that $\mathbf{NN}'$ is always a diagonal matrix as long as the aggregation units do not overlap. Obviously, Eq.(15) is a GWR whose disturbances are weighted by (diagonals of) $\mathbf{NN}'$, and accordingly, the estimates of $\boldsymbol{\beta}_i$ are given as

$$\hat{\boldsymbol{\beta}}_i = (\overline{\mathbf{X}}'\overline{\mathbf{W}}_i^{1/2}(\mathbf{NN}')^{-1}\overline{\mathbf{W}}_i^{1/2}\overline{\mathbf{X}})^{-1}\overline{\mathbf{X}}'\overline{\mathbf{W}}_i^{1/2}(\mathbf{NN}')^{-1}\overline{\mathbf{W}}_i^{1/2}\overline{\mathbf{y}}. \tag{16}$$

where $\overline{\mathbf{W}}_i$ is a diagonal matrix whose m-th element describes the connectivity between site s_i and the m-th aggregation unit. Following the COSP studies, we assume that $\overline{\mathbf{W}}_i = \mathbf{NW}_i\mathbf{N}'$ (see Figure 1).

The constructed model applies to the change of support itself (e.g., spatial unit conversion/downscaling), as with geostatistical methods. The BLUPs for each site are given from Eqs.(14) and (15) as

$$\hat{\mathbf{y}} = \boldsymbol{\mu} + \mathbf{N}'(\mathbf{NN}')^{-1}(\overline{\mathbf{y}} - \mathbf{N}\boldsymbol{\mu}), \tag{17}$$

$$\boldsymbol{\mu} = < \mathbf{x}_i'\boldsymbol{\beta}_i >,$$

where $\mathbf{x}_i$ is a vector of the explanatory variables at s_i, and $\langle \mathbf{x}_i'\boldsymbol{\beta}_i \rangle$ is a vector whose elements are given using $\mathbf{x}'_i\boldsymbol{\beta}_i$.

The parameters in our basic model shown in Eq.(15) are estimated using the following steps: (i) calibrate the bandwidth parameter r in $\overline{\mathbf{W}}_i$; (ii) substitute the estimated bandwidth into Eq.(16) and obtain $\hat{\boldsymbol{\beta}}_i$; and (iii) calculate $\hat{\mathbf{y}}$ using Eq.(17).

Because our method considers the aggregation process explicitly it might be robust to MAUP as with Nagle et al. (2011). In addition, it is also a method for COSPs. More precisely, the method provides non-aggregate level parameter estimates ($\hat{\boldsymbol{\beta}}_i$) and BLUP of explanatory variables ($\hat{\mathbf{y}}$). However, because of information loss due to aggregations, it is generally impossible to recover the non-aggregation level processes perfectly, and the parameter estimates and the predictors must be inefficient, depending on the underlying spatial processes of the data and aggregation processes.

Thus, the next section examines the effectiveness of the proposed method using a simulation study.

5. Simulation study

5.1. Assumptions

This section examines the effectiveness of the proposed model by applying it to the aggregated datasets obtained from simulations.

Non-aggregate level explained variables and explanatory variables are generated for each of the 30 × 30 sites. The explanatory variables include an intercept and two variables, $x_1(s_i)$ and $x_2(s_i)$, which are generated independently from $N(0,1)$ respectively. The explained variables are generated from Eq.(18):

$$y(s_i) = \alpha(s_i) + x_1(s_i)\beta_1(s_i) + x_2(s_i)\beta_2(s_i) + \varepsilon(s_i) \qquad \varepsilon(s_i) \sim N(0, \sigma^2), \tag{18}$$

where $\alpha(s_i)$, $\beta_1(s_i)$, and $\beta_2(s_i)$ are spatially varying parameters.

True values of $\alpha(s_i)$, $\beta_1(s_i)$, and $\beta_2(s_i)$ are required to evaluate validities for MAUP. In this study, following Finley (2010), these parameters are given using Gaussian processes whose covariance functions are τ^2exp(-$d(s_i,s_j)^2/range$), where τ^2 for $\alpha(s_i)$ and $\beta_2(s_i)$ are both 2.0 and τ^2 for $\beta_1(s_i)$ is 0.5. τ^2 describe the effectiveness of the explanatory variables: the intercept and $x_2(s_i)$ corresponding to τ^2 = 2.0 explain $y(s_i)$ effectively, while $x_1(s_i)$ does not. *range* in the equation represents the coarseness of the Gaussian processes. Note that *range* differs from the bandwidth parameter h in GWR. The resulting true distributions of $\alpha(s_i)$, $\beta_1(s_i)$, and $\beta_2(s_i)$ are as shown in Figure 2. This study would evaluate the effectiveness of the proposed method by comparing its parameter estimates with their true distributions.

