Statistics
in Medicine

Research Article

Received 26 February 2016, Accepted 27 October 2016 Published online 22 November 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7172

Cluster detection of spatial regression coefficients

Junho Lee,[a] Ronald E. Gangnon[b*†] and Jun Zhu[c]

Popular approaches to spatial cluster detection, such as the spatial scan statistic, are defined in terms of the responses. Here, we consider a varying-coefficient regression and spatial clusters in the regression coefficients. For varying-coefficient regression, such as the geographically weighted regression, different regression coefficients are obtained for different spatial units. It is often of interest to the practitioners to identify clusters of spatial units with distinct patterns in a regression coefficient, but there is no formal statistical methodology for that. Rather, cluster identification is often ad-hoc such as by eyeballing the map of fitted regression coefficients and discerning patterns. In this paper, we develop new methodology for spatial cluster detection in the regression setting based on hypotheses testing. We evaluate our methods in terms of power and coverages for true clusters via simulation studies. For illustration, our methodology is applied to a cancer mortality dataset. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: geographically weighted regression; hypothesis testing; spatial cluster detection; spatial scan statistic; varying coefficient regression.

1. Introduction

Cluster detection, the identification of spatial units adjacent in space that are associated with distinctive patterns of data of interest relative to background variation, is an important problem in disciplines such as spatial epidemiology and disease surveillance. For count data, clusters have distinctive risks of an event of interest: typically elevated, but possibly reduced, relative to background variation. For continuous data, clusters show higher or lower mean values than the background.

Spatial scan statistics [1,2] and their variants [3–11] are popular approaches to cluster detection within a frequentist hypothesis testing framework. The scan statistic is the maximum likelihood ratio test statistic based on a large collection of potential clusters of a particular regular geometric form (e.g., circles). Significance is evaluated via Monte Carlo simulation under an assumed null hypothesis, such as a constant risk over the entire spatial domain.

An alternative approach to spatial cluster detection uses Bayesian models for the underlying event rates that incorporate explicit spatial clusters associated with distinctive, either elevated or lowered, risks [12–18]. These models allow for formal inference regarding the number, locations, and risks associated with clusters relative to a model-specified and possibly non-uniform background risk. The aforementioned spatial cluster detection approaches, however, are all defined in terms of the responses. Here, we consider a new problem, namely, cluster detection of spatial regression coefficients.

In a spatial regression framework, it is plausible that a subdomain has a different relationship between the response and a covariate than the background. Such a subdomain can be considered a spatial cluster with different regression coefficients inside/outside the cluster. Alternatively, one can consider varying-coefficient regression such as the geographically weighted regression (GWR) [19, 20]. For example, GWR allows the relationship between a response and covariates to vary geographically by considering

[a]*Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A.*
[b]*Department of Biostatistics and Medical Informatics and Department of Population Health Sciences, University of Wisconsin, Madison, WI 53726, U.S.A.*
[c]*Department of Statistics and Department of Entomology, University of Wisconsin, Madison, WI 53706, U.S.A.*
[*]*Correspondence to: Ronald E. Gangnon, Department of Biostatistics and Medical Informatics and Department of Population Health Sciences, University of Wisconsin, Madison, WI 53726, U.S.A.*
[†]*E-mail: ronald@biostat.wisc.edu*

Copyright © 2016 John Wiley & Sons, Ltd.

locally weighted regression coefficients. Then, cluster identification can be carried out by eyeballing the smooth map of fitted regression coefficients. This method does not directly model clustering of regression coefficients. In addition, Lawson *et al.* [21] proposed discrete grouping of regression coefficients by considering a prior distribution for spatial grouping in a Bayesian framework. While this method directly provides grouping of regression coefficients, the number of groups needs to be specified in advance. Here, we propose new approach that enables the detection of an unknown number of spatial clusters in terms of the relationships between the response and the covariate.

In particular, we focus on spatially varying coefficient regression models and develop new methodology for spatial cluster detection with a covariate. For a single cluster, we consider testing potential circular clusters of regression coefficients against the null hypothesis that the regression coefficient is the same over the entire spatial domain by an F statistic. The p-value of our test is obtained via a Monte Carlo simulation. For multiple clusters, we adopt the sequential detection approach as Zhang *et al.* [22] proposed. Further, we propose two methods to detect multiple clusters sequentially in the regression setting. The first method detects significant clusters in the slopes and the intercepts simultaneously. In the second method, significant clusters in the slopes are detected first, and then in the intercepts. We believe that our method is the first of its kind to cluster the relationship between the response and the covariate in space. With a unified modeling framework for spatial clusters of covariates in relation to the response, it is more rigorous to discern heterogeneity of the relationship in terms of spatial clusters and more intuitive to interpret the spatial patterns than GWR. The main challenge in developing our method is computing time. A large number of matrix manipulations are involved due to the large number of potential clusters. However, we resolve the computational challenge by devising an efficient algorithm that reduces the computational complexity.

The remainder of the paper is organized as follows. In Section 2, we develop a test for spatial cluster effects in a simplified set, and propose a simultaneous detection method in intercepts and slopes. For multiple clusters, we also propose a two-stage method in Section 3. In Section 4, we evaluate these methods in terms of power and coverages for true clusters via simulation studies. For illustration, our proposed methodology is applied to a cancer mortality dataset in the Southeast of U.S.A in Section 5. Details about the computation are given as Appendix.

2. Simultaneous Spatial Cluster Detection in Intercepts and Slopes

2.1. Test for Spatial Cluster Effects in a Simplified Setting

Let D denote a spatial domain of interest in $\mathbb{R}^2$. Let N denote the number of cells that partition the spatial domain D and form a spatial lattice. For cell $i = 1, \dots, N$, let y_i denote the ith response variable. We model the response variable as $y_i = \mu_i + \varepsilon_i$, where ε_i is a random error and the ε_i's are independently and identically distributed (*iid*) as $N(0, \sigma^2)$ for a variance component $\sigma^2 > 0$. Let J denote the number of clusters on the spatial lattice and the clusters are denoted $C_1, \dots, C_J$ such that

$$C_j = \left\{ i \mid d(\boldsymbol{s}_i, \boldsymbol{c}_j) \leqslant r_j \right\},$$

where $j = 1, \dots, J$, $\boldsymbol{s}_i = (s_{1i}, s_{2i})^T$ denotes the coordinates of the geographical centroid of cell i, $\boldsymbol{c}_j$ and r_j are the center and radius of the spatial extent of cluster C_j, and $d(\cdot, \cdot)$ is the distance between two locations. Then, the mean response μ_i follows a varying-coefficient model

$$\mu_i = \begin{cases} \beta_0 + \beta_1 x_i & \text{if } i \notin \bigcup_{j=1}^{J} C_j \\ (\beta_0 + \theta_{1,0}) + (\beta_1 + \theta_{1,1}) x_i & \text{if } i \in C_1 \\ \vdots & \\ (\beta_0 + \theta_{J,0}) + (\beta_1 + \theta_{J,1}) x_i & \text{if } i \in C_J \end{cases}, \tag{1}$$

where x_i is the ith covariate, β_0 and β_1 are the intercept and the slope for the background (i.e., non-cluster), $\theta_{j,0}$ and $\theta_{j,1}$ are the cluster C_j effect in the intercepts and in the slopes. We begin with a single cluster $C \equiv C_1$ (i.e., $J = 1$) and assume that the cluster C is known *a priori*. Then, model (1) can be rewritten as

$$\mu_i = \begin{cases} \beta_0 + \beta_1 x_i & \text{if } i \notin C \\ (\beta_0 + \theta_0) + (\beta_1 + \theta_1) x_i & \text{if } i \in C \end{cases}, \tag{2}$$

Copyright © 2016 John Wiley & Sons, Ltd.

Next, we develop hypothesis testing for the cluster effect, which will be extended to test for an unknown cluster in the subsequent sections. For model (2) and a fixed cluster C, we may consider four possible hypotheses: $H_0 : \theta_0 = \theta_1 = 0$, $H_1 : \theta_0 \neq 0$, $\theta_1 = 0$, $H_2 : \theta_0 \neq 0$, $\theta_1 \neq 0$, and $H_3 : \theta_0 = 0$, $\theta_1 \neq 0$. The model under H_0 is the standard constant-coefficient (no cluster) regression; the model under H_1 has different intercepts but a common slope; the model under H_2 has different intercepts and different slopes; and the model under H_3 has a common intercept but different slopes. Among these four possible hypotheses, we will only consider H_0, H_1, and H_2 because, in a regression setting, the inference about slopes is generally of more interest than the intercept when evaluating the patterns of relationships between the response and the covariate relative to the background.

