Modeling and Regionalization of China's PM2.5 Using Spatial-Functional Mixture Models

Article *in* Journal of the American Statistical Association · June 2020

DOI: 10.1080/01621459.2020.1764363

4 authors, including:

Decai Liang
Peking University
1 PUBLICATION **2** CITATIONS
SEE PROFILE

Xiaohui Chang
Oregon State University
11 PUBLICATIONS **34** CITATIONS
SEE PROFILE

Hui Huang
Sun Yat-Sen University
6 PUBLICATIONS **194** CITATIONS
SEE PROFILE

Modeling and Regionalization of China's $PM_{2.5}$ Using Spatial-Functional Mixture Models

Decai Liang, Haozhe Zhang, Xiaohui Chang, and Hui Huang

Abstract

Severe air pollution affects billions of people around the world, particularly in developing countries such as China. Effective emission control policies rely primarily on a proper assessment of air pollutants and accurate spatial clustering outcomes. Unfortunately, emission patterns are difficult to observe as they are highly confounded by many meteorological and geographical factors. In this study, we propose a novel approach for modeling and clustering $PM_{2.5}$ concentrations across China. We model observed concentrations from monitoring stations as spatially dependent functional data and assume latent emission processes originate from a functional mixture model with each component as a spatio-temporal process. Cluster memberships of monitoring stations are modeled as a Markov random field, in which confounding effects are controlled through energy functions. The superior performance of our approach is demonstrated using extensive simulation studies. Our method is effective in dividing China and the Beijing-Tianjin-Hebei region into several regions based on $PM_{2.5}$ concentrations, suggesting that separate local emission control policies are needed.

Keywords: Latent emission process; Model-based clustering; Markov random field; Environmental policies.

Decai Liang is a Ph.D. candidate at the School of Mathematical Science and Center for Statistical Science, Peking University, Beijing, P.R. China, 100871 (Email: `liangdecai@pku.edu.cn`). Haozhe Zhang is a Data & Applied Scientist at Microsoft Corporation, Redmond, WA 98052 (Email: `haozhe.zhang@microsoft.com`). Xiaohui Chang is an assistant professor of Business Analytics at the College of Business, Oregon State University, Corvallis, OR 97331 (Email: `xiaohui.chang@oregonstate.edu`). Hui Huang is a professor of Statistics at the School of Mathematics, Sun Yat-sen University, Guangzhou, P.R. China, 510275 (Email: `huangh89@mail.sysu.edu.cn`). For correspondence, please contact Hui Huang.

1 Introduction

Among all air pollutants, fine particulate matters with aerodynamic diameters less than 2.5 μm, also known as $PM_{2.5}$, are generally regarded as the most health-damaging because they easily penetrate the lung barrier and directly enter into the circulatory system. Numerous studies have shown that chronic exposure to high concentrations of $PM_{2.5}$ contributes to the risk of developing cardiovascular and respiratory diseases and lung cancers (Pope et al., 2002; Hoek et al., 2013; Lelieveld et al., 2015). Global Burden of Disease estimated that long-term exposure to $PM_{2.5}$ caused 4.2 million deaths worldwide in 2015, making it the fifth-ranked global risk factor that year (Cohen et al., 2017).

Due to the rapid industrialization and urbanization in recent decades, many areas of China have experienced the most chronic and severe air pollution in the world with the highest $PM_{2.5}$ levels (van Donkelaar et al., 2010). In the first quarter of 2013, extremely severe smog affected more than 800 million people in China. About 70% of the days in January registered daily average $PM_{2.5}$ concentrations that exceeded 75 $\mu g/m^3$ in numerous cities (Huang et al., 2014), more than seven times the World Health Organization's (WHO) recommended level of 10 $\mu g/m^3$. In response to the consistently poor air quality, the Chinese government directed massive efforts to assess air quality and evaluate the health impacts of air pollution for the entire country. For instance, real-time high-quality air pollutant concentration measurements have been collected from a large national monitoring network since 2013. This dataset quickly became one of the key pillars for the development of environmental policies and emission control strategies (Zhang et al., 2017). Unfortunately, the measurements may not provide an accurate depiction of the true characteristics of air pollutant emission, as the distribution and transmission patterns of $PM_{2.5}$ are highly confounded by factors including meteorological conditions, topography, local emissions, secondary aerosols, and regional transportation (Liang et al., 2015). These uncertainties, along with the large variability of $PM_{2.5}$ in space, bring challenges for assessing and monitoring $PM_{2.5}$ in China. Thus, to carry out the "coordinated inter-regional prevention and control efforts" initiated by the Chinese government (Li, 2015), a comprehensive statistical methodology that examines underlying emission patterns and incorporates spatio-temporal variations is urgently needed.