Our simulations are performed by altering σ^2, r, and the number of aggregation units M as follows: σ^2 = {0.25, 1.00, 2.00}, *range* = {5, 10, 20}, M = {50, 100, 250}. In each case, the parameters are estimated 50 times and the results are compared among the models, as explained below. Following Nagle et al. (2010), the

M aggregation units are created using the Voronoi tessellation, and the parameters in the proposed model (GWR_COSP) are estimated using the aggregations of $y(s_i)$, $x_1(s_i)$, and $x_2(s_i)$.

5.2. Results

We estimated $\alpha(s_i)$, $\beta_1(s_i)$, and $\beta_2(s_i)$ using the proposed model shown in Eq.(15) (GWR_COSP). The estimates obtained in each of the first attempts are shown in Figures 3 (*range* = 5), 4 (*range* = 10), and 5 (*range* = 20). Figure 3 shows that the interpolated parameters are over-smoothed when the true parameters have local spatial variations and the number of aggregation units M is small. This figure also demonstrates that the spatial distributions of the interpolated parameters vary depending on M. In contrast, Figures 4 and 5 suggest that the imputation results of $\alpha(s_i)$ or $\beta_2(s_i)$, which have large powers (τ^2), are similar across aggregation scales, and these distributions are also similar to their true distributions (see Figure 2). These results might indicate that GWR_COSP is robust for MAUP if the spatial variations in the parameters are significant (τ^2 is large) and global compared to the spatial scales of aggregations.

The accuracy of these imputed parameter estimates are measured by evaluating the differences between these estimates and their true distributions (Figure 2) using the mean R-squared value (R^2: Table 1) and the root mean square error (RMSE: Table 2). For comparison, they are also evaluated using non-aggregate level conventional GWR Eq.(1) (GWR_NAg). Note that the results of GWR_NAg will always be better than GWR_COSP because it is based on non-aggregated variables. The calculation results of mean R^2 show that the estimates of $\alpha(s_i)$ and $\beta_2(s_i)$ in GWR_COSP describe their non-aggregate level true distributions well on the condition that both M and *range* are not small. For example, the R^2 means are at least 0.7 when $M \geq 100$ and *range* ≥ 10. On the other hand, the accuracy tends to decrease significantly when M = 50 and/or *range* = 5. The results of the RMSEs indicate similar tendencies. Thus, it is verified that, as expected, GWR_COSP is robust for MAUP if the spatial scales of the parameter distributions are global compared to the scales of aggregations.

Finally, we discuss the influence of aggregations on the bandwidth parameter estimates by comparing the estimates of GWR_COSP and an aggregate level standard GWR (GWR_Ag). Here, the differences between these estimates and the estimates of GWR_NAg are measured using RMSE (see Table 3). The results suggest that the RMSEs of GWR_COSP are smaller than GWR_Ag in all cases. In addition, the value of "ratio of improve" in Table 3 suggests that the RMSEs of GWR_COSP are more efficient in almost all the attempts. However, the estimated bandwidths of GWR_COSP (and GWR_Ag) are still larger than those of GWR_NAg in most cases (not shown in Table 3), and their gaps are large when M is small. The large range might be one reason for the over-smoothed estimates of $\alpha(s_i)$, $\beta_1(s_i)$, and $\beta_2(s_i)$ that appeared when M was small (see the first line of Figure 3).

In summary, the proposed method is robust for MAUP under certain conditions but has the limitation that the estimates are inefficient if the scales of aggregations are large compared to the spatial scales of variability of the true parameter distributions.