We consider a simultaneous test for the cluster effect in both the slopes and the intercepts:

$$H_0 : \theta_0 = \theta_1 = 0 \quad \text{versus} \quad H_2 : \theta_0 \neq 0,\ \theta_1 \neq 0. \tag{3}$$

Define a test statistic as $F = \left\{(\mathrm{SSE}_0 - \mathrm{SSE}_2)/2\right\} / \left\{\mathrm{SSE}_2/(N-4)\right\}$, where SSE_0 is the sum of squared errors (SSE) under H_0 equal to $\sum_{i=1}^N y_i^2 - \left(\sum_{i=1}^N \boldsymbol{x}_i y_i\right)^T \left(\sum_{i=1}^N \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \left(\sum_{i=1}^N \boldsymbol{x}_i y_i\right)$, and $\boldsymbol{x}_i$ is the ith covariate vector $(1, x_i)^T$. Further, SSE_2 is the SSE under H_2 equal to $\sum_{i=1}^N y_i^2 - \left(\sum_{i\in C} \boldsymbol{x}_i y_i\right)^T \left(\sum_{i\in C} \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \left(\sum_{i\in C} \boldsymbol{x}_i y_i\right) - \left(\sum_{i\notin C} \boldsymbol{x}_i y_i\right)^T \left(\sum_{i\notin C} \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \left(\sum_{i\notin C} \boldsymbol{x}_i y_i\right)$. Under H_0, the F statistic follows an F distribution with degrees of freedom $\mathrm{df}_1 = 2$ and $\mathrm{df}_2 = N - 4$.

Hypothesis testing involving the three hypotheses H_0, H_1, and H_2 will be further discussed in Section 3.

2.2. *Single Cluster*

In Section 2.1, a fixed cluster is assumed to be known *a priori*. Now, we relax this assumption and consider spatial cluster detection in the regression coefficients without assuming a fixed cluster. Let $\mathcal{C} = \{C_1, C_2, \dots\}$ denote the set of all possible clusters. For an unknown single cluster $C \in \mathcal{C}$, let

$$\mu_i = \begin{cases} \beta_0 + \beta_1 x_i & \text{if } i \notin C \\ (\beta_0 + \theta_{C,0}) + (\beta_1 + \theta_{C,1}) x_i & \text{if } i \in C \end{cases}, \tag{4}$$

where $\theta_{C,0}$ and $\theta_{C,1}$ are the cluster effect in the intercepts and in the slopes, respectively, of the cluster C.

For $C_k \in \mathcal{C}$, $k = 1, 2, \dots$, we first consider the null hypothesis H_0 versus a cluster specific local alternative hypothesis H_{C_k}:

$$H_0 : \theta_{C_k,0} = \theta_{C_k,1} = 0 \quad \text{versus} \quad H_{C_k} : \theta_{C_k,0} \neq 0,\ \theta_{C_k,1} \neq 0, \tag{5}$$

where $\theta_{C_k,0}$ and $\theta_{C_k,1}$ are the cluster effect in the intercepts and in the slopes, respectively, of the cluster C_k. For a given cluster C_k, this setting is the same as (3). Thus, an F test statistic can be defined as

$$F(C_k) = \left\{(\mathrm{SSE}_0 - \mathrm{SSE}_{C_k})/2\right\} / \left\{\mathrm{SSE}_{C_k}/(N-4)\right\}$$

and follows an F distribution with degrees of freedom $\mathrm{df}_1 = 2$ and $\mathrm{df}_2 = N - 4$ under H_0, where SSE_{C_k} is the SSE under H_{C_k}.

Next, we consider a global alternative hypothesis for an unknown generic cluster

$$H_A : \theta_{C,0} \neq 0,\ \theta_{C,1} \neq 0 \ \textit{for a cluster}\ \ C \in \mathcal{C}.$$

From the F test statistics for all the possible local hypotheses given in (5), we define the test statistic H_0 versus H_A to be

$$T = \max_{C\in\mathcal{C}} F(C). \tag{6}$$

To compute a p-value, a Monte Carlo method in the spirit of a parametric bootstrap is adopted. First, we compute the unbiased estimates of the parameters under H_0 and obtain $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$. Second, we generate Monte Carlo samples $y_i^{new} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i^{new}$, where $\varepsilon_i^{new} \sim iid\ N(0, \hat{\sigma}^2)$ for $i = 1, \dots, N$. Third, we compute the test statistic (6) for each Monte Carlo sample. Suppose there are S random Monte Carlo samples. The p-value is $R/(S+1)$, where R is the rank of the test statistic (6) for the original dataset in comparison with all the Monte Carlo samples, and the largest number acquires a rank of 1.

Copyright © 2016 John Wiley & Sons, Ltd.

The test statistic (6) is for all the possible clusters in $\mathcal{C} = \{C_1, C_2, \ldots\}$. Among those clusters, the cluster that corresponds to the test statistic T in (6) is considered to be the cluster estimate $\hat{C}$. That is,

$$\hat{C} = \arg\max_{C \in \mathcal{C}} F(C).$$

Here, the set of potential clusters, $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$, is pre-defined by circular clusters centered at the N sites in the data with various radii. We restrict the radius to be between 0 and a maximum radius, say $R_{\max}$. For a particular centroid of, say cell i, the potential clusters centered are chosen to have radii $0 = r_{i,1} < r_{i,2} < \ldots < r_{i,m_i} \leqslant R_{\max}$. Essentially, there are m_i distinct potential clusters with radii $r_{i,1}, r_{i,2}, \ldots, r_{i,m_i}$. With $K = \sum_{i=1}^{N} m_i < \infty$, there are a total of K potential clusters for the N cells.

The computational complexity and algorithm are described in Appendix A.

2.3. Multiple Clusters

To detect potential additional clusters, we propose a sequential algorithm. That is, we estimate the first cluster $\hat{C}_1 = \arg\max_{C \in \mathcal{C}} F(C)$, where $\mathcal{C}$ is pre-defined with N cells on the spatial lattice and the maximum radius is $R_{\max}$. To test $H_0 : \boldsymbol{\theta}_C = \mathbf{0}$ *for any cluster* $C \in \mathcal{C}$ versus $H_A : \boldsymbol{\theta}_C \neq \mathbf{0}$ *for a cluster* $C \in \mathcal{C}$ where $\boldsymbol{\theta}_C = (\theta_{C,0}, \theta_{C,1})^T$, the single cluster method in Section 2.2 is applied. Next, after removing the effect of $\hat{C}_1$ from the data, we estimate the second cluster $\hat{C}_2 = \arg\max_{C \in \mathcal{C}} F(C)$. To test $H_0 : \boldsymbol{\theta}_C = \mathbf{0}$ *for any cluster* $C \in \mathcal{C}$ versus $H_A : \boldsymbol{\theta}_C \neq \mathbf{0}$ *for a cluster* $C \in \mathcal{C}$, the single cluster method in Section 2.2 is again applied. Then, after removing the effect of $\hat{C}_2$ from the data again, we find the third cluster estimate $\hat{C}_3 = \arg\max_{C \in \mathcal{C}} F(C)$ and perform the single cluster test for $H_0 : \boldsymbol{\theta}_C = \mathbf{0}$ *for any cluster* $C \in \mathcal{C}$ versus $H_A : \boldsymbol{\theta}_C \neq \mathbf{0}$ *for a cluster* $C \in \mathcal{C}$, etc. In the end, a set of cluster estimates, $\{\hat{C}_1, \hat{C}_2, \hat{C}_3, \ldots\}$, is obtained. Because these cluster estimates are obtained sequentially, the corresponding p-values are also computed in a sequential fashion. The detailed algorithm has the following steps.

(1) Estimate the background coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$ under H_0 (no cluster) and compute the residuals $e_{0i} = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$.
(2) Pre-define $\mathcal{C}$ with N cells on the spatial lattice and the maximum radius $R_{\max}$.
(3) Obtain the cluster $\hat{C} = \arg\max_{C \in \mathcal{C}} F(C)$ with the residuals as the responses, its p-value, and corresponding coefficients $\hat{\boldsymbol{\theta}}_{\hat{C}} = (\hat{\theta}_{\hat{C},0}, \hat{\theta}_{\hat{C},1})^T$.
(4) Update the residuals by removing the cluster effect such as $e_{ji} = e_{(j-1)i} - \boldsymbol{x}_i^T \hat{\boldsymbol{\theta}}_{\hat{C}} \cdot I\{i \in \hat{C}\}$, where e_{ji}'s are the residuals from the model with the jth cluster and $I(\cdot)$ is the indicator function.
(5) Repeat steps 3–4 until p-value $> \alpha$. That is, stop only if the p-value in step 3 is greater than the significance level α.

The detected clusters using the sequential method above can overlap with each other. To obtain multiple non-overlapping clusters, we update the set of potential clusters for the jth cluster to be $\mathcal{C}_j = \mathcal{C} \setminus \bigcup_{k=1}^{j-1} \mathcal{K}_k$, where $\mathcal{K}_k$ is a set of clusters that overlap with the kth cluster estimate $\hat{C}_k$.

The previously proposed methodology for multiple clusters, overlapping or not, is based on F tests for the cluster effect in both the slopes ($\theta_{C,1}$) and the intercepts ($\theta_{C,0}$) of each potential cluster $C \in \mathcal{C}$. The detected clusters could have significant cluster effects in the intercepts only, or in both the slopes and the intercepts. Thus, we will refer to this cluster detection as the simultaneous detection to distinguish from an alternative sequential approach to be developed in the next section.