In this work, we propose a novel approach for modeling China's $PM_{2.5}$ data collected from the national monitoring network. The observed $PM_{2.5}$ concentration at each station is

modeled as functional data (Ramsay and Bernard, 2005). The unobserved true emission is assumed to be a latent process that employs a functional finite mixture model to account for spatial heteroscedasticity. Each component of the mixture model is a spatio-temporal process with a temporal functional principal component (FPC) expansion and spatially correlated FPC scores from different stations. Under our framework, stations are clustered into various regions based on their emission patterns. The cluster memberships (weights of components) follow a Markov random field (MRF) model, and topological factors are also exploited to define the similarity measures between stations. This approach makes inferences in both space and time while performing a model-based clustering for $PM_{2.5}$ emission.

In environmental science, cluster analysis has gained considerable attention in recent years due to its wide applicability to many real-life environmental problems. Clustering is frequently referred to as "regionalization" in the field because the outcomes are specifications of locations or regions. Researchers organize environmental units into homogeneous zones with the goal of establishing local environmental control strategies in different regions. Some applications that regionalization has played a significant role are dust storms (Qian et al., 2004), precipitation (Zhang et al., 2016), and air pollution (Wang et al., 2015). The conventional regionalization methods adopted in the field include empirical orthogonal functions (EOF) and its rotated version (REOF) (Zhang et al., 2012; Wang et al., 2015), which are basically spatial principal component analysis and the corresponding component rotations, respectively. Several new regionalization techniques have emerged, such as self-organizing maps (SOM) (Li et al., 2000) and k-means (Li et al., 2015). There are clear drawbacks to these approaches. The EOF and its extensions may be useful for initial data exploration but are unsuitable for investigation and interpretation of data characteristics. The determination of cluster boundaries using EOF is subjective. Moreover, it is very challenging to handle multi-scale data like ours through EOF, SOM or k-means. The most serious drawback of these approaches is that they either completely ignore or pay little attention to the intrinsic spatio-temporal structures of data, precluding accurate inferences for data with strong space- or time-varying features.

Cluster analysis has been well studied in the functional data analysis literature for its practical applications. For instance, James and Sugar (2003) developed a flexible model-based procedure for sparsely sampled longitudinal data. Chiou and Li (2007) proposed k-center functional clustering that is a functional version of k-means. Peng and Müller (2008)

introduced a distance-based method with multi-dimensional scaling. For high-dimensional functional data, some of the commonly used clustering methods largely rely on penalized likelihood (Pan and Shen, 2007), high-dimensional data clustering (Bouveyron and Jacques, 2011), an approximation for the density of functional random variables (Jacques and Preda, 2013), or wavelets (Giacofci et al., 2013). Nevertheless, these methods were all designed for independent curves and are unsuitable for spatially dependent data.

Earlier works on clustering spatial-functional data are relatively scarce. Romano et al. (2013) considered the spatial dependence among functions based on variogram models. Giraldo et al. (2012) proposed a hierarchical approach based on a dissimilarity matrix among curves. A recent technique introduced by Jiang and Serban (2012) incorporated an MRF into the modeling process to characterize spatial correlation and cluster dependence. MRF originated from the field of statistical physics and is a general version of the Ising model (Kindermann and Snell, 1980). In cluster analysis, especially for model-based methods, cluster memberships of stations are usually assumed to be random and their probabilities are modeled using a multinomial distribution. In this work, we employ the MRF-based approach to model both the spatial dependence and cluster memberships, and k-nearest neighbors combined with geographical information are also included for neighborhood definition.

This work contributes to the literature in the following five dimensions. First, we introduce a unified framework for joint modeling and clustering. The spatial dependence among the latent emission processes is embedded into the functional mixture model while the cluster memberships are assigned using an MRF model. Second, our method allows for heteroscedastic spatial dependence structures for different clusters, a much more realistic assumption compared to having the same spatial structure for all clusters in Jiang and Serban (2012), and greatly enhances the flexibility and applicability of our method. Third, this procedure has numerous practical advantages over other regionalization methods in real-life applications, including but not limited to, improved interpretability and clear cluster boundaries with stations connected within the same cluster, easy adaption to multi-scale data, possible extension to multi-pollutant regionalization, and more comprehensive statistical inferences on data features. Fourth, the numerical performance of this method is shown to be superior compared to others using extensive simulation studies. Last but not least, we also propose a Monte Carlo EM approach to compute the likelihood in the presence of multiple latent variables.

The structure of this paper is as follows. In Section 2, we introduce two $PM_{2.5}$ datasets from China and Beijing-Tianjin-Hebei (BTH) region, one of the most populated and polluted areas in the country, and discuss why regionalization is needed. We describe our method and the estimation procedures in Sections 3 and 4, respectively. The simulation studies are presented in Section 5. In Section 6, we demonstrate two examples of the application of the method using data from China and the BTH region. The paper concludes with a discussion in Section 7. Technical details of the Monte Carlo EM algorithm are in the Appendices. Other details including the datasets, R code, and additional results can be found in the Supplementary Material online.