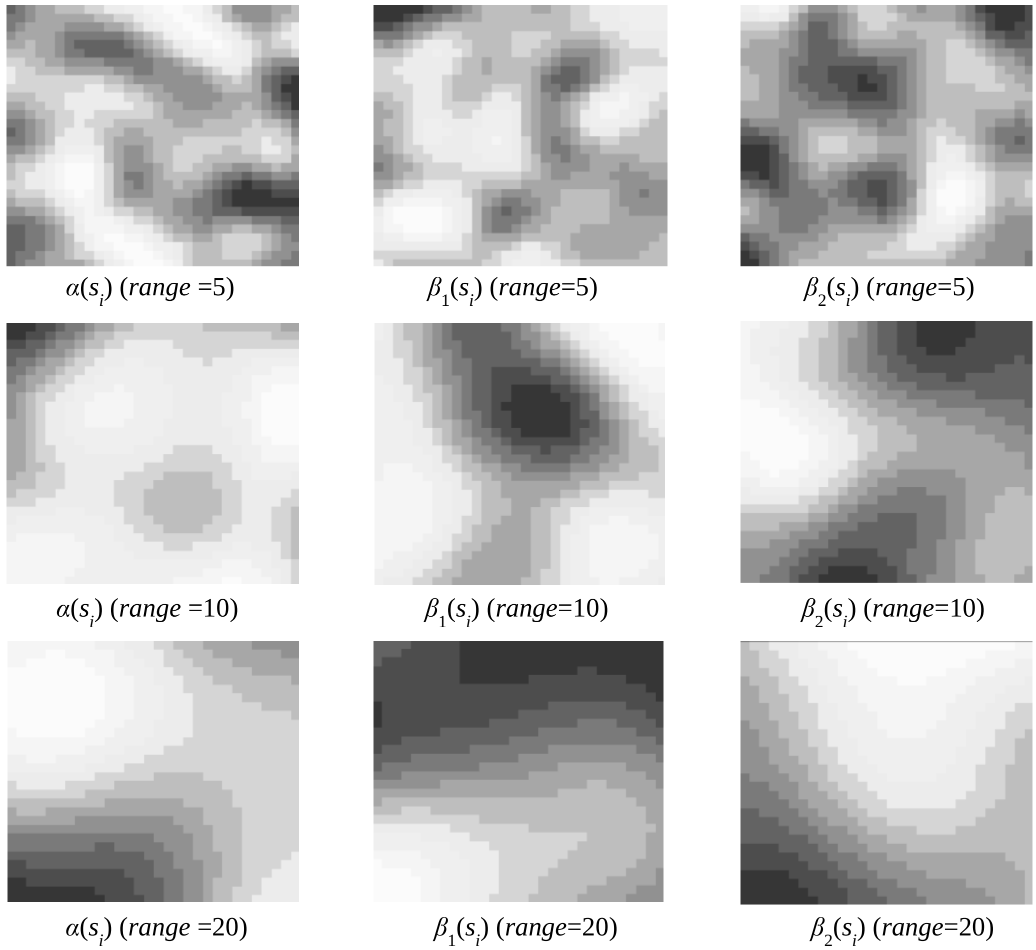

Figure 2: (Standardized) true distributions of $\alpha(s_i)$, $\beta_1(s_i)$, and $\beta_2(s_i)$