3. Two–Stage Spatial Cluster Detection in Intercepts and Slopes

In a regression setting, inference about a slope is generally of more interest than the intercept. The test statistic (6) allows the detection of spatial clusters in both the slopes and the intercepts, but it is not straightforward to determine whether the cluster effects are in the slopes or in the intercepts. To study the potential spatial pattern in the slopes, we now develop an alternative, two-stage approach to detecting multiple clusters. In particular, spatial clusters in the slopes will be detected in the first stage regardless of intercept effect. Then, in the second stage, spatial clusters in the intercepts will be detected. Henceforth, this alternative approach will be referred to as the two-stage detection.

Copyright © 2016 John Wiley & Sons, Ltd.

3.1. Test for Spatial Cluster Effects in a Simplified Setting

Assume model (2) with a fixed cluster C which is known *a priori*. We perform hypotheses testing in two steps: first the cluster effect in the slopes and then the cluster effect in the intercepts. That is,

$$H_1 : \theta_0 \neq 0,\ \theta_1 = 0 \text{ versus } H_2 : \theta_0 \neq 0,\ \theta_1 \neq 0, \tag{7}$$
$$H_0 : \theta_0 = \theta_1 = 0 \text{ versus } H_1 : \theta_0 \neq 0,\ \theta_1 = 0. \tag{8}$$

The test statistics for (7) and (8) are, respectively,

$$F^{\text{slope}} = (\text{SSE}_1 - \text{SSE}_2)/\left\{\text{SSE}_2/(N-4)\right\},$$
$$F^{\text{int}} = (\text{SSE}_0 - \text{SSE}_1)/\left\{\text{SSE}_1/(N-3)\right\},$$

where SSE_1 is the SSE under H_1 and equivalent to $\sum_{i=1}^N y_i^2 - \left(\sum_{i=1}^N \boldsymbol{w}_i y_i\right)^T \left(\sum_{i=1}^N \boldsymbol{w}_i \boldsymbol{w}_i^T\right)^{-1} \left(\sum_{i=1}^N \boldsymbol{w}_i y_i\right)$, and $\boldsymbol{w}_i$ is defined as the column vector $(1, x_i, 1)^T$ for $i \in C$ and $(1, x_i, 0)^T$ for $i \notin C$. Under H_1, the test statistic F^{slope} follows an F distribution with degrees of freedom $\text{df}_1 = 1$ and $\text{df}_2 = N - 4$, whereas the test statistic F^{int} follows an F distribution with degrees of freedom $\text{df}_1 = 1$ and $\text{df}_2 = N - 3$ under H_0.

3.2. First Stage: Spatial Cluster in the Slopes

From now, we assume model (4). Of interest is the cluster effect in the slopes ($\theta_{C,1}$) for an unknown single cluster $C \in \mathcal{C}$. For $C_k \in \mathcal{C}$, $k = 1, 2, \ldots$, we first consider the null hypothesis H_0^{slope} versus a cluster specific local alternative hypothesis $H_{C_k}^{\text{slope}}$ for the slopes:

$$H_0^{\text{slope}} : \theta_{C_k,1} = 0 \text{ versus } H_{C_k}^{\text{slope}} : \theta_{C_k,1} \neq 0. \tag{9}$$

For a given cluster C_k, this setting is the same as (7). Thus, we define

$$F^{\text{slope}}(C_k) = (\text{SSE}_{0,\text{slope}} - \text{SSE}_{C_k,\text{slope}})/\left\{\text{SSE}_{C_k,\text{slope}}/(N-4)\right\}. \tag{10}$$

The test statistic $F^{\text{slope}}(C_k)$ in (10) follows an F distribution with degrees of freedom $\text{df}_1 = 1$ and $\text{df}_2 = N - 4$ under H_0^{slope}, where $\text{SSE}_{0,\text{slope}}$ and $\text{SSE}_{C_k,\text{slope}}$ are the SSEs under H_0^{slope} and $H_{C_k}^{\text{slope}}$, respectively.

As in the simultaneous method, we consider a global alternative hypothesis

$$H_A^{\text{slope}} : \theta_{C,1} \neq 0 \text{ for a cluster } C \in \mathcal{C}$$

for an unknown generic cluster. From the F test statistics for all the possible local hypotheses given in (9), we define the test statistic for H_0^{slope} versus H_A^{slope} and the corresponding cluster estimate to be

$$T^{\text{slope}} = \max_{C \in \mathcal{C}} F^{\text{slope}}(C), \tag{11}$$

$$\hat{C} = \arg\max_{C \in \mathcal{C}} F^{\text{slope}}(C). \tag{12}$$

To compute a p-value, a Monte Carlo method is applied in a manner similar to Section 2.2.

To detect potential additional clusters in the slopes, we propose a sequential algorithm with the cluster estimate (12). That is, we estimate the first cluster $\hat{C}_1 = \arg\max_{C \in \mathcal{C}} F^{\text{slope}}(C)$. Then, we iteratively estimate the next cluster $\hat{C}_{j+1} = \arg\max_{C \in \mathcal{C}} F^{\text{slope}}(C)$ after removing the effect of $\hat{C}_j$ from the data, where $j = 1, 2, \ldots$. The iteration continues until there is not any more significant cluster in the slopes. Then, we move to the second stage to find clusters in the intercepts.

Copyright © 2016 John Wiley & Sons, Ltd.

3.3. Second Stage: Spatial Cluster in the Intercepts

In the second stage, of interest is the cluster effect in the intercepts ($\theta_{C,0}$), for an unknown single cluster $C \in \mathcal{C}$. Thus, a varying-intercept but constant-slope model is considered.

For $C_k \in \mathcal{C}$, $k = 1, 2, \ldots$, we first consider the null hypothesis H_0^{int} versus a cluster specific local alternative hypothesis $H_{C_k}^{\text{int}}$ for the intercepts:

$$H_0^{\text{int}} : \theta_{C_k,0} = \theta_{C_k,1} = 0 \quad \text{versus} \quad H_{C_k}^{\text{int}} : \theta_{C_k,0} \neq 0,\ \theta_{C_k,1} = 0. \tag{13}$$

For a given cluster C_k, this setting is the same as (8). Thus, an F test statistic can be defined as $F^{\text{int}}(C_k) = (\text{SSE}_0 - \text{SSE}_{C_k,\text{int}}) / \left\{\text{SSE}_{C_k,\text{int}}/(N-3)\right\}$ and follows an F distribution with degrees of freedom $\text{df}_1 = 1$ and $\text{df}_2 = N - 3$ under H_0^{int}, where $\text{SSE}_{C_k,\text{int}}$ is the SSE under $H_{C_k}^{\text{int}}$.

Next, we consider a global alternative hypothesis for an unknown generic cluster

$$H_A^{\text{int}} : \theta_{C,0} \neq 0 \ \textit{for a cluster}\ \ C \in \mathcal{C}.$$

From the F test statistics for all the possible local hypotheses given in (13), we define the test statistic for H_0^{int} versus H_A^{int} and corresponding cluster estimate to be

$$T^{\text{int}} = \max_{C \in \mathcal{C}} F^{\text{int}}(C), \tag{14}$$

$$\hat{C} = \arg\max_{C \in \mathcal{C}} F^{\text{int}}(C). \tag{15}$$

The p-value of the test statistic (14) is again computed via a Monte Carlo method.

Suppose a total of J_1 significant clusters in the slopes are detected in the first stage. Then, in the second stage, we could consider a sequential algorithm with the cluster estimate (15) to detect potential additional clusters in the intercepts. That is, after removing the effects of $\{\hat{C}_1, \ldots, \hat{C}_{J_1}\}$ from the data, we estimate the (J_1+1)th cluster $\hat{C}_{J_1+1} = \arg\max_{C \in \mathcal{C}} F^{\text{int}}(C)$. We again estimate the next cluster $\hat{C}_{J_1+2} = \arg\max_{C \in \mathcal{C}} F^{\text{int}}(C)$ after removing the effect of $\hat{C}_{J_1+1}$, and so on and so forth. In the end, a set of cluster estimates, $\{\hat{C}_1, \hat{C}_2, \hat{C}_3, \ldots\}$, is identified, where the first set of cluster estimates $\{\hat{C}_1, \ldots, \hat{C}_{J_1}\}$ is the effect in the slopes while the second set $\{\hat{C}_{J_1+1}, \hat{C}_{J_1+2}, \ldots\}$ is the effect in the intercepts.

For multiple non-overlapping clusters, we update the set of potential clusters for the jth cluster to be $\mathcal{C}_j = \mathcal{C} \setminus \bigcup_{k=1}^{j-1} \mathcal{K}_k$, where $\mathcal{K}_k$ is a set of clusters that overlap with the kth cluster estimate $\hat{C}_k$.

4. Simulation Study

We conducted a simulation study to evaluate our previous methodology for a single cluster or two clusters that have either overlapping or non-overlapping cells. We consider a 25×25 square grid in the unit square $[0, 1] \times [0, 1]$, which is partitioned into 625 cells with 25 rows and 25 columns. The width of each cell is $1/25 = 0.04$. The centroids of the cells are $\{0.02, 0.06, \ldots, 0.98\} \times \{0.02, 0.06, \ldots, 0.98\}$. The set of potential clusters consists of 41,493 circular clusters centered at the 625 cell centroids with radii ranging from 0 to 0.2. The single covariate, x, follows the standard normal distribution. The regression coefficients in the background are set to be $\boldsymbol{\beta} = (\beta_0, \beta_1)^T = (0, 0)^T$, and the variance of the random error ε_i is set to be $\sigma^2 = 1$. We will evaluate the power of the cluster detection tests in a single cluster setting and will evaluate the coverage of the true clusters in a two-cluster setting.

