Spatial cluster detection of regression coefficients in a mixed-effects model

Spatial Cluster Detection of Regression Coefficients in a Mixed Effect Model

Junho Lee[a,*], Ying Sun[a], Howard H. Chang[b]

[a] *Statistics Program, CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia*

[b] *Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia 30322, United States*

Abstract

Identifying spatial clusters of different regression coefficients is a useful tool for discerning the distinctive relationship between a response and covariates in space. Most of the existing cluster detection methods aim to identify the spatial similarity in responses, and the standard cluster detection algorithm assumes independent spatial units. However, the response variables are spatially correlated in many environmental applications. We propose a mixed effect model for spatial cluster detection that takes spatial correlation into account. Compared to a fixed effect model, the introduced random effect explains extra variability among the spatial responses beyond the cluster effect, thus reducing the false positive rate. The developed method exploits a sequential searching scheme and is able to identify multiple potentially overlapping clusters. We use simulation studies to evaluate the performance of our proposed method in terms of the true and false positive rates of a known cluster, and the identification of multiple known clusters. We apply our proposed methodology to particulate matter ($PM_{2.5}$) concentration data from the Northeastern U.S. in order to study the weather effect on $PM_{2.5}$ and to investigate the association between the simulations from a numerical model and the satellite-derived aerosol optical depth (AOD) data. We find geographical hot spots which show distinct features comparing to the background.

Keywords: blockwise random effect; hot spot; mixed effect model; spatial cluster

[*] Corresponding author

Email addresses: junho.lee@kaust.edu.sa (Junho Lee), ying.sun@kaust.edu.sa (Ying Sun), howard.chang@emory.edu (Howard H. Chang)

detection; spatial scan statistic; varying coefficient regression.

1. Introduction

Fine particulate matter less than 2.5 μm in aerodynamic diameter ($PM_{2.5}$) is a well-known harmful air pollutant. For example, many recent studies have shown that high concentrations of $PM_{2.5}$ are a risk factor for mortality (Samoli et al., 2008), various zcardiopulmonary diseases (Dominici et al., 2006; Pope and Dockery, 2006), and preterm birth (Chang et al., 2012). Air quality, such as the $PM_{2.5}$ concentration, is typically associated with meteorological conditions since the complex chemistry and physics in the atmosphere can influence the fate and transport of pollutants. Regression models are often used to identify the association between air pollutant concentrations and weather predictors, such as temperature, relative humidity, and wind speed (Jacob and Winner, 2009; Porter et al., 2015; Russell et al., 2017).

When analyzing spatially varying pollutants and predictors, the regression model may not have exactly the same coefficients across the entire region of study due to differences in emission sources, air pollution composition, and missing confounders. A subdomain that shows distinct patterns in the regression coefficients relative to the rest of the region is called a spatial cluster or a hot spot, and the rest of the region is referred to as the background. Kulldorff and Nagarwalla (1995) and Kulldorff (1997) proposed the spatial scan statistic for identifying spatial clusters that have distinctive risks in the Poisson process or the Bernoulli process. Later, many variants of the scan statistic (Duczmal and Assunção, 2004; Gangnon and Clayton, 2004; Tango and Takahashi, 2005; Assunção et al., 2006; Kulldorff et al., 2006, 2009; Jung, 2009; Gangnon, 2010b; Neill, 2012; Shu et al., 2012; Xu and Gangnon, 2016; Lin et al., 2016) have been proposed. Especially, Kulldorff et al. (2009) proposed a spatial scan statistic for the Gaussian data and, further, Jung (2009) proposed a covariate-adjusted spatial scan statistic based on generalized linear models for identifying spatial clusters in the intercepts only, while assuming the slopes associated with the covariates are identical for the cluster and the background. Most of these spatial scan statistics are implemented in the SaTScan™ software (http://www.satscan.org), and covariate-adjusted spatial scan statistics are available with the DClusterm package (Gómez-Rubio et al., 2018) which is implemented in R (R Core Team, 2017). Alternatively, Bayesian models were introduced for the detection of spatial clusters in disease mapping (Gangnon and Clayton, 2000, 2003; Knorr-Held and Raßer, 2004; Gangnon and Clayton, 2007; Lawson, 2000; Clark and Lawson, 2002; Yan and Clayton, 2006; Wakefield and Kim, 2013). However, the aforementioned spatial scan statistics and Bayesian cluster detection approaches are focusing on the cluster in the response or the intercepts only, but not the association between the response and covariates.

Example approaches for clustering regression coefficients, fused lasso methods (Tibshirani et al., 2005; Friedman et al., 2007; Yang et al., 2012; Tang and Song, 2016; Wang et al., 2016) have been proposed, as well as the other penalized regression models (Shen and Huang, 2010; Ke et al., 2015; Shin et al., 2016; Tutz and Oelker, 2017). Alternatively, Berger and Tutz (2018) proposed a tree-structured clustering approach to split regression coefficients into several groups. However, all of the aforementioned methods are not readily applicable to spatial data to achieve geographic clusters. Further, these clustering approaches mainly focus on grouping, but do not aim to identify spatial hot spots.

In the spatial regression analysis, the geographically weighted regression (GWR) (Brunsdon et al., 1996; Fotheringham et al., 2002) is one of the popular approaches to a spatial varying coefficient model. GWR provides regression coefficient estimates which are locally weighted and vary across space. Alternatively, in a Bayesian framework, Lawson et al. (2014) proposed an approach to a grouped spatial varying coefficient regression when the total number of groups is known *a priori*. However, neither method is directly applicable to detection of hot spot.

Recently, Lee et al. (2017a) proposed detecting an unknown number of spatial clusters in the spatial regression coefficients, which can be useful for informing subsequent spatially varying coefficient regression based on the detected spatial hot spots. However, as other spatial scan statistic approaches, Lee et al. (2017a) assume independent observations, which are not necessarily applicable to spatial data. Although some approaches to spatial cluster detection addressed the issue of spatial correlation by considering a spatial random effect (Kleinman et al., 2005; Loh and Zhu, 2007; Zhang and Lin, 2009; Lin et al., 2016), they define the spatial clusters for either the response or the intercepts only, not for the regression slopes. Alternatively, Kim et al. (2005) adopted k-means (MacQueen, 1967) and Reich and Bondell (2010) considered Dirichlet process mixture for clustering with air mass data and population genetic data, respectively. But, they use spatial information to classify individual observation or air mass back-trajectories, not to detect spatial subdomains or hot spots.

In this paper, we propose a method of detecting spatial clusters of regression coefficients using a mixed effect model. It can be viewed as an extension of the method proposed by Lee et al. (2017a), who used a fixed effect model and assumed the spatial observations to be independent. Here, we address the issue of spatial correlation when identifying spatial clusters in regression slopes. This is crucial because when the spatial correlation is ignored, any significant differences observed in the slope may simply be due to an inflated similarity among the covariate effects caused by the residual spatial correlation, rather than a true spatial cluster. Therefore, to take the spatial dependence into account, we introduce a spatial blockwise random effect, which leads to the mixed effect model. Our proposed approach enables not only the detection of spatial hot spots, which show distinct features in the coefficients comparing to the background, but also addressing the potential spatial dependency.

The remainder of the paper is organized as follows. In Section 2, we describe the motivating data which includes $PM_{2.5}$ concentration estimates, weather drivers, and satellite-derived aerosol optical depth (AOD) from the Northeastern U.S. In Section 3, we define our mixed effect model with a spatial blockwise random effect and develop corresponding hypothesis tests for the spatial cluster effects. For multiple clusters, we develop a sequential detection scheme. In Section 4, we perform simulations to evaluate our proposed method in terms of false positive or power, and identification of true clusters by comparing its results to those of the fixed effect model approach (Lee et al., 2017a). We provide a detailed analysis of the mixed effect model's application to the real data in Section 5. Section 6 contains a discussion and final conclusions.

2. Data description

Our study domain, which is defined by the National Climatic Data Center (Karl and Koss, 1984), covers the Northeastern U.S. (Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and District of Columbia).

The Environmental Protection Agency (EPA) provides the $PM_{2.5}$ values, which are generated via the Community Multiscale Air Quality Modeling System (CMAQ); both the raw and fused CMAQ data are available at https://www.epa.gov/cmaq. The fused CMAQ $PM_{2.5}$ is bias-corrected by combining the raw CMAQ data with the monitoring data (Berrocal et al., 2010a,b, 2011).

For the weather variables, we obtain the meteorological drivers of $PM_{2.5}$ from the North American Regional Reanalysis (NARR, https://www.esrl.noaa.gov/psd). We select three covariates based on information from previous similar researches (Jacob and Winner, 2009; Porter et al., 2015; Russell et al., 2017), as shown in Table 1.

[Table 1 about here.]

Another proxy measurement of particle air pollution data is the AOD obtained from satellites. Previous studies on predicting $PM_{2.5}$ concentration from AOD were reviewed by Chu et al. (2016). And, those regression based models showed that $PM_{2.5}$ concentrations have positive relationships with AOD (Liu et al., 2005; Paciorek et al., 2008; Kloog et al., 2011; Lee et al., 2011; Kloog et al., 2012; Chang et al., 2013; Ma et al., 2016; Yu et al., 2017; Grantham et al., 2018) because AOD measures light extinction due to particles (e.g., dust, smoke, pollution) in the atmospheric column. We also obtain satellite-measured AOD data for the Northeastern U.S. from the Moderate Resolution Imaging Spectroradiometer (MODIS) in order to investigate its relationship to the $PM_{2.5}$ concentrations in the fused CMAQ data via our mixed effect model with a spatial blockwise random effect.

While the CMAQ is available at the 12×12 km^2 grid, the NARR data is available on a 32×32 km^2 grid. Thus, a 32×32 km^2 grid is the common resolution available when we consider a regression modelling with the data from these different sources (CMAQ and NARR). We organize the monthly mean data for June, July, and August 2012, on a 32×32 km^2 grid by averaging each variable (raw CMAQ PM$_{2.5}$, fused CMAQ PM$_{2.5}$, TMP, RH, WS, and AOD) up to match the NARR grid cell.