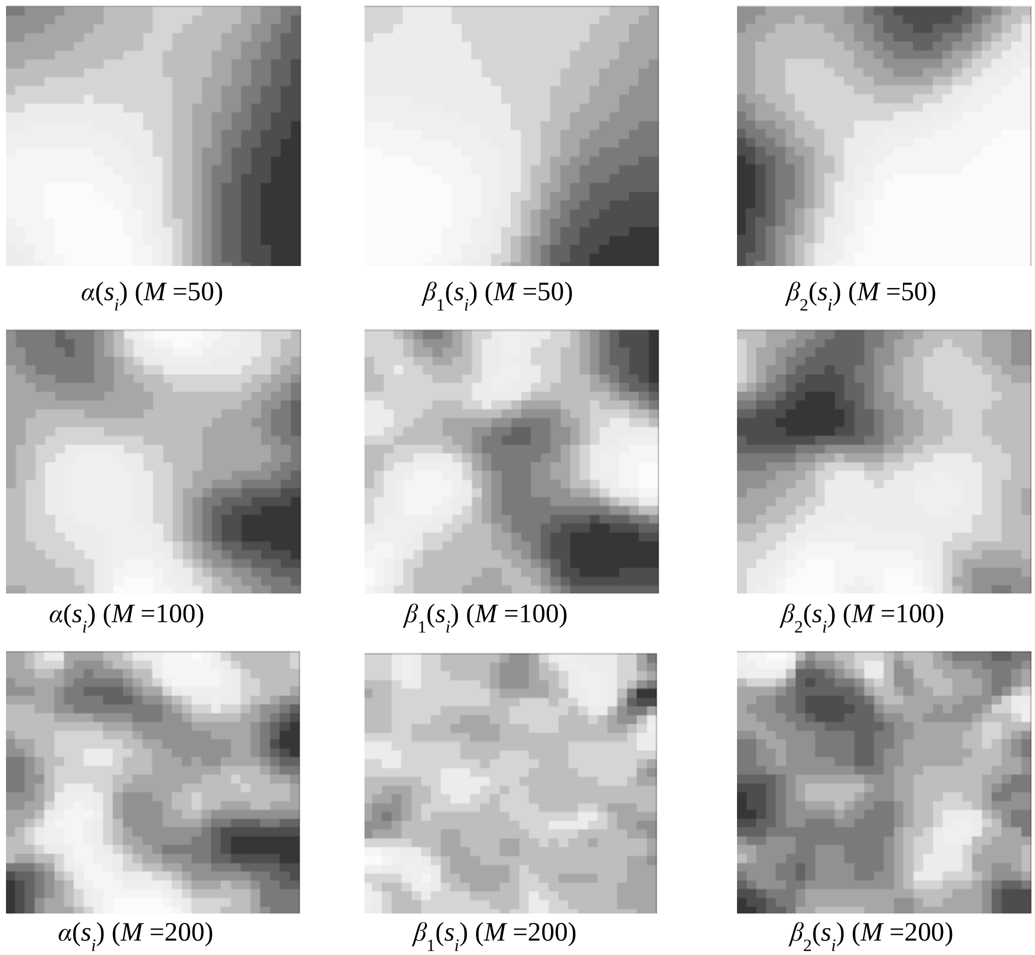

Figure 3: Estimated distributions of $\alpha(s_i)$, $\beta_1(s_i)$, and $\beta_2(s_i)$ (range=5): true distributions of these parameters are appeared on the first row in Figure 2