4.1. Evaluation of Power of Tests

For a single true cluster detection, we define power to be the proportion of simulations in which the global null hypothesis, $H_0 : \boldsymbol{\theta}_C = (\theta_{C,0}, \theta_{C,1})^T = (0, 0)^T$ *for any cluster* $C \in \mathcal{C}$, is rejected at the significance level α. There are different ways to define power for cluster detection tests in the literature, incorporating different views on how to define a correct cluster identification. However, the different definitions of power do not have much impact on the results [4, 9, 23, 24].

Here, we consider a total of nine different circular clusters which are defined by nine centroids and the same radius of 3/25 unit. One centroid is at the center (0.50, 0.50) of the unit square, four centroids are away from the center to the bottom, and the other fours are away from the center to the lower left

Copyright © 2016 John Wiley & Sons, Ltd.

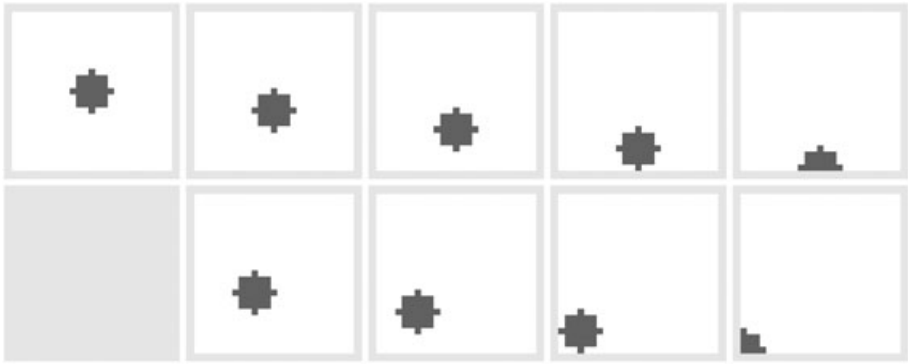

Figure 1. The nine-cluster settings with different centroids and the same radius of 3/25 unit for evaluation of power of detecting true clusters.

Table I. Power in percentage for cluster detection on the 25 × 25 square grid with the max cluster radius $R_{\max} = 1/5$. The error standard deviation is $\sigma = 1$.

		Signal-to-noise ratio (SNR: θ/σ)		
Centroid	Cells	2	1	1/2
(0.50, 0.50)	29	100.0	99.0	23.0
(0.50, 0.38)	29	100.0	99.0	22.5
(0.50, 0.26)	29	100.0	99.0	22.8
(0.50, 0.14)	29	100.0	99.0	24.1
(0.50, 0.02)	18	100.0	77.5	11.1
(0.38, 0.38)	29	100.0	99.0	22.6
(0.26, 0.26)	29	100.0	99.0	23.0
(0.14, 0.14)	29	100.0	99.1	23.4
(0.02, 0.02)	11	100.0	48.9	8.3

corner. A complete circular cluster consists of 29 cells. The half circular cluster with a centroid at the bottom, (0.05, 0.02), has 18 cells, whereas the quarter circular cluster with a centroid at the lower left corner, (0.02, 0.02), has only 11 cells. These cluster settings are illustrated in Figure 1. The cluster effect in the slope is set to be the same as in the intercept. That is $\boldsymbol{\theta} = (\theta, \theta)^T$ where θ is set to be 2, 1, or 1/2 for strong, medium, or weak cluster effect, respectively, relative to the error standard deviation $\sigma = 1$. We simulated 1000 datasets for the different combinations of centroids and cluster effects θ.

We identified the critical value of the test statistic (6), by the null distribution, which was generated from 10,000 null simulations, at $\alpha = 0.05$ with the max radius 1/5 unit. We used this critical value to test the detected cluster in each simulated dataset. The simultaneous detection, developed in Section 2, was used to find a significant cluster.

Table I provides the results of the power calculation for each simulation setting. Our cluster detection method has a 100% power when the signal-to-noise ratio (SNR: θ/σ) is 2 even for a half or a quarter circular cluster. With SNR 1, the power is around 99% for complete circular clusters, 78% for half circles with 18 cells, and 49% for quarter circles.

4.2. Evaluation of Coverage of the True Clusters

For two true clusters, we evaluated the coverage of detected clusters. We considered a total of three different two cluster settings. The two circular clusters have the same radius 3/25 unit. The two clusters are adjacent each other in the first setting and are apart in the second setting. The third setting has two overlapping clusters. These three cluster settings are illustrated in Figure 2. Further, we set two different scenarios for the cluster effects, one such that the cluster effects are in the slopes and the intercepts for each cluster and the other such that the cluster effects are in the slopes and the intercepts for one cluster, while there is the cluster effect in the intercepts only for the second cluster. The cluster effect is set to

Copyright © 2016 John Wiley & Sons, Ltd.

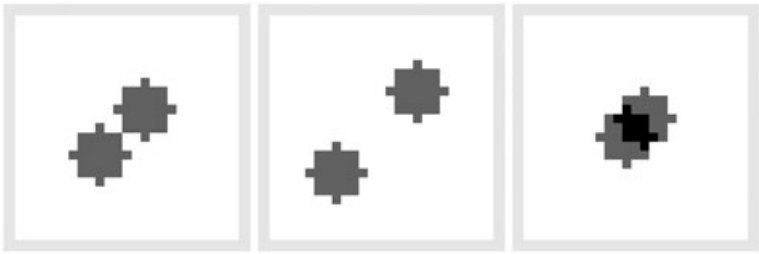

Figure 2. Two clusters are adjacent to, apart from and overlapping with each other, respectively, with the same radius of 3/25 unit for evaluation of coverage of true clusters.

be $\theta = 2$. That is, $\boldsymbol{\theta}_{C_1} = \boldsymbol{\theta}_{C_2} = (2,2)^T$ in the first scenario, and $\boldsymbol{\theta}_{C_1} = (2,2)^T$ and $\boldsymbol{\theta}_{C_2} = (2,0)^T$ in the second scenario. We simulated 1000 datasets for a total of six different combinations of cluster settings and cluster effect scenarios. For each simulated dataset, we estimated the regression coefficients for the detected clusters, and we mapped the mean coefficient estimates in comparison with the true values.

To detect clusters, we applied four methods: simultaneous detection or two-stage detection with non-overlapping or overlapping clusters. We used the critical values for the test statistics (6), (11), and (14) for testing in each simulated dataset. The null distribution of each test statistic was generated from 10,000 null simulations, at $\alpha = 0.05$ of the max radius 1/5 unit.

Figure 3 provides the maps of the mean coefficient estimates based on each of the four cluster detection methods for the simulated data with two true overlapping clusters. Columns 1 and 3 are for the mean slope estimates, whereas columns 2 and 4 are for the mean intercept estimates. In the first two columns, $\boldsymbol{\theta}_{C_1} = \boldsymbol{\theta}_{C_2} = (2,2)^T$. In the last two columns, $\boldsymbol{\theta}_{C_1} = (2,2)^T$ and $\boldsymbol{\theta}_{C_2} = (2,0)^T$. Row 1 is the oracle, namely, the true coefficients. Rows 2 and 3 are from the simultaneous detection method with non-overlapping or overlapping clusters. Rows 4 and 5 are from the two-stage detection method with non-overlapping or overlapping clusters. The results for the other two cluster settings, adjacent or apart, are omitted because the findings are similar in the sense that all the cluster detection methods perform well and the corresponding mean coefficient estimates are close to true clusters and the true regression coefficients.

Figure 3 shows that, when true clusters overlap with each other, it is hard to identify all of the true clusters under the non-overlapping clusters assumption while the results under the overlapping assumption indicate clusters that are close to the truth. Thus, detecting clusters under the overlapping assumption seems to be the safer choice for identifying true clusters, whether overlapping or not. However, the overlapping assumption requires more computation to detect multiple clusters than the non-overlapping assumption. The set of potential clusters for the jth cluster could be $\mathcal{C} \setminus \{\hat{C}_1, \ldots, \hat{C}_{j-1}\}$ when we assume overlapping clusters, while that is $\mathcal{C} \setminus \bigcup_{k=1}^{j-1} \mathcal{K}_k$ under the non-overlapping assumption, where $\mathcal{K}_k$ is a set of clusters that overlap with the kth cluster estimate $\hat{C}_k$. We have more potential clusters to examine under the overlapping assumption, $|\mathcal{C} \setminus \{\hat{C}_1, \ldots, \hat{C}_{j-1}\}| - |\mathcal{C} \setminus \bigcup_{k=1}^{j-1} \mathcal{K}_k| = |\bigcup_{k=1}^{j-1} \mathcal{K}_k| - (j-1)$, where $|\cdot|$ denotes the cardinality of a set. Further, this difference in the number of potential clusters, $|\bigcup_{k=1}^{j-1} \mathcal{K}_k| - (j-1)$, increases as j increases. That is, overlapping assumption requires more computation as the number of clusters increases. In our simulation study, identifying clusters under the overlapping assumption is about 10% slower than the non-overlapping assumption in both of the simultaneous detection and the two-stage detection.