2 Data Description

We analyze two datasets of different spatio-temporal scales: (1) city-level daily $PM_{2.5}$ concentration data for the entire country, and (2) station-level monthly concentration data collected from the BTH region. Performing regionalization for the entire country is highly challenging due to the widely diverse landscapes and drastically different meteorological conditions that are critical for modeling air pollutant data. In contrast, the smaller BTH region is more homogeneous in landscape and meteorological conditions, and this region is adopted to demonstrate the performance of the proposed methodology.

To smooth data variability and reduce extreme values, we apply a logarithmic transformation to pollutant measurements. For both datasets, the topographic information (including longitude, latitude, and elevation) is available. Pairwise distances between locations are calculated using the great circle distance that is defined as the shortest distance between two points on the surface of a sphere measured along the surface of the sphere (Porcu et al., 2016).

2.1 China

China's Ministry of Ecology and Environment has established a large monitoring network for air quality assessment since 2013. This national network had expanded to more than 1,500 monitoring stations in 338 cities in 2015 and 2016. Real-time measurements of major pollutants are continuously recorded and directly transferred to the China National Environmental Monitoring Center (CNEMC). Pollutant measurements are collected using con-

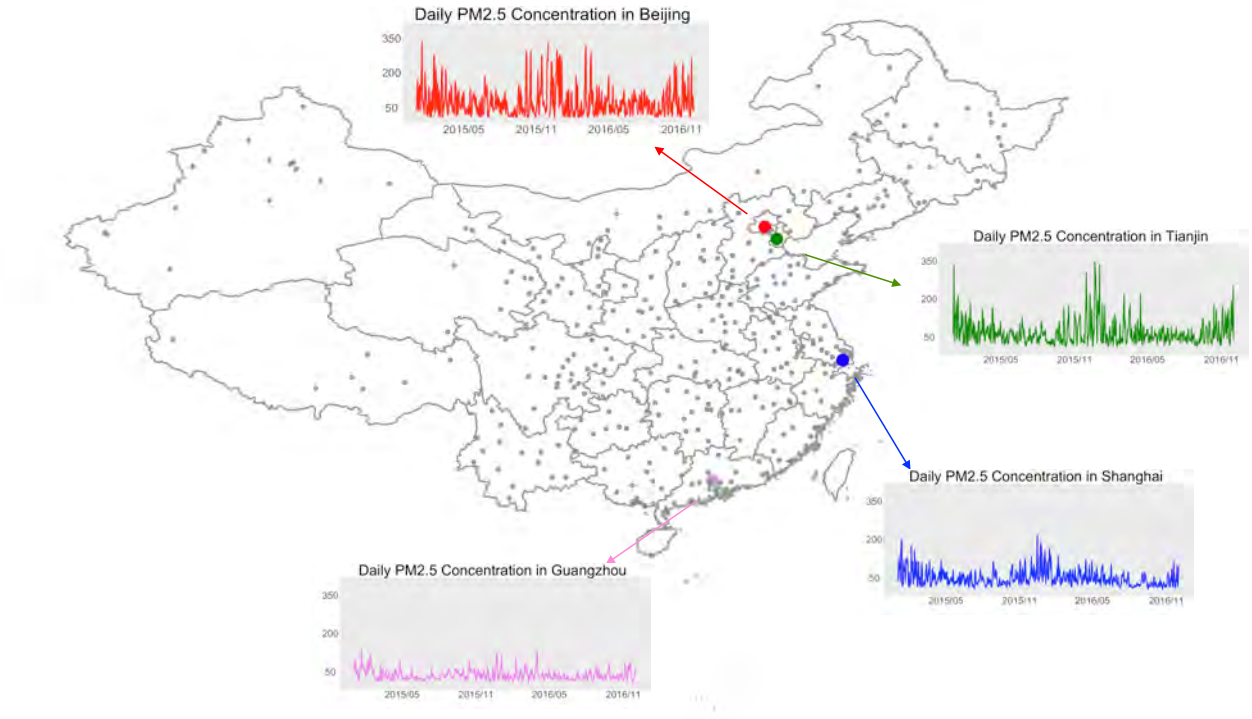

Figure 1: Locations of 338 cities (marked by dots) in China with monitoring stations, and the time series subplots of $PM_{2.5}$ daily concentrations ($\mu g/m^3$) of four megacities (Beijing, Tianjin, Shanghai, and Guangzhou) from January 2015 to December 2016.

tinuous automated methods through either tapered element oscillating microbalance or Beta ray attenuation (Wang et al., 2015). All equipment meet the standards of CNEMC.

Our city-level daily data are obtained by averaging hourly $PM_{2.5}$ concentrations from all monitoring stations in each city. A total of 731 measurements are available for each of the 338 cities from January 1, 2015, to December 31, 2016. We also remove 25 cities with low data quality from Xinjiang and Tibet. For the remaining 313 cities, missing data ($< 0.6\%$) are imputed using linear interpolation.