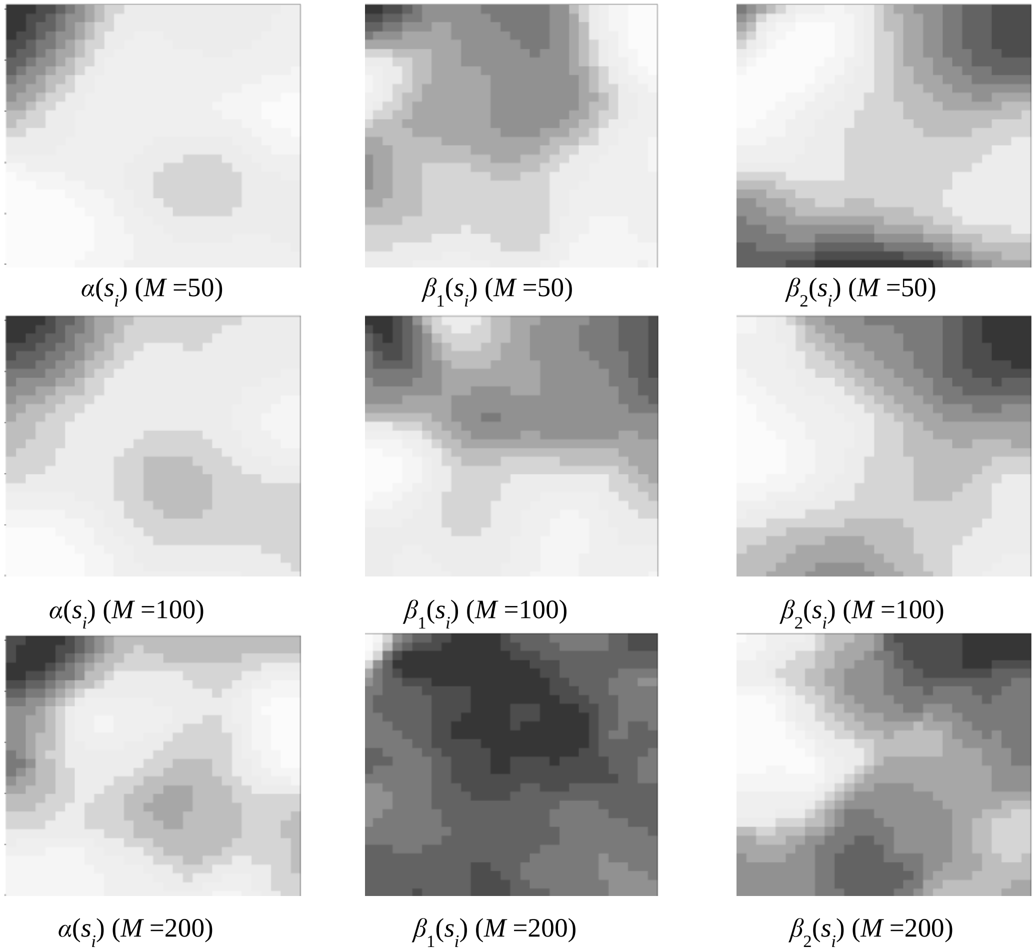

Figure 4: Estimated distributions of $\alpha(s_i)$, $\beta_1(s_i)$, and $\beta_2(s_i)$ (range=10): true distributions of these parameters are appeared on the second row in Figure 2

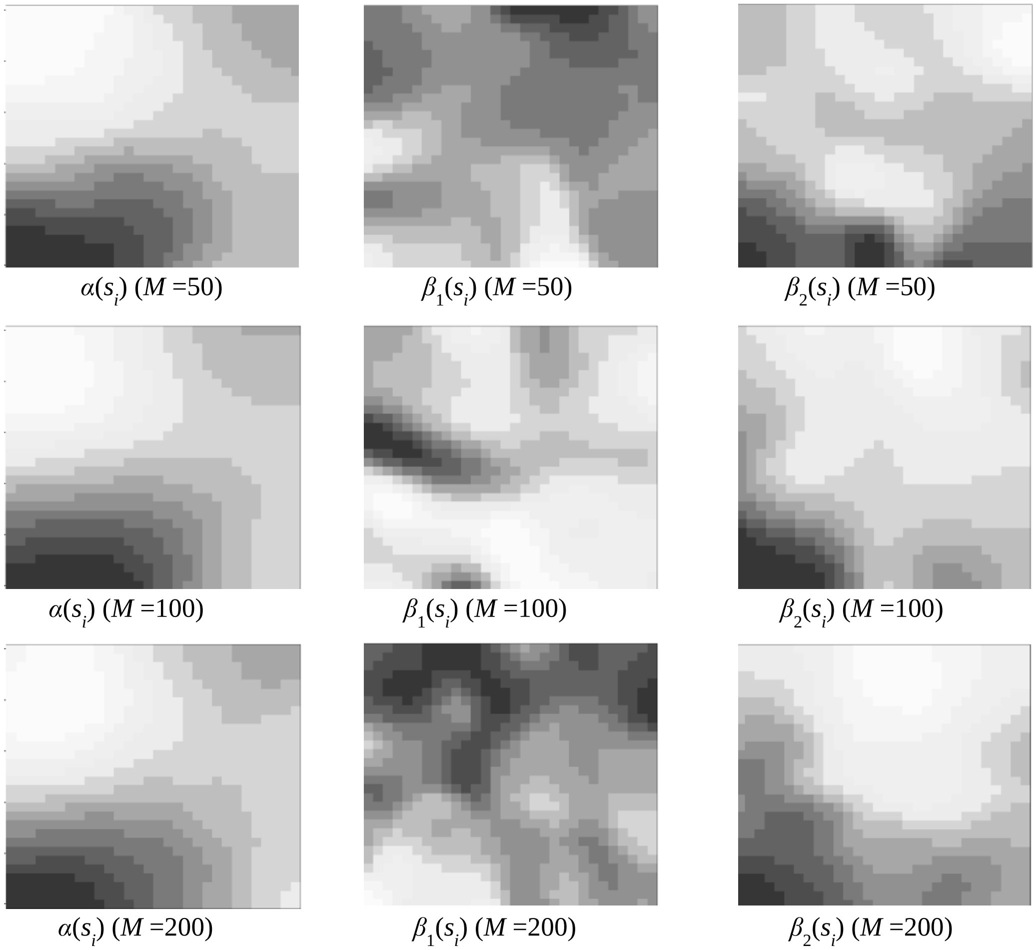

Figure 5: Estimated distributions of $\alpha(s_i)$, $\beta_1(s_i)$, and $\beta_2(s_i)$ (range=20): true distributions of these parameters are appeared on the third row in Figure 2