Under the overlapping assumption, both of the simultaneous detection and the two-stage detection show similar performances in terms of identifying true clusters in Figure 3. Because the two-stage detection is more computationally intensive, the simultaneous detection is appealing.

5. Data Example

5.1. Southeast U.S.A Cancer Mortality Data

The dataset comprises 616 counties in seven U.S. states: Alabama, Florida, Georgia, Mississippi, North Carolina, South Carolina, and Tennessee. For each county, the cancer mortality rate is defined as the number of deaths of cancer patients per 100,000 population per year over 2008–2012 and age-adjusted to the 2000 U.S standard population (http://www.statecancerprofiles.cancer.gov/). In addition, the dataset contains information about the extent of urban versus rural areas in terms of the proportion of the population in urban areas in census year 2000 (http://www.census.gov/). We considered regression models with the

Copyright © 2016 John Wiley & Sons, Ltd.

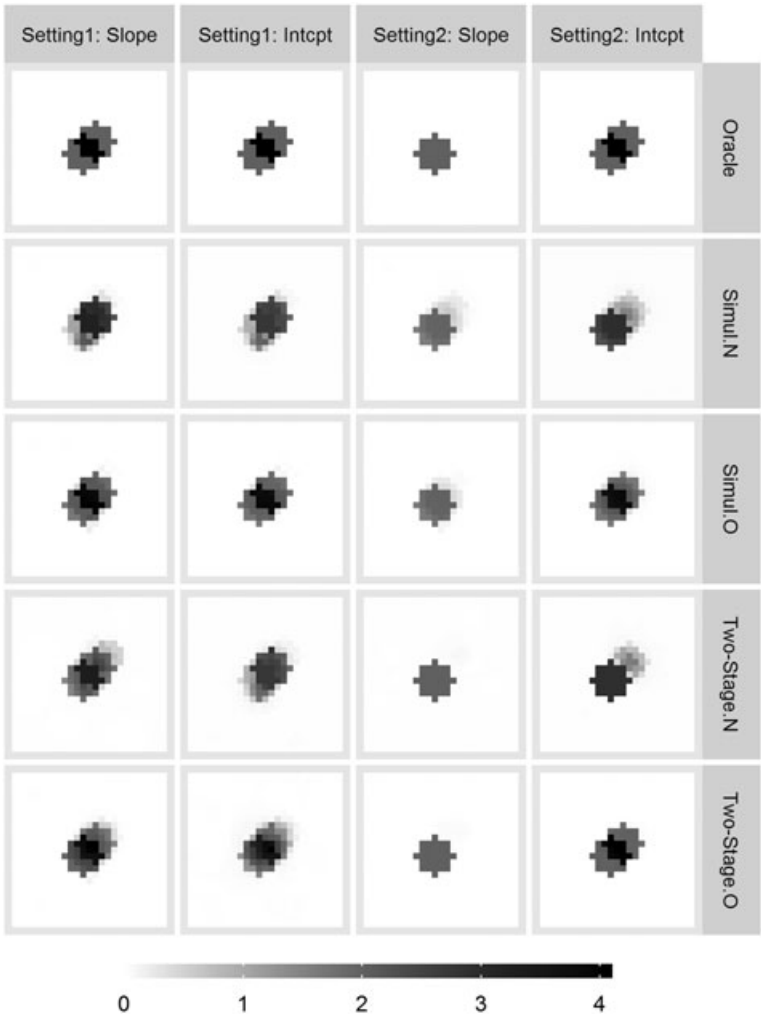

Figure 3. Maps of the mean coefficient estimates for each cell from the 1000 simulated datasets with two overlapping clusters and in the first two columns, $\boldsymbol{\theta}_{C_1} = \boldsymbol{\theta}_{C_2} = (2,2)^T$ and in the last two columns, $\boldsymbol{\theta}_{C_1} = (2,2)^T$ and $\boldsymbol{\theta}_{C_2} = (2,0)^T$. Row 1 is the oracle. Rows 2 and 3 are simultaneous detection with non-overlapping and overlapping clusters. Rows 4 and 5 are two-stage detection with non-overlapping and overlapping clusters.

log cancer mortality rate (`logMortality`) as the response variable and the proportion of the population in urban areas (`purban`) as the covariate. For $y_i = \log r_i$, where r_i is the rate for the ith county, it can be shown that $Var(y_i) \approx (n_i\rho_i)^{-1} + \sigma^2$, where n_i is the county population and $\rho_i = E(r_i)$. For county populations in the thousands, the first term $(n_i\rho_i)^{-1}$ is negligible, and thus, we assume a constant variance. In addition, the residuals did not provide evidence of clusters based on spatial scan statistics or nonnormality. Thus, the assumption of independent errors seems reasonable. The slope estimate of the ordinary regression with no cluster is -0.096 with its standard error of 0.018. Thus, the overall trend shows that there is a negative relationship between cancer mortality and proportion of urban area.

The map of the log cancer mortality rate in Figure 4 shows that Union county in northern Florida has the highest log cancer mortality rate of nearly 6.00. In addition, some highly urbanized counties such as Fulton county in northern Georgia and Hillsborough and Miami-Dade counties in southern Florida have relatively low cancer mortality rates. There is no other obvious geographical clusters of the cancer mortality rate in relation to proportion of urban area.

The result of GWR are mapped in Figure 5, where the log cancer mortality rate and the proportion of the population in urban areas are the response and the covariate, respectively. It appears that there are several potential geographical clusters in the relationship between cancer mortality and proportion of urban area: negative relationship in central Florida, coastal South Carolina and central Tennessee; positive relationship in northern Mississippi; no relationship in southern Florida and North Carolina. But, still, it is not clear how to delineate clusters and interpret the corresponding regression coefficients

Copyright © 2016 John Wiley & Sons, Ltd.

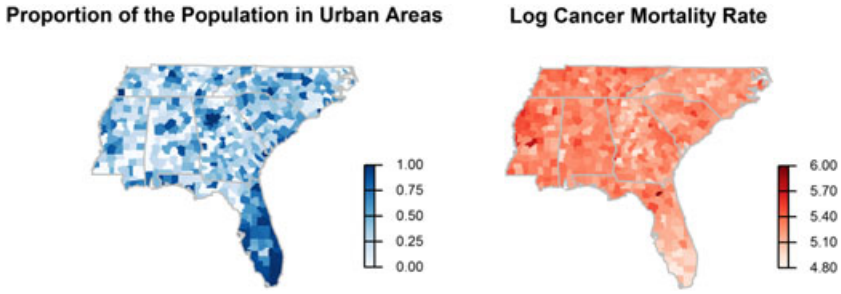

Figure 4. The proportion of the population in urban areas and the log cancer mortality rate for each county in the states of Alabama, Florida, Georgia, Mississippi, North Carolina, South Carolina, and Tennessee.

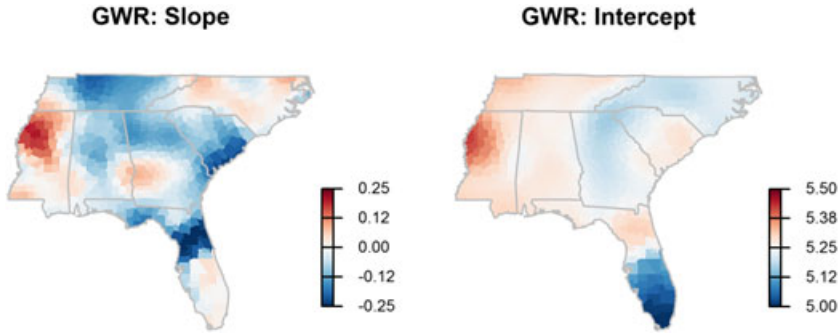

Figure 5. Coefficients Estimates from the geographically weighted regression (GWR).

Table II. Detected clusters (1) via the simultaneous cluster detection method at $\alpha = 0.05$, and (2) via the two-stage cluster detection method at $\alpha = 0.05$. The response is the log cancer mortality rate and the covariate is the proportion of the population in urban areas in a county. In the two-stage cluster detection's result, clusters $\hat{C}_3$ and $\hat{C}_5$ share one common county.

	(1) Simultaneous detection				(2) Two-stage detection				
C	Centroid	Radius	Counties	p-value	Centroid	Radius	Counties	p-value	Stage
$\hat{C}_1$	Sunflower, MS	122	22	0.001	Calhoun, MS	176	58	0.008	1st
$\hat{C}_2$	Union, FL	31	3	0.002	Columbia, FL	58	8	0.013	1st
$\hat{C}_3$	Habersham, GA	95	33	0.001	Habersham, GA	95	33	0.001	2nd
$\hat{C}_4$	Glades, FL	214	28	0.002	Glades, FL	214	28	0.001	2nd
$\hat{C}_5$	Peach, GA	128	59	0.001	Monroe, GA	101	39	0.006	2nd
$\hat{C}_6$	Person, NC	251	79	0.003	–	–	–	–	–

estimates. However, we could identify multiple clusters using our proposed methodology. The covariate is centered to have a zero mean in both of GWR and our methods. The set of potential clusters consists of 93,450 circular clusters centered at the 616 county centroids with radii ranging from 0 to 300 km. We detected multiple clusters by the simultaneous detection in Section 2 and the two-stage detection in Section 3 in terms of relations between the log cancer mortality rate and the proportion of the population in urban areas. We assumed overlapping clusters because the simulation results in Section 4.2 showed that the coverage of the true clusters under the overlapping assumption is better than those under the non-overlapping assumption even though its computation is somewhat slower. The p-values were obtained from 1000 Monte Carlo samples. The maximum radius for a potential cluster is set to be $R_{\max} = 300$ km because the largest circular cluster with $R_{\max}$ is large enough to cover all or the majority of each of the seven states.