3. Spatial cluster detection

We let $\mathcal{D}$ denote a spatial domain of interest in $\mathbb{R}^2$ and be partitioned into g subregions, $D_1, \ldots, D_g$. We let n_i denote the number of cells that partition subregion D_i with geographical centroids $\boldsymbol{s}_{ij} = (s_{1ij}, s_{2ij})^T$ for $i = 1, \ldots, g$, $j = 1, \ldots, n_i$. For cell (i, j), $y(\boldsymbol{s}_{ij})$, the response variable at $\boldsymbol{s}_{ij}$, is denoted by y_{ij} for simplicity. Then, we model the response variable with the mean response and the random error as $y_{ij} = \mu_{ij} + \varepsilon_{ij}$ for $i = 1, \ldots, g$, $j = 1, \ldots, n_i$. Further, we decompose the random error ε_{ij} as $\varepsilon_{ij} = b_i + e_{ij}$, where b_i's and e_{ij}'s are *iid* $\mathcal{N}(0, \sigma_b^2)$ and *iid* $\mathcal{N}(0, \sigma_e^2)$ with variance components $\sigma_b^2 > 0$ and $\sigma_e^2 > 0$, respectively, and b_i's and e_{ij}'s are independent. As we can see, this is a mixed effect model with a spatial blockwise random effect b_i that is associated with the subregion D_i and a purely random error e_{ij}. Thus, the correlation between two observations is

$$corr(y_{i_1 j_1}, y_{i_2 j_2}) = \begin{cases} \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}, & i_1 = i_2,\ j_1 \neq j_2 \quad \text{(within subregion)}, \\ 0, & i_1 \neq i_2 \quad\quad\quad\quad \text{(across subregions)}. \end{cases}$$

The spatial correlations could be defined further across subregions similarly to Bai et al. (2014); however, we do not pursue this direction in this paper. Our proposed model simply introduces the spatial dependence through a blockwise random effect b_i. With N observations and p covariates, the mixed effect model requires the computational complexity of $\mathcal{O}(Npm)$, where m is the rank of the covariance matrix for the random effect (Darnell et al., 2017; Tan et al., 2018). Thus, our mixed effect model has the computational complexity $\mathcal{O}(Npg)$ whereas a Gaussian process regression typically needs the higher computational cost as $\mathcal{O}(N^3)$. Further, in our model, subregions are given, and there is no longer any spatially correlated structure across b_i's so that we can easily fit our model like a linear mixed model to non-spatial data. We use these subregions to approximate the spatial dependence; this will be examined further in terms of a continuous spatial process in Section 4.

We define the subregions D_i by grouping the coordinates $\boldsymbol{s}_{ij}$. For example, in the simulation studies performed on a square grid in Section 4, we divide the study region into smaller squares and define these as the subregions. In practice, we may define the subregions by clustering x-coordinates and y-coordinates (or longitudes and latitudes). For example, k-means clustering with xy-coordinates could be an option to define the subregions as in Section 5.

Further, we do not assume that the covariates and random effects are spatially correlated. Therefore, we do not consider the potential issue of spatial confounding, where multicollinearity between the spatially random effects and covariates changes the regression coefficients estimates in a mixed effect model (Paciorek, 2010; Hodges and Reich, 2010; Hughes and Haran, 2013). In practice, however, if the spatial confounding is suspected, then the restricted spatial regression model can be applied (Hodges and Reich, 2010).

Similar to Lee et al. (2017a), for a spatial cluster C, we define a circular window with a center $\boldsymbol{c}$ and radius r such that

$$C = \big\{(i,j) \mid d(\boldsymbol{s}_{ij}, \boldsymbol{c}) \leqslant r\big\}, \tag{1}$$

where $d(\cdot,\cdot)$ is the distance between two locations. That is, those cells which geographical centroids $\boldsymbol{s}_{ij}$ are within r distance of $\boldsymbol{c}$ form the circular cluster C. Although numerous studies have reviewed and compared the performance of various window shapes (Huang et al., 2008; Goujon-Bellec et al., 2011; Grubesic et al., 2014), we consider the circular window because it can be simply defined even in the irregular grid data (e.g., county level data). However, in practice, it can be substituted by other shapes, such as ellipses and squares (Tango and Takahashi, 2005; Assunção et al., 2006; Kulldorff et al., 2006; Murray et al., 2014; Yin and Mu, 2018).

We let $\mathcal{C} = \{C_1, C_2, \ldots\}$ denote the set of all the candidate clusters of the form (1). Then, with a spatial cluster $C \in \mathcal{C}$, the mean response μ_{ij} follows a varying coefficient model:

$$\mu_{ij} = \boldsymbol{x}_{ij}^T\big(\boldsymbol{\beta} + \boldsymbol{\theta} \cdot \mathcal{I}\{(i,j) \in C\}\big), \tag{2}$$

where $\boldsymbol{x}_{ij}$ denotes the corresponding covariates, $\boldsymbol{\beta}$ is the regression coefficient vector for the background (i.e., the non-cluster), $\boldsymbol{\theta}$ is the cluster effect associated with C, and $\mathcal{I}(\cdot)$ is the indicator function.

Furthermore, within a subregion D_i, for $i = 1, \ldots, g$, we can represent the response variable in a vector form as $\boldsymbol{y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^T$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^T \sim \mathcal{N}_{n_i}(\mathbf{0}_{n_i}, \boldsymbol{\Lambda}_i)$ is multivariate normally distributed. The $n_i \times n_i$ within-subregion covariance matrix $\boldsymbol{\Lambda}_i$ is represented by $\boldsymbol{\Lambda}_i = Var(\boldsymbol{\varepsilon}_i) = Var(\boldsymbol{b}_i) + Var(\boldsymbol{e}_i) = \sigma^2 \cdot \boldsymbol{\Sigma}_i$, where $\sigma^2 = \sigma_b^2 + \sigma_e^2$, $\boldsymbol{\Sigma}_i = (1-\rho)\mathbf{I}_{n_i} + \rho \mathbf{J}_{n_i}$ is the $n_i \times n_i$ within-subregion correlation matrix, $\rho = \sigma_b^2 / \sigma^2$, $\mathbf{I}_{n_i}$ is an $n_i \times n_i$ identity matrix, and $\mathbf{J}_{n_i}$ is an $n_i \times n_i$ matrix of ones.

By combining all the subregions D_i and their corresponding cells, our model is $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_g^T)^T$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_g^T)^T$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^T, \ldots, \boldsymbol{\varepsilon}_g^T)^T \sim \mathcal{N}_N(\mathbf{0}, \boldsymbol{\Lambda})$, and $N = \sum_{i=1}^{g} n_i$. The covariance matrix $\boldsymbol{\Lambda}$ is an $N \times N$ block-diagonal matrix of the form

$$\boldsymbol{\Lambda} = diag\{\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_g\} = \sigma^2 \cdot diag\{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_g\} = \sigma^2 \cdot \boldsymbol{\Sigma}, \tag{3}$$

where $\boldsymbol{\Sigma}$ is the block-diagonal correlation matrix. The off-block-diagonal entries in both $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}$ are all zero because we assume independent subregions.

Since we include both spatially clustered intercept and a spatial blockwise random effect in the model, there will be an identifiability issue between these two when the cluster C equals to one of subregions. However, this identifiability problem can be solved by redefining the set of the candidate clusters $\mathcal{C}^*$ by avoiding those subregions, if any, from $\mathcal{C}$ as $\mathcal{C}^* = \mathcal{C} \setminus \{D_1, \ldots, D_g\}$.

3.1. Single cluster detection

We begin with a cluster-specific local hypothesis test for the kth candidate cluster $C_k \in \mathcal{C}$, $k = 1, 2, \ldots$:

$$H_{0_k} : \boldsymbol{\theta}_k = \mathbf{0} \quad \text{versus} \quad H_{A_k} : \boldsymbol{\theta}_k \neq \mathbf{0},$$

$$\lambda(C_k) = \frac{\mathcal{L}(\hat{\boldsymbol{\zeta}}_{A_k})}{\mathcal{L}(\hat{\boldsymbol{\zeta}}_{0_k})},$$

where $\boldsymbol{\theta}_k$ is the cluster effect associated with C_k, $\mathcal{L}(\boldsymbol{\zeta})$ is the likelihood evaluated at $\boldsymbol{\zeta} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, \sigma^2, \rho)^T$, and $\lambda(C_k)$ is the likelihood ratio test (LRT) statistic. In the implementation, we used the lmer function in the lme4 package (Bates et al., 2015) for R to achieve $\hat{\boldsymbol{\zeta}}_{0_k}$ and $\hat{\boldsymbol{\zeta}}_{A_k}$ which are the estimates of $\boldsymbol{\zeta}$ under H_{0_k} and H_{A_k}, respectively.

Now, we consider a global hypothesis test to locate an unknown generic cluster C among all the candidate clusters in $\mathcal{C}$:

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \;\; \textit{for all } C \in \mathcal{C} \qquad \text{versus} \qquad H_A : \boldsymbol{\theta} \neq \mathbf{0} \;\; \textit{for some } C \in \mathcal{C}, \tag{4}$$

$$\nu = \max_{C \in \mathcal{C}} \{\lambda(C)\}, \qquad \hat{C} = \arg\max_{C \in \mathcal{C}} \{\lambda(C)\}, \tag{5}$$

where ν is the test statistic for (4) defined as the largest value of the LRT statistics for all the cluster-specific local hypothesis tests, and $\hat{C}$ is the cluster estimate corresponding to the test statistic ν. Since the null distribution of ν in (5) does not exit in a closed form, we adopt a Monte Carlo method to compute a p-value (Lee et al., 2017a,b). First, we estimate $\boldsymbol{\zeta} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, \sigma^2, \rho)^T$ under H_0 in (4) to be $\hat{\boldsymbol{\zeta}}_0$. Second, we generate S Monte Carlo samples with $\hat{\boldsymbol{\zeta}}_0$ under H_0. Third, we compute the test statistic ν in (5) for each of the Monte Carlo samples, which are denoted by $\nu_1, \ldots, \nu_S$ in descending order so that ν_1 has rank 1. Finally, the observed test statistics are denoted by ν_{obs}, and R is the rank of ν_{obs}. We define the p-value as $R/(S + 1)$.