Table 1: Averages of R-squared (R^2) values between estimates of $\alpha(s_i)$, $\beta_1(s_i)$, and $\beta_2(s_i)$ and their true values (see Figure 2): **Bold**: average $R^2 > 0.5$

range = 5

Num. of agg. units	σ^2	$\alpha(s_i)$		$\beta_1(s_i)$		$\beta_2(s_i)$	
		GWR NAg	GWR COSP	GWR NAg	GWR COSP	GWR NAg	GWR COSP
	0.25	**0.97**	0.20	**0.70**	0.04	**0.95**	0.07
50	1.00	**0.95**	0.17	**0.62**	0.03	**0.93**	0.05
	2.00	**0.93**	0.16	**0.54**	0.03	**0.91**	0.06
	0.25	**0.97**	**0.54**	**0.71**	0.06	**0.94**	0.17
100	1.00	**0.95**	**0.58**	**0.61**	0.05	**0.93**	0.15
	2.00	**0.93**	0.49	**0.55**	0.04	**0.90**	0.15
	0.25	**0.97**	**0.88**	**0.69**	0.15	**0.94**	**0.57**
200	1.00	**0.95**	**0.86**	**0.61**	0.14	**0.92**	**0.57**
	2.00	**0.93**	**0.84**	**0.54**	0.13	**0.90**	**0.56**

range = 10

Num. of agg. units	σ^2	$\alpha(s_i)$		$\beta_1(s_i)$		$\beta_2(s_i)$	
		GWR NAg	GWR COSP	GWR NAg	GWR COSP	GWR NAg	GWR COSP
	0.25	**0.96**	**0.64**	**0.95**	0.11	**0.99**	0.46
50	1.00	**0.93**	**0.59**	**0.91**	0.13	**0.98**	0.43
	2.00	**0.89**	**0.58**	**0.87**	0.12	**0.97**	0.43
	0.25	**0.96**	**0.86**	**0.95**	0.32	**0.99**	**0.74**
100	1.00	**0.92**	**0.82**	**0.90**	0.33	**0.98**	**0.72**
	2.00	**0.89**	**0.79**	**0.87**	0.27	**0.97**	**0.70**
	0.25	**0.96**	**0.94**	**0.95**	**0.71**	**0.99**	**0.94**
200	1.00	**0.93**	**0.91**	**0.90**	**0.66**	**0.98**	**0.92**
	2.00	**0.89**	**0.87**	**0.86**	**0.59**	**0.97**	**0.90**

range = 20

Num. of agg. units	σ^2	$\alpha(s_i)$		$\beta_1(s_i)$		$\beta_2(s_i)$	
		GWR NAg	GWR COSP	GWR NAg	GWR COSP	GWR NAg	GWR COSP
	0.25	**0.99**	**0.95**	**0.94**	0.19	**0.99**	0.43
50	1.00	**0.98**	**0.94**	**0.88**	0.18	**0.98**	0.43
	2.00	**0.97**	**0.93**	**0.82**	0.17	**0.97**	0.41
	0.25	**0.99**	**0.98**	**0.94**	0.32	**0.99**	**0.77**
100	1.00	**0.98**	**0.97**	**0.89**	0.31	**0.98**	**0.72**
	2.00	**0.97**	**0.96**	**0.83**	0.32	**0.97**	**0.73**
	0.25	**0.99**	**0.99**	**0.94**	**0.71**	**0.99**	**0.94**
200	1.00	**0.98**	**0.98**	**0.88**	**0.59**	**0.98**	**0.92**
	2.00	**0.97**	**0.97**	**0.85**	**0.56**	**0.97**	**0.89**