5.2. *Simultaneous Detection*

Table II's left panel and Table III's top panel provide the significant clusters and the corresponding coefficient estimates that were detected via the simultaneous detection method at $\alpha = 0.05$. There are a total

Copyright © 2016 John Wiley & Sons, Ltd.

Table III. Coefficients estimates for sequentially detected clusters (1) via the simultaneous cluster detection method at $\alpha = 0.05$, and (2) via the two-stage cluster detection method at $\alpha = 0.05$. The response is the log cancer mortality rate and the covariate is the proportion of the population in urban areas in a county.

	Number of $\hat{C}_j$	0	1	2	3	4	5	6
(1) Simultaneous detection	$\hat{\beta}_0$	5.242	5.234	5.233	5.240	5.246	5.253	5.260
	$\hat{\beta}_1$	-0.096	-0.105	-0.106	-0.115	-0.083	-0.096	-0.113
	$\hat{\theta}_{\hat{C}_1,0}$	–	0.213	0.213	0.213	0.213	0.213	0.213
	$\hat{\theta}_{\hat{C}_1,1}$	–	0.261	0.261	0.261	0.261	0.261	0.261
	$\hat{\theta}_{\hat{C}_2,0}$	–	–	0.143	0.143	0.143	0.143	0.143
	$\hat{\theta}_{\hat{C}_2,1}$	–	–	5.223	5.223	5.223	5.223	5.223
	$\hat{\theta}_{\hat{C}_3,0}$	–	–	–	-0.123	-0.123	-0.123	-0.123
	$\hat{\theta}_{\hat{C}_3,1}$	–	–	–	0.002	0.002	0.002	0.002
	$\hat{\theta}_{\hat{C}_4,0}$	–	–	–	–	-0.185	-0.185	-0.185
	$\hat{\theta}_{\hat{C}_4,1}$	–	–	–	–	0.104	0.104	0.104
	$\hat{\theta}_{\hat{C}_5,0}$	–	–	–	–	–	-0.074	-0.074
	$\hat{\theta}_{\hat{C}_5,1}$	–	–	–	–	–	0.146	0.146
	$\hat{\theta}_{\hat{C}_6,0}$	–	–	–	–	–	–	-0.059
	$\hat{\theta}_{\hat{C}_6,1}$	–	–	–	–	–	–	0.167
(2) Two-stage detection	$\hat{\beta}_0$	5.242	5.235	5.234	5.240	5.246	5.253	–
	$\hat{\beta}_1$	-0.096	-0.115	-0.118	-0.127	-0.094	-0.092	–
	$\hat{\theta}_{\hat{C}_1,0}$	–	0.096	0.096	0.096	0.096	0.096	–
	$\hat{\theta}_{\hat{C}_1,1}$	–	0.361	0.361	0.361	0.361	0.361	–
	$\hat{\theta}_{\hat{C}_2,0}$	–	–	0.274	0.274	0.274	0.274	–
	$\hat{\theta}_{\hat{C}_2,1}$	–	–	1.286	1.286	1.286	1.286	–
	$\hat{\theta}_{\hat{C}_3,0}$	–	–	–	-0.125	-0.125	-0.125	–
	$\hat{\theta}_{\hat{C}_4,0}$	–	–	–	–	-0.135	-0.135	–
	$\hat{\theta}_{\hat{C}_5,0}$	–	–	–	–	–	-0.096	–

of six clusters with no overlapping region. The maps of the coefficients estimates for the detected clusters are given in Figure 6, and corresponding scatter plots are illustrated in Figure 7. Different clusters have different coefficient estimates. Each cluster has a different slope and intercept from the background except the third cluster $\hat{C}_3$ that covers Georgia, North Carolina, and South Carolina. This third cluster differs from the background in the intercept but not quite in the slope. The slopes are negative in the background and in the third cluster $\hat{C}_3$ but are positive in the first two clusters, $\hat{C}_1$ and $\hat{C}_2$. Further, the slopes are close to zero in the last three clusters, $\hat{C}_4$, $\hat{C}_5$, and $\hat{C}_6$. The negative slopes, in the background and in $\hat{C}_3$, suggest a negative association between cancer mortality and proportion of urban areas. Among the clusters with almost zero slopes, southern Florida ($\hat{C}_4$), central Georgia ($\hat{C}_5$), and most of North Carolina and several counties of South Carolina ($\hat{C}_6$), have lower intercepts than the background. The cluster in northwestern Mississippi ($\hat{C}_1$) has the distinct pattern of a positive slope and a higher intercept than the background. In this cluster, there are 0% urban area (`purban` = 0) in the least urbanized county, while 83% urban area (`purban` = 0.829) in the most urbanized county. In addition, the difference in the fitted log cancer mortality rates between these two counties, $\hat{y}(x_{\max}) - \hat{y}(x_{\min})$, is 0.123 while the difference is -0.080 when the ordinary regression with no cluster is considered. A small cluster, which consists of three counties in northern Florida ($\hat{C}_2$), has a positive, but steep slope, possibly due to Union county that has the highest cancer mortality rate. In $\hat{C}_2$, there are 35% urban area (`purban` = 0.349) and 47% urban area (`purban` = 0.474) in the least urbanized county and the most urbanized county, respectively. These two

Copyright © 2016 John Wiley & Sons, Ltd.

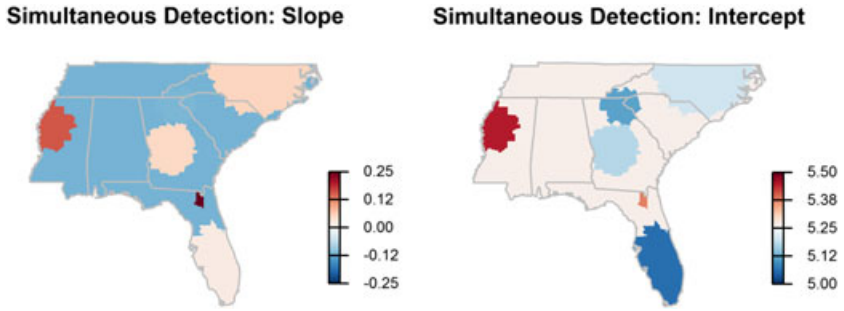

Figure 6. Coefficients estimates with overlapping clusters that were significant at $\alpha = 0.05$ via the simultaneous cluster detection.

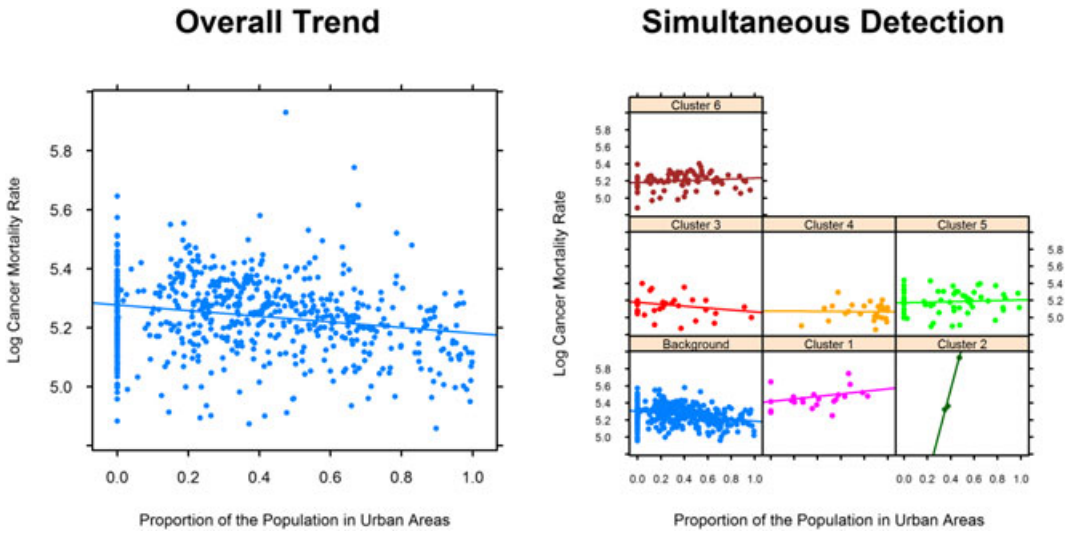

Figure 7. Scatter plots with fitted regression lines with overlapping clusters which were significant at $\alpha = 0.05$ via the simultaneous cluster detection.

counties show the difference of 0.639 in the fitted log cancer mortality rates while that is -0.012 from the ordinary regression with no cluster.