The locations of all 338 cities are presented in Figure 1. As an illustration, we also highlight four megacities (i.e., Beijing, Tianjin, Shanghai, and Guangzhou) and display their average daily observations from January 2015 to December 2016. The $PM_{2.5}$ time series of Beijing and Tianjin are similar and highly correlated, especially during winter, partially due to their close geographical proximity. On cold days, particle pollution is severe in North China as a result of coal-burning for heating, and the large variation is explained

by the frequent and strong northern winds that can blow away air pollutants. The $PM_{2.5}$ concentrations of Shanghai and Guangzhou are quite stable throughout the year but with some clear differences in their mean functions and variations. Refer to Liang et al. (2016) for a detailed discussion on $PM_{2.5}$ patterns and weather influences in these Chinese megacities. In this paper, we focus on feature extraction and separation of long-term and large-spatial-scale patterns of $PM_{2.5}$ across regions.

2.2 Beijing-Tianjin-Hebei

The Beijing-Tianjin-Hebei (BTH) region is one of the most polluted areas in the world, mostly due to the emission of primary pollutants and weather conditions (Chen et al., 2018). It is the national capital region of China, consisting of two of the most populated Chinese cities (Beijing and Tianjin) and eleven cities in Hebei Province. Serious environmental concerns have been raised when preparing for the 2008 Summer Olympics in Beijing, and many studies focusing on this area have followed since then (Lin et al., 2008; Wang et al., 2009; Xu et al., 2011).

There is a clear disparity in the air pollution level in the BTH region. For example, the northern cities, including Zhangjiakou, Chengde, and Qinhuangdao, are located in a mountainous area and marginally affected by emissions from factories and plants, whereas the cities to the south of Beijing are burdened with emissions from steel and cement factories, coal mines, and coking plants in Hebei Province. Therefore, the regionalization of the BTH region using air pollutant data would provide significant information for both policy makers and researchers.

The BTH region consists of a total of 73 national monitoring stations; see Figure 2 for a map of all stations and cities. These stations are a subset of the stations in the national network described in Section 2.1 and identical to those examined in earlier studies, e.g., Chen et al. (2018). Due to the high proportions of missing values in the first half of 2013, we limit our analysis to between June 2013 and December 2016. We also use monthly-averaged data instead of daily measurements to eliminate any local patterns and focus on the long-term trend. There are 43 months in total.

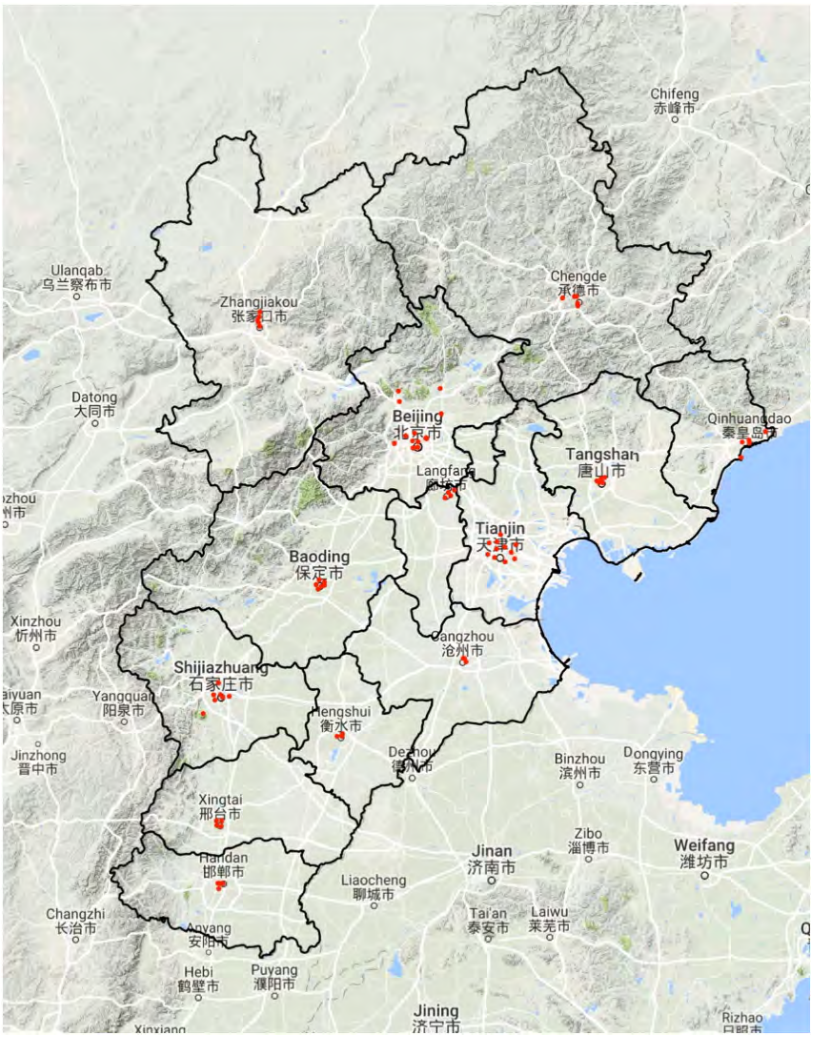

Figure 2: Locations of 13 cities and 73 monitoring stations (marked by dots) in the Beijing-Tianjin-Hebei (BTH) region. The lines represent the administrative borders of the cities.