Table 2: Root mean square errors (RMSE) the estimates of $\alpha(s_i)$, $\beta_1(s_i)$, and $\beta_2(s_i)$: Each RMSE are evaluated using their true values (see Figure 2): **Bold**: estimates of GWR_COSP whose RMSE if less than twice of GWR_NAg

range = 5

Num. of agg. units	σ^2	$\alpha(s_i)$		$\beta_1(s_i)$		$\beta_2(s_i)$	
		GWR NAg	GWR COSP	GWR NAg	GWR COSP	GWR NAg	GWR COSP
50	0.25	0.39	1.82	0.36	2.26	0.44	2.66
	1.00	0.48	1.85	0.43	1.72	0.52	2.50
	2.00	0.56	1.82	0.47	2.20	0.59	2.52
100	0.25	0.39	1.44	0.36	1.94	0.45	2.40
	1.00	0.48	1.42	0.43	2.05	0.52	2.47
	2.00	0.56	1.43	0.48	2.19	0.60	2.34
200	0.25	0.38	0.77	0.36	1.07	0.45	1.25
	1.00	0.47	**0.81**	0.42	1.04	0.52	1.25
	2.00	0.57	**0.88**	0.47	1.12	0.59	1.30

range = 10

Num. of agg. units	σ^2	$\alpha(s_i)$		$\beta_1(s_i)$		$\beta_2(s_i)$	
		GWR NAg	GWR COSP	GWR NAg	GWR COSP	GWR NAg	GWR COSP
50	0.25	0.17	0.66	0.16	1.61	0.18	3.04
	1.00	0.25	0.60	0.21	1.62	0.25	1.88
	2.00	0.30	0.62	0.25	1.33	0.30	1.74
100	0.25	0.17	0.37	0.16	0.94	0.19	0.94
	1.00	0.24	**0.39**	0.21	0.93	0.26	1.02
	2.00	0.30	**0.43**	0.25	0.95	0.30	1.00
200	0.25	0.17	**0.23**	0.16	0.39	0.18	0.43
	1.00	0.25	**0.29**	0.21	0.43	0.25	**0.49**
	2.00	0.30	**0.35**	0.25	**0.50**	0.30	**0.55**

range = 20

Num. of agg. units	σ^2	$\alpha(s_i)$		$\beta_1(s_i)$		$\beta_2(s_i)$	
		GWR NAg	GWR COSP	GWR NAg	GWR COSP	GWR NAg	GWR COSP
50	0.25	0.12	0.32	0.10	1.25	0.13	1.23
	1.00	0.17	0.35	0.15	1.32	0.18	1.39
	2.00	0.22	**0.38**	0.17	1.26	0.22	1.35
100	0.25	0.12	**0.21**	0.10	0.55	0.12	0.60
	1.00	0.17	**0.24**	0.14	0.62	0.18	0.64
	2.00	0.21	**0.27**	0.18	0.61	0.22	0.67
200	0.25	0.11	**0.14**	0.10	0.24	0.13	0.27
	1.00	0.17	**0.19**	0.14	0.28	0.19	**0.34**
	2.00	0.21	**0.23**	0.17	**0.34**	0.22	**0.38**

Table 3: RMSE of the bandwidth parameters: Here, bandwidth parameters estimated by GWR_Nag are regarded as their true values. Ratio of improve. means the ratio that RMSE of GWR_COSP is smaller than GWR_Ag

Num. of agg. units	σ^2	*range* = 5			*range* = 10			*range* = 20		
		GWR Ag	GWR COSP	Ratio of improve.	GWR Ag	GWR COSP	Ratio of improve.	GWR Ag	GWR COSP	Ratio of improve.
50	0.25	4.99	4.72	0.94	2.73	2.71	0.90	2.15	1.88	0.90
	1.00	3.63	3.33	0.98	2.69	2.43	0.98	1.88	1.61	0.92
	2.00	5.21	4.87	1.00	3.02	2.74	1.00	1.65	1.35	1.00
100	0.25	2.33	2.13	0.94	1.33	1.08	0.86	1.08	0.91	0.94
	1.00	2.21	2.06	0.96	1.30	1.12	0.98	0.95	0.75	1.00
	2.00	2.37	2.20	0.98	1.25	1.10	1.00	0.89	0.76	0.98
200	0.25	0.69	0.57	0.92	0.55	0.44	0.98	0.47	0.39	0.96
	1.00	0.57	0.45	0.98	0.48	0.37	0.98	0.47	0.41	0.98
	2.00	0.55	0.44	1.00	0.55	0.46	1.00	0.64	0.59	1.00

Reference

Anselin, L. and Le Gallo, J. (2006) Interpolation of Air Quality Measures in hedonic house price models: spatial aspects. *Spatial Economic Analysis*, **1** (1), 31-52.

Berrocal, V.J., Gelfand, A.E., Holland, D.M. (2010) A spatio-temporal downscaler for outputs from numerical models. *Journal of Agricultural, Biological and Environmental Statistics*, **15** (2), pp.176-197.