5.3. Two-Stage Detection

Table II's right panel and Table III's bottom panel provide the significant clusters and the corresponding coefficient estimates that were detected via the two-stage detection method at $\alpha = 0.05$. There are a total of five detected clusters with one overlapping region. The maps of the coefficients estimates for the detected clusters are given in Figure 8, and corresponding scatter plots are illustrated in Figure 9. The first two detected clusters, $\hat{C}_1$ and $\hat{C}_2$, are significant in the slopes, and the next three clusters, $\hat{C}_3$–$\hat{C}_5$, are significant in the intercepts only. A big cluster in North Carolina, which was significant in the simultaneous detection, is not identified via the two-stage detection. Other than that, however, the detected clusters are quite similar to those from the simultaneous detection. The first cluster ($\hat{C}_1$) is centered at a county in Mississippi, and the second cluster ($\hat{C}_2$) is in northern Florida including the Union county. The third cluster ($\hat{C}_3$) covers Georgia, North Carolina, and South Carolina and shares one county (Oconee county, Georgia) with another cluster in central Georgia ($\hat{C}_5$). There is also a cluster in southern Florida ($\hat{C}_4$). In Figure 8, the first map shows two clusters that have different slopes from the background, while the second map indicates that all the clusters have different intercept estimates. The two clusters in northern Mississippi with several counties of Alabama and Tennessee ($\hat{C}_1$), and in northern Florida with a county of Georgia ($\hat{C}_2$), have positive slopes and higher intercepts than the background. In $\hat{C}_1$, there are 0% urban area (`purban` = 0) in the least urbanized county while 97% urban area (`purban` = 0.967) in the most urbanized county. In addition, the difference in the fitted log cancer mortality rates between these two counties, $\hat{y}(x_{\max}) - \hat{y}(x_{\min})$, is 0.260 while the difference is -0.093 when the ordinary regression with no

Copyright © 2016 John Wiley & Sons, Ltd.

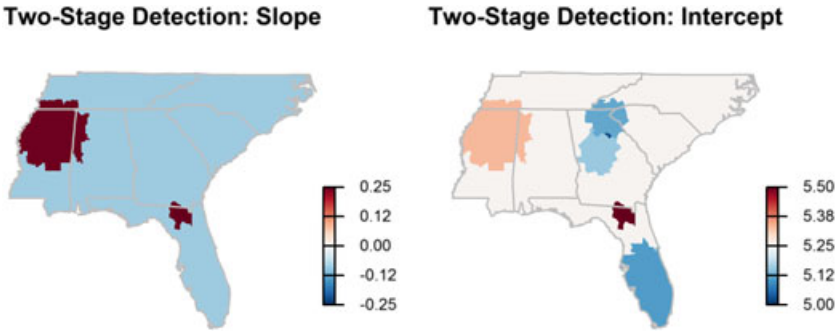

Figure 8. Coefficients estimates with overlapping clusters that were significant at $\alpha = 0.05$ via the two-stage cluster detection.

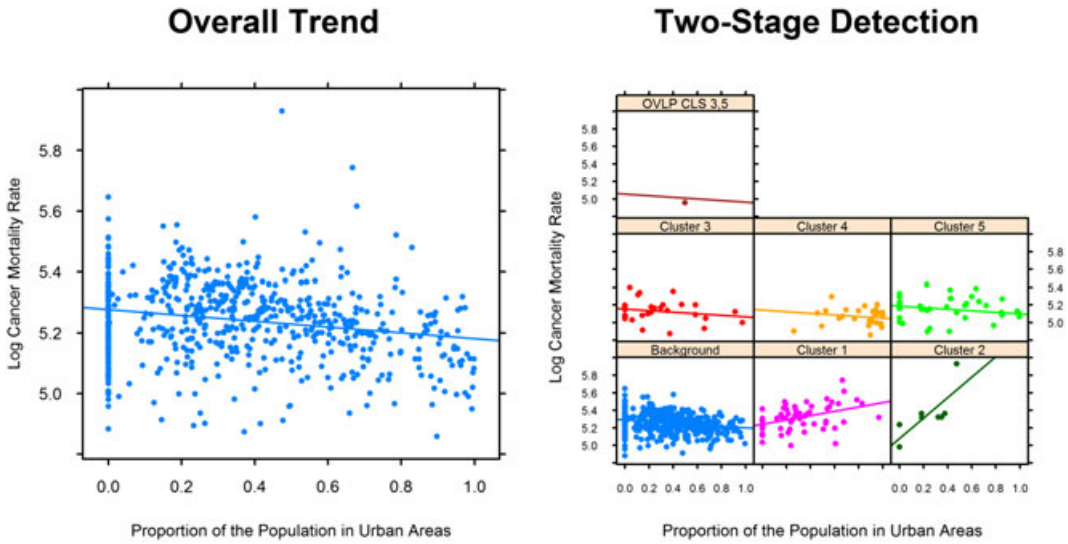

Figure 9. Scatter plots with fitted regression lines with overlapping clusters that were significant at $\alpha = 0.05$ via the two-stage cluster detection. The third and the fifth clusters share one common county, Oconee county in Georgia (OVLP CLS 3,5).

cluster is considered. In $\hat{C}_2$, there are 0% urban area (`purban` = 0) and 47% urban area (`purban` = 0.474) in the least urbanized county and the most urbanized county, respectively. These two counties show the difference of 0.566 in the fitted log cancer mortality rates, while the difference is -0.045 from the ordinary regression with no cluster. The other three clusters, $\hat{C}_4$, $\hat{C}_5$, and $\hat{C}_6$, have lower intercepts than the background, while they have the same negative slopes as the background.

6. Conclusions and Discussion

We have developed in this paper a new methodology to detect spatial clusters in the regression coefficients. Both the simultaneous detection and the two-stage detection methods can be used to find geographic regions that have different relationship between a response variable and a covariate in a varying-coefficient regression setting. Although it is a common practice to use circular clusters as we have performed here, our methods can be modified to consider other shapes, such as ellipses and squares (e.g., [5–7]).

Our simulation study, which evaluated the power and the coverage of true clusters, suggests satisfactory performance of both methods. The simultaneous detection method is faster to compute than the two-stage detection. In the simultaneous cluster detection, the regression coefficient estimates are obtained for both the intercepts and the slopes in any detected cluster. However, some of the slope estimates may not differ significantly from the background. In contrast, the two-stage detection produces slope estimates for only those clusters that have the slope estimates significantly different from the background. For those

Copyright © 2016 John Wiley & Sons, Ltd.

clusters, in which only the intercept is significantly different from the background but not the slope, only the intercept estimates are reported. Because this latter method consists of two separate stages, it is slower to compute than the simultaneous detection.

The simultaneous cluster detection and the two-stage cluster detection methods provide different results, but qualitatively the interpretation in both the locations and the coefficient estimates of the clusters is similar. Thus, between the two methods, we may choose one that is more suitable for the application.

For further research, we will consider more than one covariate. While the simultaneous detection can be readily extended to a multiple regression model, it is not easy to derive a multiple stage detection method from the two-stage detection, as the computational time increases greatly with more covariates.

Appendix A: Computational Aspects

A.1 Computational Complexity

The test statistic $T = \max_{C\in\mathcal{C}} F(C)$ in (6) is based on F statistics for the local hypotheses for all the possible clusters in $\mathcal{C} = \{C_1, C_2, \ldots\}$. We consider a multiple regression model with $(p-1)$ covariates such that $\boldsymbol{x}_i = (1, x_{1i}, \ldots, x_{(p-1)i})^T$. Then, for a given cluster C_k, $F(C_k)$ is defined as $F(C_k) = \left\{(\mathrm{SSE}_0 - \mathrm{SSE}_{C_k})/p\right\} / \left\{\mathrm{SSE}_{C_k}/(N-2p)\right\}$. Thus, while a single calculation of SSE_0 is enough because SSE_0 is identical for all C_k, SSE_{C_k} needs to be calculated for every given cluster C_k, which can be time consuming. Thus, we rewrite SSE_{C_k} as

$$\begin{aligned}\mathrm{SSE}_{C_k} = &\sum_{i=1}^{N} y_i^2 - \left(\sum_{i\in C_k} \boldsymbol{x}_i y_i\right)^T \left(\sum_{i\in C_k} \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \left(\sum_{i\in C_k} \boldsymbol{x}_i y_i\right) \\ &- \left(\sum_{i=1}^{N} \boldsymbol{x}_i y_i - \sum_{i\in C_k} \boldsymbol{x}_i y_i\right)^T \left(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T - \sum_{i\in C_k} \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{x}_i y_i - \sum_{i\in C_k} \boldsymbol{x}_i y_i\right).\end{aligned} \tag{A.1}$$

The components of SSE_{C_k} in (A.1) for a given cluster C_k are $\sum_{i=1}^N y_i^2$, $\sum_{i=1}^N \boldsymbol{x}_i \boldsymbol{x}_i^T$, $\sum_{i=1}^N \boldsymbol{x}_i y_i$, $\sum_{i\in C_k} \boldsymbol{x}_i \boldsymbol{x}_i^T$, and $\sum_{i\in C_k} \boldsymbol{x}_i y_i$. Among these components, the first three need to be computed just once, but $\sum_{i\in C_k} \boldsymbol{x}_i \boldsymbol{x}_i^T$ and $\sum_{i\in C_k} \boldsymbol{x}_i y_i$ need to be calculated for every C_k. Thus, the last two components, $\sum_{i\in C_k} \boldsymbol{x}_i \boldsymbol{x}_i^T$ and $\sum_{i\in C_k} \boldsymbol{x}_i y_i$, are bottlenecks in the computation of the test statistic T.