3.2. Multiple cluster detection

There may be more than one cluster in the area of interest. To find those unknown numbers of spatial clusters, we adopt the sequential detection approach (Zhang et al., 2010; Lee et al., 2017a), detailed below:

(i) Predefine $\mathcal{C}$ with N cells on the spatial lattice, and the minimum and maximum radii, $r_{\min}$ and $r_{\max}$, respectively.

(ii) Obtain the cluster $\hat{C} = \arg\max_{C \in \mathcal{C}} \{\lambda(C)\}$, its p-value, and the residuals $\hat{e}_{ij} = y_{ij} - \hat{\mu}_{ij} - \hat{b}_i$, where $\hat{\mu}_{ij} = \boldsymbol{x}_{ij}^T(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\theta}} \cdot \mathcal{I}\{(i, j) \in \hat{C}\})$.

(iii) Update the responses as the residuals, i.e., replace y_{ij} with $\hat{e}_{ij}$.

(iv) Repeat steps (ii)–(iii) until the p-value $> \alpha$. That is, stop only if the p-value in step (ii) is greater than the significance level α.

Here, we allow multiple overlapping clusters. However, this approach also can handle the detection of multiple non-overlapping clusters by updating $\mathcal{C}$ with $\mathcal{C} \setminus \{C \mid C \cap \hat{C} \neq \emptyset\}$ in step (iii), where $\hat{C}$ is the cluster estimate from step (ii) and $\emptyset$ is the empty set.

3.3. Two–stage multiple cluster detection

The approach proposed in Sections 3.1–3.2 uses hypothesis testing on the cluster effects to determine the spatial cluster estimates in both the intercepts and the slopes. Therefore, when the slope is of primary interest, it is unclear whether the cluster effects stem from the intercepts or the slopes. To solve this problem, we adopt the two–stage approach of Lee et al. (2017a), which enables us to detect multiple spatial clusters in the slopes and the intercepts separately.

Again, we begin with two cluster-specific local hypothesis tests on the slopes and the intercepts for the kth candidate cluster $C_k \in \mathcal{C}$, $k = 1, 2, \ldots$:

$$H_{0_k}^{\text{slp}} : \boldsymbol{\theta}_{k,-0} = \mathbf{0} \qquad \text{versus} \qquad H_{A_k}^{\text{slp}} : \boldsymbol{\theta}_{k,-0} \neq \mathbf{0}, \tag{6}$$

$$H_{0_k}^{\text{int}} : \boldsymbol{\theta}_{k,-0} = \mathbf{0},\ \theta_{k,0} = 0 \qquad \text{versus} \qquad H_{A_k}^{\text{int}} : \boldsymbol{\theta}_{k,-0} = \mathbf{0},\ \theta_{k,0} \neq 0, \tag{7}$$

where $\boldsymbol{\theta}_k = (\theta_{k,0}, \theta_{k,1}, \ldots, \theta_{k,(p-1)})^T \in \mathbb{R}^p$ is the cluster effect associated with C_k, and $\boldsymbol{\theta}_{k,-0} = (\theta_{k,1}, \ldots, \theta_{k,(p-1)})^T \in \mathbb{R}^{p-1}$. The LRT statistics for (6) and (7) are $\lambda^{\text{slp}}(C_k) = \mathcal{L}(\hat{\boldsymbol{\zeta}}_{A_k}^{\text{slp}})/\mathcal{L}(\hat{\boldsymbol{\zeta}}_{0_k}^{\text{slp}})$ and $\lambda^{\text{int}}(C_k) = \mathcal{L}(\hat{\boldsymbol{\zeta}}_{A_k}^{\text{int}})/\mathcal{L}(\hat{\boldsymbol{\zeta}}_{0_k}^{\text{int}})$, respectively. The estimates under $H_{0_k}^{\text{slp}}$, $H_{A_k}^{\text{slp}}$, $H_{0_k}^{\text{int}}$, and $H_{A_k}^{\text{int}}$ are $\hat{\boldsymbol{\zeta}}_{0_k}^{\text{slp}}$, $\hat{\boldsymbol{\zeta}}_{A_k}^{\text{slp}}$, $\hat{\boldsymbol{\zeta}}_{0_k}^{\text{int}}$, and $\hat{\boldsymbol{\zeta}}_{A_k}^{\text{int}}$, respectively.

Then, in the global hypothesis test, we define the test statistic and the cluster estimate for an unknown generic cluster $C \in \mathcal{C}$ in the slopes as

$$\begin{aligned} & H_0^{\text{slp}} : \boldsymbol{\theta}_{-0} = \mathbf{0} \ \ \textit{for all } C \in \mathcal{C} \\ \text{versus} \quad & H_A^{\text{slp}} : \boldsymbol{\theta}_{-0} \neq \mathbf{0} \ \ \textit{for some } C \in \mathcal{C}, \\ & \nu^{\text{slp}} = \max_{C \in \mathcal{C}} \{\lambda^{\text{slp}}(C)\}, \qquad \hat{C} = \arg\max_{C \in \mathcal{C}} \{\lambda^{\text{slp}}(C)\}. \end{aligned} \tag{8}$$

Similarly, we define the global hypothesis testing for an unknown generic cluster $C \in \mathcal{C}$ in the intercepts as:

$$\begin{aligned} & H_0^{\text{int}} : \boldsymbol{\theta}_{-0} = \mathbf{0},\ \theta_0 = 0 \ \textit{ for all } C \in \mathcal{C} \\ \text{versus} \quad & H_A^{\text{int}} : \boldsymbol{\theta}_{-0} = \mathbf{0},\ \theta_0 \neq 0 \ \textit{ for some } C \in \mathcal{C}, \\ & \nu^{\text{int}} = \max_{C \in \mathcal{C}} \{\lambda^{\text{int}}(C)\}, \qquad\qquad \hat{C} = \arg\max_{C \in \mathcal{C}} \{\lambda^{\text{int}}(C)\}, \end{aligned} \tag{9}$$

Similar to Section 3.1, the p-values of ν^{slp} in (8) and of ν^{int} in (9) can be computed via the Monte Carlo method.

For multiple potential clusters, we adopt the sequential detection scheme again. This time, however, the first stage is reserved for the slopes, and the second for the intercepts. Further, we adjust the p-value with the Bonferroni correction since this detection method consists of dual tests for the first and second stages (Dunn, 1961). That is, we

(i) Predefine $\mathcal{C}$ with N cells on the spatial lattice, and the minimum and maximum radii, $r_{\min}$ and $r_{\max}$, respectively.

— *First stage* —

(ii) Obtain the cluster $\hat{C} = \arg\max_{C \in \mathcal{C}} \{\lambda^{\text{slp}}(C)\}$, its p-value, and the residuals.

(iii) Update the responses as the residuals.

(iv) Repeat steps (ii)–(iii) until the p-value $> \alpha/2$. That is, stop and go to step (v) only if the p-value in step (ii) is greater than $\alpha/2$.

— *Second stage* —

(v) Obtain the cluster $\hat{C} = \arg\max_{C \in \mathcal{C}} \{\lambda^{\text{int}}(C)\}$, its p-value, and the residuals.

(vi) Update the responses as the residuals.

(vii) Repeat steps (v)–(vi) until the p-value $> \alpha/2$. That is, stop only if the p-value in step (v) is greater than the significance level $\alpha/2$.

4. Simulation studies

In this section, we conduct two simulation studies to evaluate our proposed method. The first one is for the false positive and power computation, and the second study is to evaluate how the proposed approach identifies the true clusters well enough. In each simulation study, we compare our mixed effect model approach to the fixed effect model approach (Lee et al., 2017a). Both the mixed effect and fixed effect approaches are implemented in R, and the lme4 package is used to fit the mixed effect model.

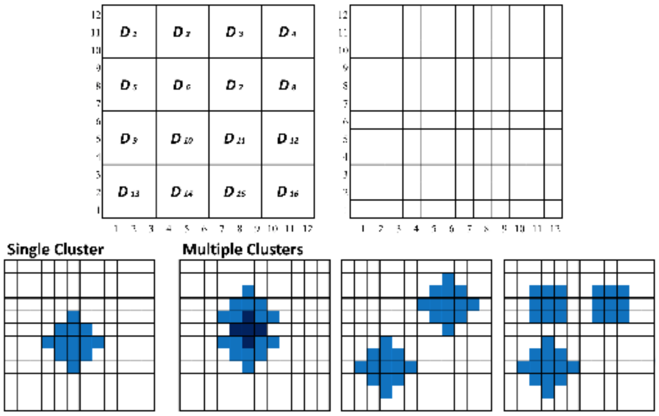


Figure 1: Top row: Sixteen subregions in a 12×12 square grid; each subregion is defined as a 3×3 square grid. Bottom row: A single cluster setting for the false positive and power computation, and three multiple clusters settings for the true clusters identification.