Berrocal, V.J., Gelfand, A.E., Holland, D.M. (2012) Space-time data fusion under error in computer model output: an application to modeling air quality. *Biometrics*, **68** (3), pp.837-848.

Dominici, F., Samet, J.M., Zeger, S.L. (2000) Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **163** (3), pp.263-302.

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

Fotheringham, S., Brunsdon, C., Charlton, M. (2002) *Geographically weighted regression: The analysis of spatially varying relationships*. Wiley.

Fuentes, M. and Raftery, A.E. (2005) Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, **61**, pp.36-45.

Gelfand, A.E. (2010) Misaligned spatial data: the change of support problem. In *Handbook of Spatial Statistics* (Gelfand, A.E., Diggle, P., Guttorp, P., Fuentes eds.), CRC Press, **29**, pp. 517-539.

Gelfand, A.E. (2012) Hierarchical modeling for spatial data problems. *Spatial Statistics*, **1**, pp.30-39.

Goodchild, M. (2001) Models of scale and scales of modeling. In Modeling scale in geographical information science (Tate, N. and Atkinson, P. eds.), John Wiley and Sons, pp. 3-10.

Gotway, C.A. and Young, L.J. (2002) Combining incompatible spatial data. *Journal of the American Statistical Association*, **97** (458), 632-648.

Kim, C.W, Phipps, T.T., Anselin, L. (2003) Measuring the benefits of air quality improvement: a spatial hedonic approach. *Journal of Environmental Econometrics and Management*, **45** (1), pp.24-39.

Lee, S.J., Yeatts, K.B., Serre, M.L. (2009) A Bayesian maximum entropy approach to address the change of support problem in the spatial analysis of childhood asthma prevalence across North Carolina. *Spatial and Spatio-temporal Epidemiology*, **1** (1), pp.49-60.

Nagle, N.N., Sweeney, S.H., Kyriakidis, P.C. (2011) A geostatistical linear regression model for small area data. *Geographical Analysis*, **43** (1), pp.38-60.

Odoi, A., Martin, W., Michel, P., Holt, J., Middleton, D., Wilson, J. (2003) Geographical and temporal distribution of human giardiasis in Ontario Canada. *International Journal of Health Geographics*, **2** (5).

Openshaw, S. (1984). *The modifiable areal unit problem*. Norwich, UK: Geo Books

Openshaw, S. and Taylor, P. (1979) A Million or so correlation coefficients: three experiments on the modifiable areal unit problem. In *Statistical Methods in the Spatial Sciences* (Wrigley, N. eds.), London: Pion, pp.127-144.

Reynolds, H.D. (1998) *The modifiable areal unit problem: empirical analysis by statistical simulation*. University of Toronto, 1998.

Sahu, S.K., Gelfand, A.E., Holland. D.M. (2010) Fusing point and areal level space–time data with application to wet deposition. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59** (1), pp.77-103.

Schabenberger, O. and Gotway, C.A. (2005) *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC.

Swift, A, Liu, L., Uber, J. (2008) Reducing MAUP bias of correlation statistics between water quality and GI illness. *Computers, Environment and Urban Systems*, **32**, pp.134-148.

Tagashira, N. and Okabe, A. (2002) The modifiable areal unit problem in a regression model whose independent variable is a distance from a predetermined point. *Geographical Analysis*, **34** (1), pp.1-19.

Tranmer, M., and Steel, D. (1998). Using census data to investigate the causes of the ecological fallacy. *Environment and Planning A*, 30, pp.817-831.

Young, L.J., Gotway, C.A., Kearney, G., Duclos, C. (2009a) Assessing uncertainty in support-adjusted spatial misalignment problems. *Communications in Statistics: Theory & Methods*, **38** (16/17), pp.3249-3264.

Young, L.J., Gotway, C.A., Yang, J. Kearney, G., Duclos, C. (2009b) Linking health and environmental data in geographical analysis: it's so much more than centroids. *Spatial and Spatio-temporal Epidemiology*, **1** (1), pp.73-84.

Wong, D. (2009) The modifiable areal unit problem (MAUP). In *The SAGE Handbook of Spatial Analysis* (Fotheringham, A.S. and Rogerson, P.A. eds.), SAGE, **7**, pp.105-124.

Wheeler, D. and Tiefelsdorf (2005) Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, **7** (2), 161-187.