3 Model and Assumptions

In this section, we propose a functional mixture model with spatially correlated random effects to measure the spatio-temporal dependencies of $PM_{2.5}$ concentrations. We also suppose the random functions of time are defined in a time domain $\mathcal{T}$ and sampled from locations in a spatial domain $\mathcal{D}$. Let $Y(\boldsymbol{s}_i, t_{ij})$ be the discrete observation at time $t_{ij} \in \mathcal{T}$ on the random curve of $PM_{2.5}$ concentration at location $\boldsymbol{s}_i \in \mathcal{D} \subset \mathbb{R}^2$, where $i = 1, \ldots, n$, and $j = 1, \ldots, m_i$. The total number of spatial locations is n, and the total number of discrete observations at location $\boldsymbol{s}_i$ is m_i.

We denote $\mathcal{Z}_i$ as the cluster membership of the random curve at location $\boldsymbol{s}_i$. In particular,

the membership $\mathcal{Z}_i$ is a random variable following a multinomial distribution with support $\{1, \ldots K\}$, and $\mathcal{Z}_i = k$ if the random curve at $\boldsymbol{s}_i$ belongs to the kth cluster, where K is the total number of clusters. An MRF model is employed for the cluster membership to account for spatial dependence intrinsically embedded in the data. For any location $\boldsymbol{s}_i$, we assume the marginal probability $P(\mathcal{Z}_i = k) = \pi_k$, where $\pi_k \geq 0$ for all k and $\sum_{k=1}^{K} \pi_k = 1$.

3.1 Reduced-Rank Functional Mixture Model for $PM_{2.5}$ Data

Conditional on the cluster membership $\mathcal{Z}_i = k$ for $k = 1, \ldots, K$, we assume the following functional mixture model for $PM_{2.5}$ measurements:

$$Y(\boldsymbol{s}_i, t_{ij}) \,\big|\, (\mathcal{Z}_i = k) = \mu_k(t_{ij}) + \eta_k(\boldsymbol{s}_i, t_{ij}) + \epsilon_{ij}, \tag{1}$$

where $\mu_k(\cdot)$ is the mean function for the kth cluster, $\eta_k(\cdot, \cdot)$ is a zero-mean spatio-temporal process on $\mathcal{D} \times \mathcal{T}$ representing a spatially correlated functional random effect for the kth cluster, and ϵ_{ij}'s are iid measurement errors with $\mathrm{E}(\epsilon_{ij}) = 0$ and $\mathrm{var}(\epsilon_{ij}) = \sigma^2_\epsilon$. We consider $\mu_k(\cdot) + \eta_k(\boldsymbol{s}_i, \cdot)$ as the latent and smooth random function of $PM_{2.5}$ concentrations at location $\boldsymbol{s}_i$ after removing the measurement errors. This framework allows for spatially correlated random effects, making it more general than the identical random effects for all clusters adopted in James and Sugar (2003). Also, we consider $\eta_k(\boldsymbol{s}, t)$ as a spatial extension of a temporal process with a standard Karhunen-Loève expansion:

$$\eta_k(\boldsymbol{s}, t) = \sum_{q=1}^{\infty} \gamma_{q,k}(\boldsymbol{s}) \psi_{q,k}(t), \tag{2}$$

where $\psi_{q,k}(\cdot)$'s are orthonormal eigenfunctions known as the functional principal components (FPC), and the functional principal component score $\gamma_{q,k}(\boldsymbol{s}) = \int_\tau \eta_k(\boldsymbol{s}, t) \psi_{q,k}(t) dt$ is the loading of $\eta_k(\boldsymbol{s}, t)$ on the qth principal component. We assume $\gamma_{q,k}(\boldsymbol{s})$'s are zero-mean and second-order stationary random fields that are independent across q and k.

Spatial structure of the function data is modeled through the spatial covariance between the FPC scores. More specifically,

$$\mathrm{cov}\{\gamma_{q,k}(\boldsymbol{s}_1), \gamma_{q',k'}(\boldsymbol{s}_2)\} = \begin{cases} \sigma^2_{\gamma,q,k} \rho(\boldsymbol{s}_1 - \boldsymbol{s}_2; \boldsymbol{\phi}_{q,k}), & \text{if } q = q' \text{ and } k = k', \\ 0, & \text{otherwise,} \end{cases}$$

where $\sigma^2_{\gamma,1,k} \geq \sigma^2_{\gamma,2,k} \geq \ldots > 0$, and $\rho(\cdot)$ is a spatial correlation function that depends on

the distance measure between two sites and some parameter vector $\boldsymbol{\phi}_{q,k}$. A wide range of spatial correlation functions can be employed to model the spatial dependence structure. For example, when the most popular Matérn function in geostatistics is selected, $\boldsymbol{\phi}_{q,k}$ includes a range parameter and a smoothness parameter (Matérn, 1960). It is worth mentioning that the parameters $\sigma^2_{\gamma,q,k}$'s, $\boldsymbol{\phi}_{q,k}$'s, and principal components $\psi_{q,k}(\cdot)$'s can vary across q and k, thus this model allows for a heteroscedastic random effect structure, which is a much more realistic assumption for many real-life applications compared to a homoscedastic random effect structure.