The setting for our simulation study on a 12×12 square grid is illustrated in Figure 1. For the blockwise random effect, we consider a total of sixteen subregions defined as 3×3 square grids. That is, $n_i = 9$ for $i = 1, \ldots, g$ where $g = 16$, and $N = \sum_{i=1}^{g} n_i = 144$. Further, we consider a single cluster setting for the false positive and power computation and three settings for the multiple cluster identification. We generate two covariates, x_{ij1} and x_{ij2}, for cell (i, j) from the standard normal distribution $\mathcal{N}(0, 1)$. The regression coefficients for the background $\boldsymbol{\beta}$ and the variance component for the pure random error e_{ij} are set to $\boldsymbol{\beta} = (0, 0, 0)^T$ and $\sigma_e^2 = 1$, respectively.

We also consider model misspecification. Instead of simulating the observations from the mixed effect model, we simulate $N = 144$ observations from a Gaussian process (GP):

$$\boldsymbol{y}_{\mathcal{GP}} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_{\mathcal{GP}}, \quad \boldsymbol{\varepsilon}_{\mathcal{GP}} \sim \ \mathcal{GP}\big(\mathbf{0}, K(\cdot)\big), \quad K(d;\phi) = \sigma^2_{\mathcal{GP}} \cdot e^{-\frac{d^2}{\phi^2}},$$

where $\sigma^2_{\mathcal{GP}} \cdot e^{-d^2/\phi^2}$ is a Gaussian covariance function, and d is the distance between two observations in units of grid spacing. We set $\sigma^2_{\mathcal{GP}}$ equal to 1 and vary the range parameter ϕ to induce different strengths of spatial correlation. Thus, we show that our mixed effect approach provides satisfactory results even when the data are simulated from a GP. We also evaluate the impact of the block size for the subregions.

4.1. False positive and power evaluation

We conduct a simulation study using the single cluster setting shown in Figure 1 for the power evaluation. The cluster effect $\boldsymbol{\theta}$ is set to $\boldsymbol{\theta} = (\theta, \theta, \theta)^T$, where $\theta \in \{0.0,\ 0.1,\ 0.2,\ \ldots,\ 2.5\}$, and the variance component for the spatial blockwise random effect b_i is set to $\sigma^2_b \in \{0.1, 0.5, 0.7, 1.0, 2.0\}$, i.e., σ^2_b/σ^2_e is 0.1, 0.5, 0.7, 1.0, or 2.0. For the GP error $\boldsymbol{\varepsilon}_{\mathcal{GP}}$, ϕ is set to 1, 2, or 3.

The power is defined as the proportion of the simulations in which the global null hypothesis (4) is rejected at the significance level $\alpha = 0.05$ (Gangnon and Clayton, 2004; Waller et al., 2006; Gangnon, 2010a, 2012; Lee et al., 2017a). We simulate 1,000 datasets from each setting.

[Table 2 about here.]

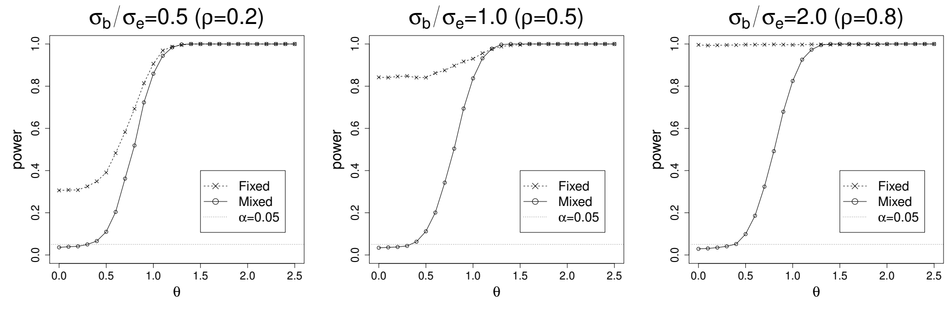


Figure 2: Power curve: $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$, $\sigma^2 = \sigma^2_b + \sigma^2_e$ and $\rho = \sigma^2_b/\sigma^2$.

Table 2 shows the empirical false positive, which is the power computation when the cluster effect does not exist, $\boldsymbol{\theta} = (0, 0, 0)^T$. Our mixed effect approach provides false positive rates close to the nominal $\alpha = 0.05$. However, the false positive in the fixed effect model becomes inflated as σ_b^2, the variation of the random effect, increases. The empirical powers are illustrated in Figure 2. The mixed effect model provides an S-shaped power curve, whereas the power curve of the fixed effect approach is inflated everywhere. Since the power is defined with such an event where the global null hypothesis, $H_0 : \boldsymbol{\theta} = \mathbf{0}$ *for all* $C \in \mathcal{C}$, is rejected, a higher power does not mean that the cluster estimate is closer to the true cluster. Identification of the true clusters will be discussed in Section 4.2.

[Table 3 about here.]

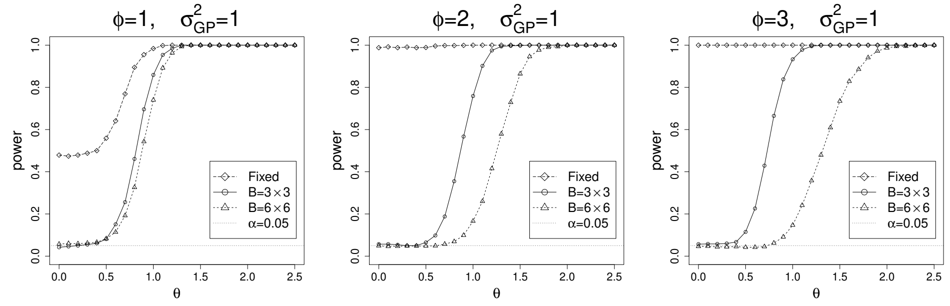


Figure 3: Power curve: $\boldsymbol{y}_{\mathcal{GP}} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_{\mathcal{GP}}$, where $\boldsymbol{\varepsilon}_{\mathcal{GP}} \sim \mathcal{GP}\big(\mathbf{0}, K(\cdot)\big)$ and $K(d;\phi) = \sigma^2_{\mathcal{GP}} \cdot e^{-\frac{d^2}{\phi^2}}$.

Table 3 and Figure 3 show the results from the simulation performed with GP errors. When we apply our mixed effect model approach to the GP data, we consider two different block sizes for the subregions: sixteen 3×3 square grids ($B = 3 \times 3$), which is the same as in Figure 1, and four 6×6 square grids ($B = 6 \times 6$). Similar to the results from the simulation with the blockwise random effect, the fixed effect model shows a high false positive rate (shown in Table 3) and inflated power (shown in Figure 3), which are caused by the spatial autocorrelation increasing as ϕ gets larger. With respect to the false positive, we also consider another Gaussian process with an exponential covariance function, $K(d; \phi) = \sigma^2_{\mathcal{GP}} \cdot e^{-d/\phi}$, where d is the distance between two observations in units of grid spacing and ϕ is set to be 0.7, 1.0 or 1.4 (shown in Table 3). With 1,000 simulation, the standard errors of the estimated false positive rates are approximately $\sqrt{(0.05)(0.95)} \approx 0.0069$. Thus, all of the estimated false positives in the mixed effect model are within, or lie slightly more than, one standard error away from the nominal $\alpha = 0.05$. Further, they are all within two standard errors away from $\alpha = 0.05$. That is, the block size for the random effect does not affect the false positive. However, it is apparent that the block size affects the power for detecting a cluster. The smaller block size ($B = 3 \times 3$) provides a higher power in the mixed effect model (Figure 3). But, we are very careful to choose the block size as small as possible since the smaller block size with the stronger spatial correlation may result in the inflated power. Indeed, in the smaller block size ($B = 3 \times 3$), the data with $\phi = 3$ provides the higher power for all the cluster effect size $\theta \in \{0.0,\ 0.1,\ 0.2,\ \ldots,\ 2.5\}$ than the data with $\phi = 1$.

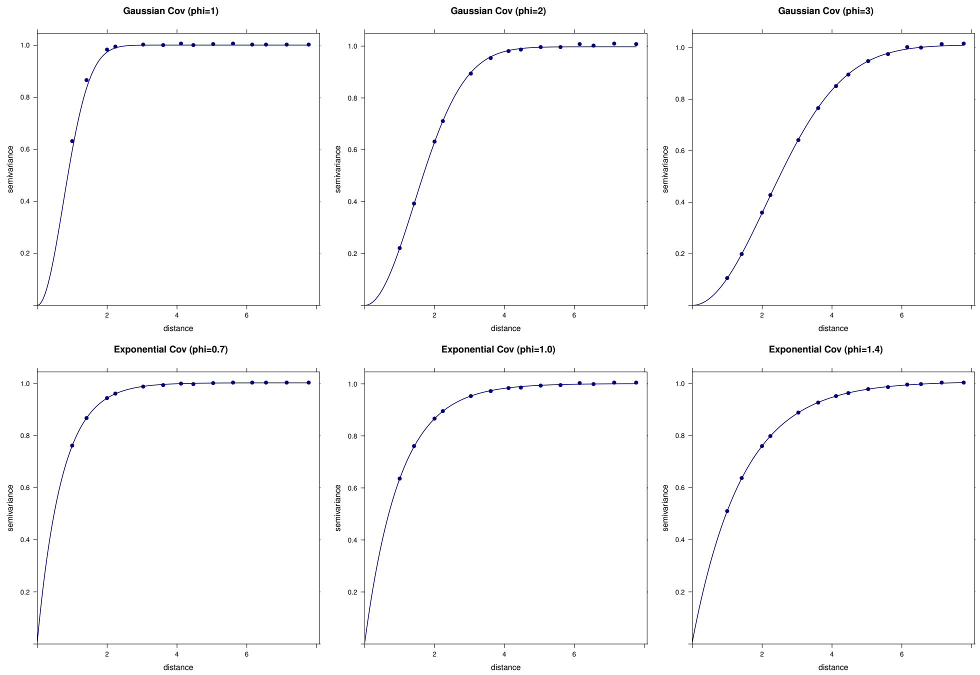


Figure 4: Variogram: $\boldsymbol{y}_{\mathcal{GP}} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_{\mathcal{GP}}$, where $\boldsymbol{\varepsilon}_{\mathcal{GP}} \sim \mathcal{GP}\big(\mathbf{0}, K(\cdot)\big)$, $K(d;\phi) = \sigma^2_{\mathcal{GP}} \cdot e^{-d^2/\phi^2}$ for the Gaussian covariance function (top row) and $K(d;\phi) = \sigma^2_{\mathcal{GP}} \cdot e^{-d/\phi}$ for the exponenetial covariance function (bottom row).