The computational complexities for these component are $O(N)$, $O(Np^2)$, $O(Np)$, $O(|C_k|p^2)$, and $O(|C_k|p)$, respectively, where $|\cdot|$ denotes the cardinality of a set. Thus, the total computational complexity for all the clusters $C_k \in \mathcal{C} = \{C_1, C_2, \ldots, C_K\}$ is

$$O\left\{N(1+p^2+p)\right\} + O\left\{\sum_{k=1}^{K} |C_k|(p^2+p)\right\} = O\left\{\sum_{k=1}^{K} |C_k|(p^2+p)\right\} \tag{A.2}$$

because $\sum_{k=1}^K |C_k| \gg N$.

A.2 Computational Algorithm

As in Section 2.2, we can consider a total of m_i potential clusters centered at site i with radii $r_{i,1}, r_{i,2}, \ldots, r_{i,m_i}$. Let $C(i, r_{i,q})$ be the cluster centered at site i with the radius $r_{i,q}$ for $i = 1, \ldots, N$ and $q = 1, \ldots, m_i$. Then, $|C(i, r_{i,q})| = q$ and $\sum_{k=1}^K |C_k| = \sum_{i=1}^N \sum_{q=1}^{m_i} |C(i, r_{i,q})|$. Thus, (A.2) can be expressed as

$$O\left\{\sum_{i=1}^{N}\sum_{q=1}^{m_i} q(p^2+p)\right\} = O\left\{\sum_{i=1}^{N} m_i(m_i+1)(p^2+p)/2\right\}. \tag{A.3}$$

Based on the fact that $C(i, r_{i,1}) \subset C(i, r_{i,2}) \subset \ldots \subset C(i, r_{i,m_i})$ for clusters with the same centroid i, the cumulative sums for $\boldsymbol{x}_i \boldsymbol{x}_i^T$ and $\boldsymbol{x}_i y_i$ can be considered to ease the bottleneck. That is

$$\sum_{i'\in C(i,r_{i,q})} \boldsymbol{x}_{i'} \boldsymbol{x}_{i'}^T = \sum_{i'\in C(i,r_{i,q-1})} \boldsymbol{x}_{i'} \boldsymbol{x}_{i'}^T + \sum_{i'\in C(i,r_{i,q})\setminus C(i,r_{i,q-1})} \boldsymbol{x}_{i'} \boldsymbol{x}_{i'}^T.$$

Copyright © 2016 John Wiley & Sons, Ltd.

For example, suppose $C(1, r_{1,1}) = \{1\}$, $C(1, r_{1,2}) = \{1, 3\}$, $C(1, r_{1,3}) = \{1, 3, 7\}$, …, with the centroid $i = 1$. Then, $\sum_{i' \in C(1,r_{1,1})} \boldsymbol{x}_{i'} \boldsymbol{x}_{i'}^T = \boldsymbol{x}_1 \boldsymbol{x}_1^T$ has the computational complexity $O(p^2)$. For the next cluster at the centroid $i = 1$, $\sum_{i' \in C(1,r_{1,2})} \boldsymbol{x}_{i'} \boldsymbol{x}_{i'}^T = \boldsymbol{x}_1 \boldsymbol{x}_1^T + \boldsymbol{x}_3 \boldsymbol{x}_3^T$. However, because $\boldsymbol{x}_1 \boldsymbol{x}_1^T$ is already calculated in the previous cluster, only $\boldsymbol{x}_3 \boldsymbol{x}_3^T$ needs to be computed with the complexity $O(p^2)$. For the next cluster $C(1, r_{1,3}) = \{1, 3, 7\}$, we only need to calculate $\boldsymbol{x}_7 \boldsymbol{x}_7^T$ and its complexity is still $O(p^2)$. Thus, by considering these cumulative sums, the number of mathematical operations for the $C(i, r_{i,q})$'s with the same centroid i can be reduced from $\sum_{q=1}^{m_i} |C(i, r_{i,q})|(p^2 + p) = \sum_{q=1}^{m_i} q(p^2 + p)$ to $\sum_{q=1}^{m_i} (p^2 + p) = m_i(p^2 + p)$. Thus, the total computational complexity for all $C_k \in \mathcal{C} = \{C_1, C_2, \ldots, C_K\}$ becomes

$$O\left\{\sum_{i=1}^{N} \sum_{q=1}^{m_i} (p^2 + p)\right\} = O\left\{\sum_{i=1}^{N} m_i (p^2 + p)\right\} = O\left\{K(p^2 + p)\right\}, \tag{A.4}$$

where $K = \sum_{i=1}^{N} m_i$.

The ratio of (A.3) and (A.4) is

$$(1/2)\left(K^{-1} \sum_{i=1}^{N} m_i^2 + 1\right) \geqslant (1/2)\left(K^{-1} \sum_{i=1}^{N} (K/N)^2 + 1\right) = (1/2)\,(K/N + 1)\,. \tag{A.5}$$

The inequality in (A.5) suggests that we can ease the bottleneck computation by reducing computation by at least $(K/N + 1)/2$ times. For a 25×25 square grid in the unit square with a total of $N = 625$ cells, if we consider circular clusters with the maximum radius $R_{\max} = 1/5$ unit, there are a total of $K = 41493$ potential clusters. Thus, we could reduce the computation by about 30 times.

Appendix B: Source Code

The algorithm of our methodology is implemented in R. The source code and an illustrative example are available in the Supporting Information.

Acknowledgements

The authors thank the editor, an associate editor, and two referees for their insightful and constructive comments. We also thank Maria Kamenetsky for her assistance with the cancer mortality dataset. Funding has been provided by a USDA Cooperative State Research, Education and Extension Service (CSREES) McIntire-Stennis project and a pilot project from the Center for Demography of Health and Aging at the University of Wisconsin-Madison.

References

1. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. *Statistics in Medicine* 1995; **14**:799–810.
2. Kulldorff M. A spatial scan statistic. *Communications in Statistics, Part A* 1997; **26**:1481–1496.
3. Duczmal L, Assuncao R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis* 2004; **45**:269–284.
4. Gangnon RE, Clayton MK. Likelihood-based tests for localized detecting spatial clustering of disease. *Environmetrics* 2004; **15**:797–810.
5. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 2005; **4**:11.
6. Assuncao R, Costa M, Tavares A, Ferreira S. Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine* 2006; **25**:723–742.
7. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. *Statistics in Medicine* 2006; **25**:3929–3943.
8. Kulldorff M, Huang L, Konty K. A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics* 2009; **8**:58.
9. Gangnon RE. Local multiplicity adjustments for spatial cluster detection. *Environmental and Ecological Statistics* 2010; **17**(1):55–71.
10. Neill DB. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society, Series B* 2012; **74**(2): 337–360.
11. Shu L, Jiang W, Tsui KL. A standardized scan statistic for detecting spatial clusters with estimated parameters. *Naval Research Logistics* 2012; **59**:397–410.
12. Gangnon RE, Clayton MK. Bayesian detection and modeling of spatial disease clustering. *Biometrics* 2000; **56**:922–935.

Copyright © 2016 John Wiley & Sons, Ltd.

13. Gangnon RE, Clayton MK. A hierarchical model for spatially clustered disease rates. *Statistics in Medicine* 2003; **22**: 3213–3228.
14. Gangnon RE, Clayton MK. Cluster detection using bayes factors from overparameterized cluster models. *Environmental and Ecological Statistics* 2007; **14**:69–82.
15. Lawson AB. Cluster modelling of disease incidence via rjmcmc methods: a comparative evaluation. *Statistics in Medicine* 2000; **19**:2361–2375.
16. Clark AB, Lawson AB. Spatio-temporal cluster modelling of small area health data. In *Spatial Cluster Modelling*. Chapman and Hall/CRC: Boca Raton, FL, 2002; 235–258.
17. Yan P, Clayton MK. A cluster model for space-time disease counts. *Statistics in Medicine* 2006; **25**:867–881.
18. Wakefield J, Kim A. A bayesian model for cluster detection. *Biostatistics* 2013; **14**(4):752–765.
19. Brunsdon C, Fotheringham AS, Charlton ME. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis* 1996; **28**:281–298.
20. Fotheringham AS, Brunsdon C, Charlton M E. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley: New York, 2002.
21. Lawson AB, Choi J, Zhang J. Prior choice in discrete latent modeling of spatially referenced cancer survival. *Statistical Methods in Medical Research* 2014; **23**(2):183–200.
22. Zhang Z, Assuncao R, Kulldorff M. Spatial scan statistic adjusted for multiple clusters. *Journal of Probability and Statistics* 2010; Article ID 642379:11.
23. Waller LA, Hill EG, Rudd RA. The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. *Statistics in Medicine* 2006; **25**:853–865.
24. Gangnon RE. Local multiplicity adjustment for the spatial scan statistic using the Gumbel distribution. *Biometrics* 2012; **68**:174–182.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

Copyright © 2016 John Wiley & Sons, Ltd.