In practice, the equation in (2) is often approximated by truncating the series using the first Q leading functional principal components. Then, the reduced-rank version of the functional mixture model can be written as:

$$Y\left(\boldsymbol{s}_i, t_{ij}\right) \Big| \left(\mathcal{Z}_i = k\right) = \mu_k(t_{ij}) + \sum_{q=1}^{Q} \gamma_{q,k}(\boldsymbol{s}_i)\psi_{q,k}(t_{ij}) + \epsilon_{ij}, \tag{3}$$

where Q is considered as a tuning parameter of interest (Li et al., 2013). A data-driven method for selecting Q is discussed in more detail in Section 4.3.

3.2 Markov Random Field Model for Cluster Membership

Cluster membership is critical for specifying the joint distribution of the complete data in our framework. Following Jiang and Serban (2012) with some modifications, we assume the clustering configuration follows a locally dependent MRF model. The Markov property in space implies that the state space (namely the cluster membership) of any given location, $\boldsymbol{s}_i$, would depend on the states of its neighboring locations, denoted by ∂i. This assumption is reasonable because air pollutant data are usually locally dependent, and the degree of dependence is highly influenced by geographical factors, such as the elevation of mountains. To account for the spatial dependence in the cluster membership, we model the probability mass function of one cluster membership, conditioning on its neighbors, as the Gibbs distribution:

$$P\left(\mathcal{Z}_i = k | \boldsymbol{\mathcal{Z}}_{\partial i}\right) = \frac{\exp\{U_{ik}(\nu)\}}{N_i(\nu)}, \tag{4}$$

where $\boldsymbol{\mathcal{Z}}_{\partial i}$ is a vector of cluster memberships of the neighbor locations of $\boldsymbol{s}_i$, $U_{ik}(\nu) = \nu \sum_{i' \in \partial i} I(\mathcal{Z}_{i'} = k)$ where $\nu \geq 0$ is known as the energy function, $I(\cdot)$ is an indicator function, and $N_i(\nu) = \sum_{k=1}^{K} \exp\{U_{ik}(\nu)\}$ is a normalizing constant. The energy function $U_{ik}(\nu)$

determines the spatial pattern of the entire region. A large value of $U_{ik}(\nu)$ corresponds to a spatial pattern where many spatially connected locations belong to the same cluster, whereas a small $U_{ik}(\nu)$ implies a weak spatial dependence in the cluster membership. The parameter ν reflects the degree of interaction among the nearby sites in the MRF: a large ν represents a highly spatially dependent cluster membership, and $\nu = 0$ when there is no spatial dependence at all with an equal chance of belonging to any cluster. This idea of using the Gibbs distribution to model dependence structure originates from statistical physics (Kindermann and Snell, 1980) and has been frequently used in spatial statistics (Clifford, 1990).

4 Estimation and Implementation

4.1 Spline Approximation

We use polynomial splines to approximate and estimate the mean functions and eigenfunctions defined in Section 3. For simplicity, we assume the time domain is $\mathcal{T} = [0, 1]$. Let $\boldsymbol{B}(t) = \{b_1(t), \ldots, b_p(t)\}^{\mathrm{T}}$ be a spline basis with dimension p (de Boor, 2001). For simplicity, we use equally spaced knots in $\mathcal{T}$. We can approximately write $\mu_k(t) = \boldsymbol{B}^{\mathrm{T}}(t)\boldsymbol{\alpha}_k$ and $[\psi_{1,k}(t), \ldots, \psi_{Q,k}(t)] = \boldsymbol{B}^{\mathrm{T}}(t)\boldsymbol{\Theta}_k$, where $\boldsymbol{\alpha}_k$ and $\boldsymbol{\Theta}_k$ are $p \times 1$ and $p \times Q$ matrix of spline coefficients, respectively. According to Zhou et al. (2008) and Zhou et al. (2010), we impose identifiability restrictions on $\boldsymbol{B}(t)$ and $\boldsymbol{\Theta}_k$ such that: $\int \boldsymbol{B}(t)\boldsymbol{B}^{\mathrm{T}}(t)dt = \mathbf{I}_{p\times p}$ and $\boldsymbol{\Theta}_k^{\mathrm{T}}\boldsymbol{\Theta}_k = \mathbf{I}_{Q\times Q}$, and further require the first nonzero element of each column of $\boldsymbol{\Theta}_k$ to be positive. See Appendix 1 of Zhou et al. (2008) on how to construct a spline basis that satisfies the orthonormal constraints described above. For each station i, we use $\boldsymbol{\gamma}_i = [\gamma_{i1}, \ldots, \gamma_{iQ}]^{\mathrm{T}}$ to represent the spatial random effect. Conditional on $\{\mathcal{Z}_i = k, \boldsymbol{\gamma}_i\}$, the reduced-rank model (3) takes the form:

$$Y\left(\boldsymbol{s}_i, t_{ij}\right) | \left(\mathcal{Z}_i = k, \boldsymbol{\gamma}_i\right) = \boldsymbol{B}^{\mathrm{T}}(t_{ij})\boldsymbol{\alpha}_k + \boldsymbol{B}^{\mathrm{T}}(t_{ij})\boldsymbol{\Theta}_k\boldsymbol{\gamma}_i + \epsilon_{ij}. \tag{5}$$

A discussion on the selection of the spline basis is given in Section 4.3.

4.2 Monte Carlo EM Algorithm

As the standard approach for model-based functional clustering (e.g., James and Sugar, 2003), we use a likelihood-based procedure and treat both the memberships $\boldsymbol{\mathcal{Z}}$ and the spatial

random effects $\boldsymbol{\gamma}$ as latent variables. To overcome the computational challenges associated with the joint distribution of $(\boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma})$, we use the Monte Carlo EM (MCEM) algorithm based on Gibbs sampling for parameter estimation (Wei and Tanner, 1990).

With the spline approximation, the collection of parameters that need to be estimated becomes $\boldsymbol{\Omega} = \{\sigma_\epsilon^2, \nu, \boldsymbol{\Omega}_k;\ k = 1, \ldots, K\}$, where $\boldsymbol{\Omega}_k = \{\boldsymbol{\alpha}_k, \boldsymbol{\Theta}_k, \boldsymbol{\phi}_{q,k}, \sigma^2_{\gamma,q,k}\}$. We represent the model in a hierarchical matrix form:

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma} &= \widetilde{\boldsymbol{B}}^{\mathrm{T}} \widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\mathcal{Z}}} + \widetilde{\boldsymbol{B}}^{\mathrm{T}} \widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}} \boldsymbol{\gamma} + \boldsymbol{\epsilon}, \\ \boldsymbol{\gamma}|\boldsymbol{\mathcal{Z}} &\sim \text{Normal } \left(\mathbf{0}, \widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}\right), \\ \boldsymbol{\mathcal{Z}} &\sim \text{Markov random fields } (\nu). \end{aligned}$$

Here, we use $\boldsymbol{Y}_i = \{Y(\boldsymbol{s}_i, t_{i1}), \ldots, Y(\boldsymbol{s}_i, t_{im_i})\}^{\mathrm{T}}$ to denote the vector of all m_i discrete observations at location $\boldsymbol{s}_i$ and then define $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^{\mathrm{T}}, \ldots, \boldsymbol{Y}_n^{\mathrm{T}}\right)^{\mathrm{T}}$. Let $\widetilde{\boldsymbol{B}}^{\mathrm{T}}$ be an $\widetilde{n} \times np$ block diagonal matrix of spline basis functions $\boldsymbol{B} = \boldsymbol{B}^{\mathrm{T}}(t)$, where $\widetilde{n} = \sum_{i=1}^n m_i$ is the total sample size. For the cluster memberships, we write $\boldsymbol{\mathcal{Z}} = (\mathcal{Z}_1, \ldots, \mathcal{Z}_n)^{\mathrm{T}}$ as a vector of all cluster memberships with $\mathcal{Z}_i$ taking values from 1 to K. This means one realization of $\boldsymbol{\mathcal{Z}}$ is $(k_1, \ldots, k_n)^{\mathrm{T}}$, where $k_i \in \{1, \cdots, K\}$ for all i. We define the spline coefficients $\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\mathcal{Z}}} = \left(\boldsymbol{\alpha}_{\mathcal{Z}_1}^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_{\mathcal{Z}_n}^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}} = \mathrm{diag}\left(\boldsymbol{\Theta}_{\mathcal{Z}_1}, \ldots, \boldsymbol{\Theta}_{\mathcal{Z}_n}\right)$. The FPC scores are $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_1^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_n^{\mathrm{T}}\right)^{\mathrm{T}}$, and errors are $\boldsymbol{\epsilon} = \left(\boldsymbol{\epsilon}_1^{\mathrm{T}}, \ldots, \boldsymbol{\epsilon}_n^{\mathrm{T}}\right)^{\mathrm{T}}$, where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im_i})^{\mathrm{T}}$. The conditional covariance matrix of $\boldsymbol{\gamma}$ given the cluster membership $\boldsymbol{\mathcal{Z}}$ is $\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}$, which depends on the parameters $\boldsymbol{\phi}_{q,k}$ and $\sigma^2_{\gamma,q,k}$. More computational details on $\boldsymbol{\gamma}$ and $\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}$ are given in Appendix A.