Hence, in practice, the strength of spatial correlation would be taken into account when choosing an appropriate number of subregions, or the block size for the random effect. However, we do not know about the true spatial dependent structure in the real data. Thus, here, we provide a rule of thumb to choose the block size based on the spatial dependence range. Based on the exploratory analysis, we suggest to choose the block size $B \approx \ell \times \ell$ for some $\ell \in (\frac{\mathcal{R}}{2}, \mathcal{R})$, where $\mathcal{R}$ the spatial dependence range. In this simulation study, variogram suggests $\mathcal{R} = 3, 5$ and 7 for the Gaussian covariance with $\phi = 1, 2$ and 3, and for the exponential covariance $\phi = 0.7, 1.0$ and 1.4, respectively (shown in Figure 4). We leave the theoretical details about the block size for future research.

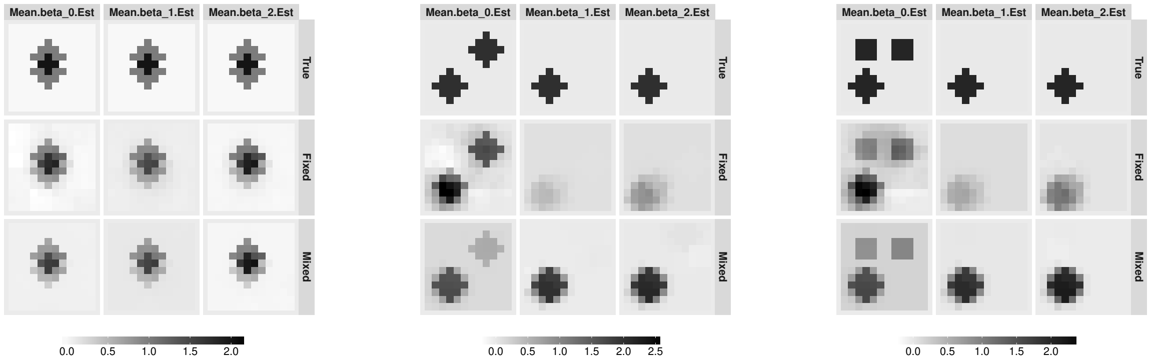
4.2. True cluster identification

To evaluate how well the proposed method identifies the true spatial clusters, we conduct a simulation study with the multiple clusters (settings shown in Figure 1). For the spatial blockwise random effect model, σ_b^2 is set to 1 (i.e., $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_e^2) = 0.5$). For the GP error model with a Gaussian covariance function, ϕ is also set to 1. We generate 100 datasets for each combination of the three multiple cluster settings and two random effect settings (i.e., blockwise random effect and GP error).

For each simulated dataset, we detect multiple clusters at the significance level $\alpha = 0.05$. During the detection, which is based on our mixed effect model and the sequential process presented in Sections 3.2–3.3, we consider a grid of sixteen 3×3 squares as displayed in Figure 1. For each simulated dataset, we estimate the regression coefficients with detected spatial clusters. Then we map the mean coefficient estimates for each multiple cluster setting as illustrated in Figure 5. The maps of the mean squared error (MSE) corresponding to each subfigure in Figure 5 are provided in Figure 6.

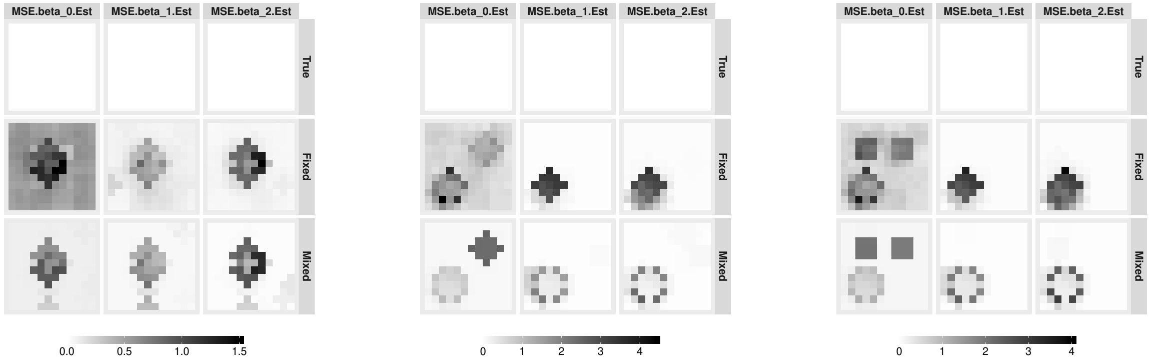
Each subfigure in Figures 5–6 contains nine maps. In each subfigure, row 1 is the map of the true coefficients, and rows 2 and 3 are from the fixed effect model and the mixed effect model, respectively. Column 1 contains the map for the intercept estimates $\hat{\beta}_0$, and columns 2 and 3 are for the slope estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.

The dual overlapping-cluster setting, where both clusters have their effects in the intercepts as well as in the slopes, is shown in Figures 5a and 6a. For the data simulated from this setting, we apply the multiple cluster detection proposed in Section 3.2. From Figure 5a, we see that both the fixed effect and the mixed effect method identify the true clusters well



(a) (b) (c)

Figure 5: Mean coefficient estimates: $\sigma_b^2 = \sigma_e^2 = 1$ ($\rho = \sigma_b^2/(\sigma_b^2 + \sigma_e^2) = 0.5$).



(a) (b) (c)

Figure 6: Mean squared error (MSE): $\sigma_b^2 = \sigma_e^2 = 1$ ($\rho = \sigma_b^2/(\sigma_b^2 + \sigma_e^2) = 0.5$).

enough. However, Figure 6a tells us that the mixed effect approach provides more consistent estimates, especially for the intercept with a small MSE, than the fixed effect model.

The settings of two or three non-overlapping clusters are shown in Figures 5b–5c and 6b–6c. Only one cluster has its effects in the intercepts as well as in the slopes whereas the other clusters exist in the intercepts only. We apply the two–stage multiple cluster detection proposed in Section 3.3 to the data simulated from these settings. In Figures 5b–5c, we see that the mixed effect model identifies the true clusters more clearly than the fixed effect method when there are cluster effects in the slopes. On the other hand, for those clusters in the intercepts only, the fixed effect model provides a better performance. This can be attributed to the random effect added to the intercept of the mixed effect model, which affects the identification of the intercept clusters. Nevertheless, we see in Figures 6b–6c that the fixed effect model has a higher MSE not only for the clusters in the slopes, but also for the intercept overall.

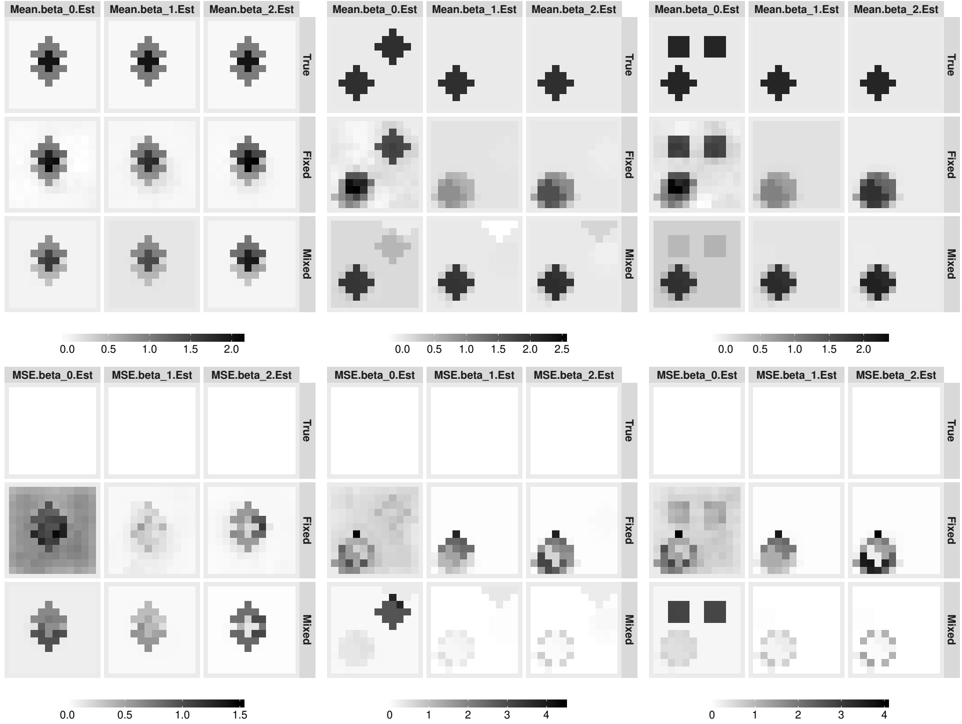


Figure 7: Top row: Mean coefficient estimates. Bottom row: Mean squared error (MSE). $\phi = 2$, $\sigma^2_{\mathcal{GP}} = 1$.

The results from the GP errors are illustrated in Figure 7. We omit the detailed discussion since these results are similar to those shown in Figures 5–6. Our proposed mixed effect model provides more consistent estimates for both the clusters in the slopes and the corresponding regression coefficients.

5. Data application

We apply our mixed effect model to the air quality data described in Section 2. For the raw CMAQ data, we take the logarithm of $\text{PM}_{2.5}$ because the exploratory analysis suggests that $\log\text{PM}_{2.5}$ shows stronger linear associations with the chosen covariates. Further, the variogram of the residuals from the linear regression fit of $\log\text{PM}_{2.5}$ with the covariates provides the spatial dependence range $\mathcal{R} = 500$ km (shown in Figure 8). To choose the appropriate subregion size, we perform an additional exploratory analysis with k-means clustering on the 448 grid cells from the raw CMAQ data. We apply k-means clustering to coordinates from the contiguous Albers equal-area conic projection to preserve the distance between two locations where $k = 7, 8, \ldots, 12$. The result shows that eight, nine or ten subregions can be considered based on the rule of thumb we suggested in the end of Section 4.1. Thus, we assign nine subregions of about 50 observations each (illustrated in Figure 8) via k-means clustering with coordinates. In each subregion, the maximum distance between two cells is around 300 km. Thus, the size of each subregion is proportional to 300×300 km^2 and this satisfies the rule of thumb for choosing the block size. Maps of the $\log\text{PM}_{2.5}$ (raw CMAQ) and $\text{PM}_{2.5}$ (fused CMAQ) data are also shown in Figure 8. Maps of the covariates are provided in Figure 9.

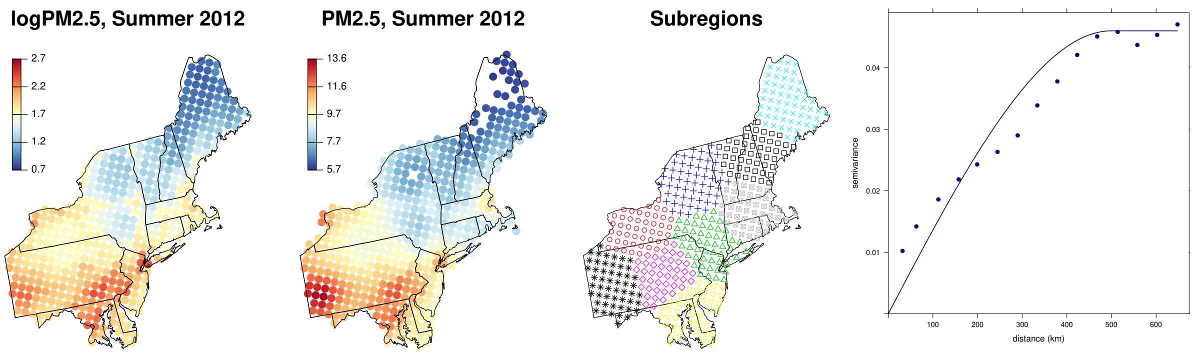


Figure 8: Log of monthly mean $\text{PM}_{2.5}$ (raw CMAQ) and monthly mean $\text{PM}_{2.5}$ (fused CMAQ) during the summer 2012, nine subregions from k-means clustering ($k = 9$), and the variogram of the residuals of the regression fit.

5.1. $\log\text{PM}_{2.5}$ and meteorological drivers

We apply the multiple cluster detection from Section 3.2 to the model, with $\log\text{PM}_{2.5}$ as the response variable and three meteorological drivers (TMP, RH, and WS) as the covariates. The set of candidate clusters $\mathcal{C}$ is predefined with a maximum radius $r_{\max} = 300$ km. A total

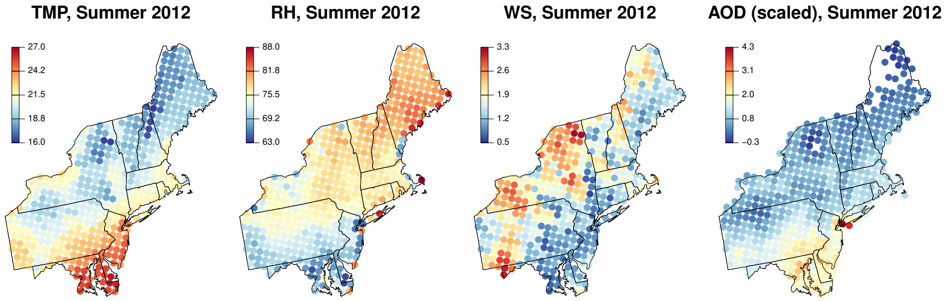


Figure 9: Monthly means of air temperature (TMP), relative humidity (RH), wind speed (WS) and aerosol optical depth (AOD) during the summer 2012.

of four clusters are detected at the significance level $\alpha = 0.05$. When we ignore the random effect for the subregions, the fixed effect model approach (Lee et al., 2017a) provides a total of ten clusters at $\alpha = 0.05$, suggesting that there might be some false positives.

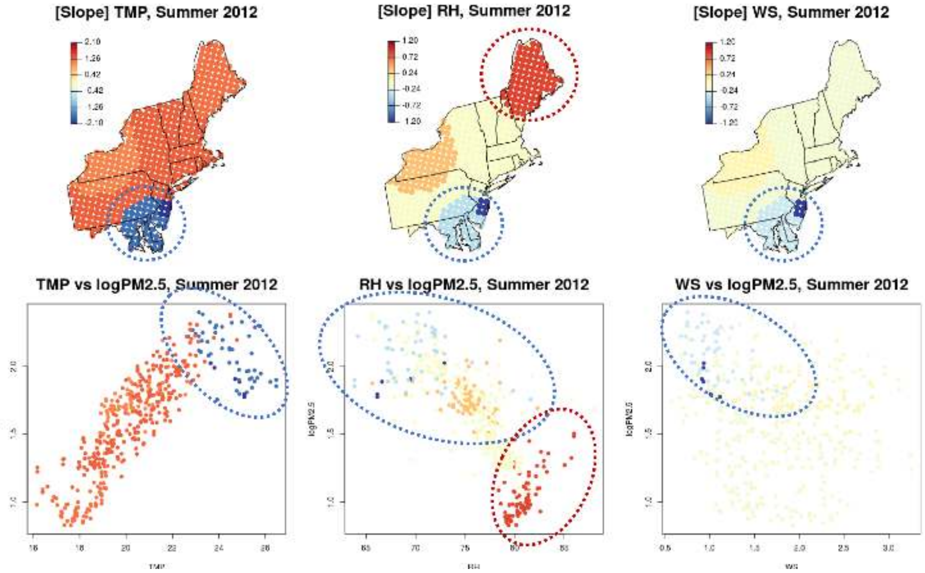


Figure 10: Top row: Maps of the slope estimates of each meteorological covariate. Bottom row: Scatter plots between each meteorological covariate and $\log\mathrm{PM}_{2.5}$. The colors in each scatter plot match the colors in the corresponding map.

Maps of the slope estimates for each meteorological covariate showing the spatial clusters detected via our mixed effect model approach are illustrated in Figure 10. The scatter plots between each covariate and the response variable $\log\mathrm{PM}_{2.5}$ are provided below the slope estimate maps. For each covariate, the colors in the map and the scatter plot match each other. The positive association between temperature and $PM_{2.5}$ is likely caused by the production of major components (e.g., sulfate) (Tai et al., 2010). Precipitation explains the negative relationship between relative humidity and $PM_{2.5}$. The wind speed slopes for $PM_{2.5}$ are also negative, though more weakly so due to the better atmospheric mixing that lead to more dispersion of pollutions.

We find two geographical regions that show distinctive patterns relative to the overall trend. The Chesapeake Bay area, indicated by blue dotted lines in Figure 10, has high temperature, low relative humidity, low wind speed, and a negative association between TMP and $PM_{2.5}$ compared to the rest of the region. One possible explanation is that these weather conditions are associated with a different wind direction. We found a positive correlation of 0.43 between temperature and wind blowing from east to west. That is, a high TMP in this region is associated with wind from the ocean moving inland, which reduces the $PM_{2.5}$ concentrations. Another possibility is that high temperatures in this region are associated with a higher boundary-layer height. Specifically, near bodies of water, high temperatures can break through an inversion, leading to lower $PM_{2.5}$ levels. Also, this area is known to have a complex meteorology in terms of air quality because of bay breezes, etc (Stauffer and Thompson, 2015).

The region around Maine, indicated by red dotted lines in Figure 10, has high relative humidity and shows a positive correlation between RH and $PM_{2.5}$. This is perhaps due to wind direction, similar to the Chesapeake Bay area. The majority of pollution and its precursors should come from the Ohio River Valley. In this area, there is a negative correlation between RH and wind speed of -0.32, and the correlation between RH and wind blowing from east to west is 0.44. Therefore, we see that the wind becomes too weak to push the pollutants from the Ohio River Valley out of the region around Maine.

Even when we apply our mixed effect model with eight or ten subregions, the results are qualitatively the same in the locations of two geographical hot spots, the Chesapeake Bay area and the region around Maine.

5.2. *$PM_{2.5}$ and AOD*

In this analysis, we consider a model where the fused CMAQ $\mathrm{PM}_{2.5}$ is the response variable and AOD is a single covariate. The fused product is used here to better reflect observed $\mathrm{PM}_{2.5}$ concentration. Both $\mathrm{PM}_{2.5}$ and AOD are centered and scaled based on the data for the whole U.S. to have zero means and standard deviations of one. We interpret the slopes as indicating the strength of the linear predictor (AOD) and the different intercepts as the biases or offsets. Thus, we apply the two–stage multiple cluster detection outlined in Section 3.3 to this single covariate model. The set of candidate clusters $\mathcal{C}$ is predefined with minimum and maximum radii of $r_{\min} = 100$ km and $r_{\max} = 300$ km, respectively. A total of four clusters are detected in the first stage, and six clusters are detected in the second stage, both at the significance level $\alpha = 0.05$.

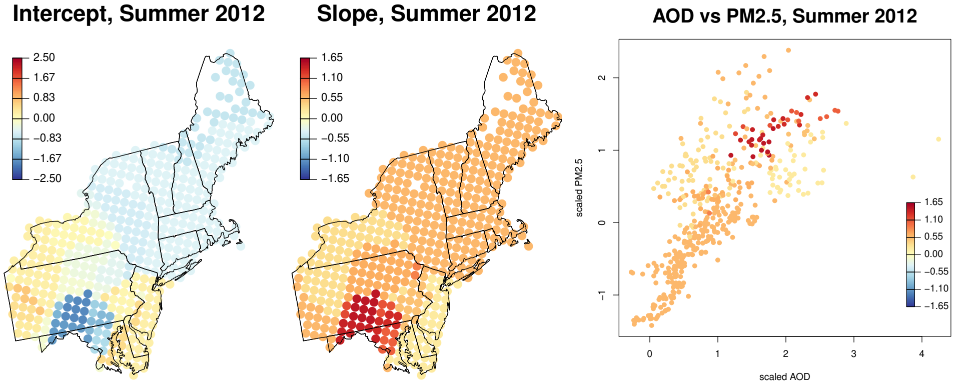


Figure 11: Maps of the intercept estimates and the slope estimates, and a scatter plot between AOD and $\mathrm{PM}_{2.5}$. The colors in the scatter plot match to the colors in the slope estimate maps.

The maps of the coefficient estimates and a scatter plot between AOD and $\mathrm{PM}_{2.5}$ (Figure 11) show the spatial clusters detected via our mixed effect model approach. The colors in the scatter plot match to the colors in the slope estimate maps. From this AOD analysis, we find that the overall slope is positive, as expected. Moreover, the slope estimate maps show potential heterogeneity in the relationship between $\mathrm{PM}_{2.5}$ and AOD, even within this small geographic region (Kloog et al., 2014).

6. Conclusion and discussion

We have developed a new approach to the detection of spatial clusters of regression coefficients within the framework of a mixed effect model. The novelty of this paper is the introduction of a spatial blockwise random effect within the subregions in the regression model, allowing our model to account for necessary spatial dependences in the data. Thus in the case of finding spatial hot spots, which aims to identify distinct features comparing to the background, our proposed method would be more easily applicable than other clustering or grouping methods that aim at splitting the space into several subregions. Further, by using a sequential searching scheme, our method is able to identify an unknown number of multiple clusters, including overlapping clusters.

Simulation studies evaluating the false positives, power, and identification of true clusters support that the proposed mixed effect model provides better results than the fixed effect model (Lee et al., 2017a). Even though we do not consider correlations across subregions, our approach performs satisfactorily, even when the data are from a GP model. When the true model is a GP model with different covariance functions or other spatial models, such as the CAR model (Besag, 1974), our model can be viewed as an approximation of the spatial dependency. Therefore, we need to choose the appropriate subregion size according to the effective spatial range, and a rule of thumb for the subregion size selection have provided in this paper. In practice, alternatively, if there is a reference variable which attribute could better explain the spatial dependency, then the exploratory spatial data analysis (ESDA) (Murray, 1999) could be considered to achieve subregions for our mixed effect model.

When we applied our cluster detection approach to real data from the Northeastern U.S., June–August, 2012, we found some geographical regions where the associations between the $\text{PM}_{2.5}$ concentrations and the weather covariates (temperature, relative humidity, and wind speed) were distinct from the overall trend. In practice, more complicated weather conditions may require additional meteorological drivers into the model as covariates. Our analysis for the $\text{PM}_{2.5}$ and AOD data tested here shows geographical heterogeneity in the AOD–$\text{PM}_{2.5}$ relationship, which might be helpful in the development of statistical models with spatially varying AOD effects when estimating $\mathrm{PM}_{2.5}$.

Acknowledgements

The authors thank the editor, an associate editor, and two referees for their insightful and constructive comments. The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST) and by US National Institutes of Health grant R01 ES027892. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Dr. James Mulholland, Dr. Armistead Russell, Dr. Yang Liu, and Ms. Niru Senthikumar for providing the CMAQ and satellite data.

References

Assunção R, Costa M, Tavares A, Ferreira S. 2006. Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine* **25**(5): 723–742.

Bai Y, Kang J, Song PX. 2014. Efficient pairwise composite likelihood estimation for spatialclustered data. *Biometrics* **70**(3): 661–670.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, Articles* **67**(1): 1–48.

Berger M Tutz G. 2018. Tree-Structured Clustering in Fixed Effects Models. *Journal of Computational and Graphical Statistics* **27**(2): 380–392.

Berrocal VJ, Gelfand AE, Holland DM. 2010a. A bivariate spacetime downscaler under space and time misalignment. *The Annals of Applied Statistics* **4**(4): 1942–1975.

—. 2010b. A Spatio-Temporal Downscaler for Output From Numerical Models. *Journal of Agricultural, Biological, and Environmental Statistics* **15**(2): 176–197.

—. 2011. SpaceTime Data fusion Under Error in Computer Model Output: An Application to Modeling Air Quality. *Biometrics* **68**(3): 837–848.

Besag J. 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2): 192–236.

Brunsdon C, Fotheringham AS, Charlton ME. 1996. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis* **28**(4): 281–298.

Chang HH, Hu X, Liu Y. 2013. Calibrating MODIS aerosol optical depth for predicting daily PM2.5 concentrations via statistical downscaling. *Journal Of Exposure Science And Environmental Epidemiology* **24**: 398.

Chang HH, Reich BJ, Miranda ML. 2012. Time-to-Event Analysis of Fine Particle Air Pollution and Preterm Birth: Results From North Carolina, 2001–2005. *American Journal of Epidemiology* **175**(2): 91–98.

Chu Y, Liu Y, Li X, Liu Z, Lu H, Lu Y, Mao Z, Chen X, Li N, Ren M, Liu F, Tian L, Zhu Z, Xiang H. 2016. A Review on Predicting Ground PM2.5 Concentration Using Satellite Aerosol Optical Depth. *Atmosphere* **7**(10).

Clark AB Lawson AB 2002. Spatio-temporal cluster modelling of small area health data. in *Spatial Cluster Modelling*, eds. Lawson AB Denison D, Boca Raton, FL: Chapman and Hall/CRC, pp. 235–258.

Darnell G, Georgiev S, Mukherjee S, Engelhardt BE. 2017. Adaptive Randomized Dimension Reduction on Massive Data. *Journal of Machine Learning Research* **18**(140): 1–30.

Dominici F, Peng RD, Bell ML, et Al. 2006. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA* **295**(10): 1127–1134.

Duczmal L Assunção R. 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis* **45**(2): 269–286.

Dunn OJ. 1961. Multiple Comparisons among Means. *Journal of the American Statistical Association* **56**(293): 52–64.

Fotheringham AS, Brunsdon C, Charlton ME. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, New York: Wiley.

Friedman J, Hastie T, Höfling H, Tibshirani R. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics* **1**(2): 302–332.

Gangnon RE. 2010a. A model for space-time cluster detection using spatial clusters with flexible temporal risk patterns. *Statistics in Medicine* **29**(22): 2325–2337.

—. 2010b. Local multiplicity adjustments for spatial cluster detection. *Environmental and Ecological Statistics* **17**(1): 55–71.

—. 2012. Local Multiplicity Adjustment for the Spatial Scan Statistic Using the Gumbel Distribution. *Biometrics* **68**(1): 174–182.

Gangnon RE Clayton MK. 2000. Bayesian Detection and Modeling of Spatial Disease Clustering. *Biometrics* **56**(3): 922–935.

—. 2003. A hierarchical model for spatially clustered disease rates. *Statistics in Medicine* **22**(20): 3213–3228.

—. 2004. Likelihood-based tests for localized spatial clustering of disease. *Environmetrics* **15**(8): 797–810.

—. 2007. Cluster detection using Bayes factors from overparameterized cluster models. *Environmental and Ecological Statistics* **14**: 69–82.

Gómez-Rubio V, Serrano PEM, Rowlingson B 2018. *DClusterm: Model-Based Detection of Disease Clusters*, R package version 0.2-1.

Goujon-Bellec S, , Demoury C, and Guyot-Goubin Aurélie, and Hémon Denis, and Clavel Jacqueline. 2011. Detection of clusters of a rare disease over a large territory: performance of cluster detection methods. *International Journal of Health Geographics* **10**(1): 53.

Grantham NS, Reich BJ, Liu Y, Chang HH. 2018. Spatial regression with an informatively missing covariate: Application to mapping fine particulate matter. *Environmetrics* **0**(0): e2499.

Grubesic A, Wei R, Murray AT. 2014. Spatial Clustering Overview and Comparison: Accuracy, Sensitivity, and Computational Expense. *Annals of the American Association of Geographers* **104**(6): 1134–1156.

Hodges JS Reich BJ. 2010. Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love. *The American Statistician* **64**(4): 325–334.

Huang L, Pickle LW, Das B. 2008. Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. *Statistics in Medicine* **27**(25): 5111–5142.

Hughes J Haran M. 2013. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(1): 139–159.

Jacob DJ Winner DA. 2009. Effect of climate change on air quality. *Atmospheric Environment* **43**(1): 51–63.

Jung I. 2009. A generalized linear models approach to spatial scan statistics for covariate adjustment. *Statistics in Medicine* **28**(7): 1131–1143.

Karl T Koss WJ. 1984. *Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983*, Asheville, N.C.: National Climatic Data Center.

Ke ZT, Fan J, Wu Y. 2015. Homogeneity Pursuit. *Journal of the American Statistical Association* **110**(509): 175–194.

Kim J, Yoon SC, Jefferson A, Zahorowski W, Kang CH. 2005. Air mass characterization and source region analysis for the Gosan super-site, Korea, during the ACE-Asia 2001 field campaign. *Atmospheric Environment* **39**(35): 6513–6523.

Kleinman KP, Abrams AM, Kulldorff M, Platt R. 2005. A model-adjusted spacetime scan statistic with an application to syndromic surveillance. *Epidemiology and Infection* **133**(3): 409–419.

Kloog I, Chudnovsky AA, Just AC, Nordio F, Koutrakis P, Coull BA, Lyapustin A, Wang Y, Schwartz J. 2014. A new hybrid spatio-temporal model for estimating daily multi-year PM2.5 concentrations across northeastern USA using high resolution aerosol optical depth data. *Atmospheric Environment* **95**: 581–590.

Kloog I, Koutrakis P, Coull BA, Lee HJ, Schwartz J. 2011. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric Environment* **45**(35): 6267–6275.

Kloog I, Nordio F, Coull BA, Schwartz J. 2012. Incorporating Local Land Use Regression And Satellite Aerosol Optical Depth In A Hybrid Model Of Spatiotemporal PM2.5 Exposures In The Mid-Atlantic States. *Environmental Science & Technology* **46**(21): 11913–11921.

Knorr-Held L Raßer G. 2004. Bayesian Detection of Clusters and Discontinuities in Disease Maps. *Biometrics* **56**(1): 13–21.

Kulldorff M. 1997. A spatial scan statistic. *Communications in Statistics, Part A* **26**: 1481–1496.

Kulldorff M, Huang L, Konty K. 2009. A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics* **8**: 58.

Kulldorff M, Huang L, Pickle L, Duczmal L. 2006. An elliptic spatial scan statistic. *Statistics in Medicine* **25**(22): 3929–3943.

Kulldorff M Nagarwalla N. 1995. Spatial disease clusters: Detection and inference. *Statistics in Medicine* **14**(8): 799–810.

Lawson AB. 2000. Cluster modelling of disease incidence via rjmcmc methods: a comparative evaluation. *Statistics in Medicine* **19**: 2361–2375.

Lawson AB, Choi J, Zhang J. 2014. Prior choice in discrete latent modeling of spatially referenced cancer survival. *Statistical Methods in Medical Research* **23**(2): 183–200.

Lee HJ, Liu Y, Coull BA, Schwartz J, Koutrakis P. 2011. A novel calibration approach of MODIS AOD data to predict PM2.5 concentrations. *Atmospheric Chemistry and Physics* **11**(15): 7991–8002.

Lee J, Gangnon RE, Zhu J. 2017a. Cluster detection of spatial regression coefficients. *Statistics in Medicine* **36**(7): 1118–1133.

Lee J, Gangnon RE, Zhu J, Liang J. 2017b. Uncertainty of a detected spatial cluster in 1D: quantification and visualization. *Stat* **6**(1): 345–359.

Lin PS, Kung YH, Clayton M. 2016. Spatial scan statistics for detection of multiple clusters with arbitrary shapes. *Biometrics* **72**(4): 1226–1234.

Liu Y, Sarnat JA, Kilaru V, Jacob DJ, Koutrakis P. 2005. Estimating Ground-Level PM2.5 in the Eastern United States Using Satellite Remote Sensing. *Environmental Science & Technology* **39**(9): 3269–3278.

Loh JM Zhu Z. 2007. Accounting for spatial correlation in the scan statistic. *The Annals of Applied Statistics* **1**(2): 560–584.

Ma Z, Liu Y, Zhao Q, Liu M, Zhou Y, Bi J. 2016. Satellite-derived high resolution PM2.5concentrations in Yangtze River Delta Region of China using improved linear mixed effects model. *Atmospheric Environment* .

MacQueen J 1967. Some methods for classification and analysis of multivariate observations. in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, Calif.: University of California Press, pp. 281–297.

Murray AT. 1999. Spatial analysis using clustering methods: Evaluating central point and median approaches. *Journal of Geographical Systems* **1**(4): 367–383.

Murray AT, Grubesic TH, Wei R. 2014. Spatially significant cluster detection. *Spatial Statistics* **10**: 103–116.

Neill DB. 2012. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(2): 337–360.

Paciorek CJ. 2010. The Importance of Scale for Spatial-Confounding Bias and Precision of Spatial Regression Estimators. *Statistical Science* **25**(1): 107–125.

Paciorek CJ, Liu Y, Moreno-Macias H, Kondragunta S. 2008. Spatiotemporal Associations between GOES Aerosol Optical Depth Retrievals and Ground-Level PM2.5. *Environmental Science & Technology* **42**(15): 5800–5806.

Pope CA Dockery DW. 2006. Health Effects of Fine Particulate Air Pollution: Lines that Connect. *Journal of the Air & Waste Management Association* **56**(6): 709–742.

Porter WC, Heald CL, Cooley D, Russell B. 2015. Investigating the observed sensitivities of air-quality extremes to meteorological drivers via quantile regression. *Atmospheric Chemistry and Physics* **15**(18): 10349–10366.

R Core Team 2017. *R: A Language and Environment for Statistical Computing.*, R Foundation for Statistical Computing, Vienna, Austria.

Reich BJ Bondell HD. 2010. A Spatial Dirichlet Process Mixture Model for Clustering Population Genetics Data. *Biometrics* **67**(2): 381–390.

Russell BT, Wang D, McMahan CS. 2017. Spatially modeling the effects of meteorological drivers of PM2.5 in the Eastern United States via a local linear penalized quantile regression estimator. *Environmetrics* **28**(5): e2448.

Samoli E, Peng R, Ramsay T, Pipikou M, Touloumi G, Dominici F, Burnett R, Cohen A, Krewski D, Samet J, Katsouyanni K. 2008. Acute Effects of Ambient Particulate Matter on Mortality in Europe and North America: Results from the APHENA Study. *Environmental Health Perspectives* **116**(11): 1480–1486.

Shen X Huang HC. 2010. Grouping Pursuit Through a Regularization Solution Surface. *Journal of the American Statistical Association* **105**(490): 727–739.

Shin S, Fine J, Liu Y. 2016. Adaptive Estimation with Partially Overlapping Models. *Statistica Sinica* **26**(1): 235–253.

Shu L, Jiang W, Tsui KL. 2012. A standardized scan statistic for detecting spatial clusters with estimated parameters. *Naval Research Logistics (NRL)* **59**(6): 397–410.

Stauffer RM Thompson AM. 2015. Bay breeze climatology at two sites along the Chesapeake bay from 19862010: Implications for surface ozone. *Journal of Atmospheric Chemistry* **72**(3): 355–372.

Tai APK, Mickley LJ, Jacob DJ. 2010. Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: Implications for the sensitivity of PM2.5 to climate change. *Atmospheric Environment* **44**(32): 3976–3984.

Tan Z, Roche K, Zhou X, Mukherjee S. 2018. Scalable Algorithms for Learning High-Dimensional Linear Mixed Models. *arXiv:1803.04431* .

Tang L Song PXK. 2016. Fused Lasso Approach in Regression Coefficients Clustering – Learning Parameter Heterogeneity in Data Integration. *Journal of Machine Learning Research* **17**(113): 1–23.

Tango T Takahashi K. 2005. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* **4**: 11.

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1): 91–108.

Tutz G Oelker MR. 2017. Modelling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures. *International Statistical Review* **85**(2): 204–227.

Wakefield J Kim A. 2013. A bayesian model for cluster detection. *Biostatistics* **14**(4): 752–765.

Waller LA, Hill EG, Rudd RA. 2006. The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. *Statistics in Medicine* **25**(5): 853–865.

Wang F, Wang L, Song PXK. 2016. Fused lasso with the adaptation of parameter ordering in combining multiple studies with repeated measurements. *Biometrics* **72**(4): 1184–1193.

Xu J Gangnon RE. 2016. Stepwise and stagewise approaches for spatial cluster detection. *Spatial and Spatio-temporal Epidemiology* **17**: 59–74.

Yan P Clayton MK. 2006. A cluster model for spacetime disease counts. *Statistics in Medicine* **25**(5): 867–881.

Yang S, Yuan L, Lai YC, Shen X, Wonka P, Ye J. 2012. Feature Grouping and Selection Over an Undirected Graph. *KDD : proceedings. International Conference on Knowledge Discovery & Data Mining* : 922–930.

Yin P Mu L. 2018. A hybrid method for fast detection of spatial disease clusters in irregular shapes. *GeoJournal* **83**(4): 693–705.

Yu W, Liu Y, Ma Z, Bi J. 2017. Improving satellite-based PM2.5 estimates in China using Gaussian processes modeling in a Bayesian hierarchical setting. *Scientific Reports* **7**(1): 7048.

Zhang T Lin G. 2009. Cluster Detection Based on Spatial Associations and Iterated Residuals in Generalized Linear Mixed Models. *Biometrics* **65**(2): 353–360.

Zhang Z, Assunção R, Kulldorff M. 2010. Spatial scan statistic adjusted for multiple clusters. *Journal of Probability and Statistics* **Article ID**: 11.

Table 1: Meteorological drivers

Name	Description
TMP	Air temperature at 2m (K)
RH	Relative humidity at 2m (%)
WS	Wind speed at 1000 mb (m/s)

Table 2: False positive: $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}_N(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma})$ and $\sigma^2 = \sigma_b^2 + \sigma_e^2$.

σ_b/σ_e	0.1	0.5	0.7	1.0	2.0
Fixed effect model	0.048	0.306	0.556	0.842	0.996
Mixed effect model	0.052	0.047	0.034	0.030	0.044

Table 3: False positive: $\boldsymbol{y}_{\mathcal{GP}} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_{\mathcal{GP}}$, where $\boldsymbol{\varepsilon}_{\mathcal{GP}} \sim \mathcal{GP}\big(\mathbf{0}, K(\cdot)\big)$.

	Gaussian covariance			exponential covariance		
ϕ	1	2	3	0.7	1.0	1.4
Fixed effect model	0.479	0.988	1.000	0.363	0.653	0.872
Mixed ($B = 3 \times 3$)	0.042	0.057	0.057	0.048	0.056	0.045
Mixed ($B = 6 \times 6$)	0.057	0.049	0.046	0.061	0.058	0.049