Based on $f(\boldsymbol{Y}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma}) = f(\boldsymbol{Y}|\boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma}) f(\boldsymbol{\gamma}|\boldsymbol{\mathcal{Z}}) f(\boldsymbol{\mathcal{Z}})$ and the assumptions made in Section 3.1, we write out the log-likelihood for the complete data:

$$\ell(\boldsymbol{\Omega}; \boldsymbol{Y}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma}) = \log\left\{f\left(\boldsymbol{Y}|\boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma}; \widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\mathcal{Z}}}, \widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}}, \sigma_\epsilon^2\right)\right\} + \log\left\{f\left(\boldsymbol{\gamma}|\boldsymbol{\mathcal{Z}}; \widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}\right)\right\} + \log\left\{f(\boldsymbol{\mathcal{Z}}; \nu)\right\}. \quad (6)$$

We adopt the pseudo-likelihood proposed in Besag (1975) to approximate the last term in (6), i.e., $\log f(\boldsymbol{\mathcal{Z}}; \nu) \approx \sum_{i=1}^n \log\left\{f(\mathcal{Z}_i|\boldsymbol{\mathcal{Z}}_{\partial i}; \nu)\right\}$. Given $\boldsymbol{\mathcal{Z}} = (\mathcal{Z}_1, \ldots, \mathcal{Z}_n)^{\mathrm{T}}$ and $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_1^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_n^{\mathrm{T}}\right)^{\mathrm{T}}$, $\boldsymbol{Y}_i's$ are conditionally independent from a m_i-dim multivariate normal distribution with mean $\boldsymbol{B}^{\mathrm{T}}\left(\boldsymbol{\alpha}_{\mathcal{Z}_i} + \boldsymbol{\Theta}_{\mathcal{Z}_i}\boldsymbol{\gamma}_i\right)$ and variance σ_ϵ^2. As a result, the complete log-likelihood in (6) can be written as:

$$\ell(\boldsymbol{\Omega}; \boldsymbol{Y}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma}) \approx \log\left\{f\left(\boldsymbol{Y}|\boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma}; \widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\mathcal{Z}}}, \widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}}, \sigma_\epsilon^2\right)\right\} + \log\left\{f\left(\boldsymbol{\gamma}|\boldsymbol{\mathcal{Z}}; \widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}\right)\right\} + \sum_{i=1}^n \log\left\{f(\mathcal{Z}_i|\boldsymbol{\mathcal{Z}}_{\partial i}; \nu)\right\}$$

$$\propto -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} I(\mathcal{Z}_i = k)\left\{m_i \log\left(\sigma_\epsilon^2\right) + \left\|\boldsymbol{Y}_i - \boldsymbol{B}^{\mathrm{T}}\left(\boldsymbol{\alpha}_{\mathcal{Z}_i} + \boldsymbol{\Theta}_{\mathcal{Z}_i}\boldsymbol{\gamma}_i\right)\right\|^2 / \sigma_\epsilon^2\right\}$$
$$- \frac{1}{2} I(\mathcal{Z}_1 = k_1, \ldots, \mathcal{Z}_n = k_n)\left(\log\left|\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}\right| + \boldsymbol{\gamma}^{\mathrm{T}}\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}^{-1}\boldsymbol{\gamma}\right)$$
$$+ \sum_{i=1}^{n}\sum_{k=1}^{K} I(\mathcal{Z}_i = k)\left[U_{ik}(\nu) - \log\left\{N_i(\nu)\right\}\right]. \tag{7}$$

The complete likelihood depends on latent random variables $\boldsymbol{\mathcal{Z}}$'s and $\boldsymbol{\gamma}$'s, so it cannot be maximized directly. Instead, we treat these latent variables as missing data and estimate the unknown parameters using the EM algorithm by iterating between the E-step and the M-step until convergence. To determine the cluster memberships, we assign location $\boldsymbol{s}_i$ to the cluster k that maximizes that location's conditional probability $\pi_{k|i} = P(\mathcal{Z}_i = k|\boldsymbol{Y}_i)$.

4.2.1 E-Step

In the E-step, the expected log-likelihood is first calculated:

$$\mathcal{Q}(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}}) = \mathrm{E}\left\{\ell(\boldsymbol{\Omega}; \boldsymbol{Y}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{\Omega}_{\text{prev}})\right\}, \tag{8}$$

where $\boldsymbol{\Omega}_{\text{prev}}$ represents the value of the parameters from the previous EM iteration.

To calculate the expectation on the right side of (8), we would need the exact joint conditional distribution $f(\boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{\Omega}_{\text{prev}})$. But because $\boldsymbol{\mathcal{Z}}$ and $\boldsymbol{\gamma}$ are not conditionally independent, this joint distribution $f(\boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{\Omega}_{\text{prev}})$ cannot be directly approximated using the product of $f(\boldsymbol{\mathcal{Z}}|\boldsymbol{Y}, \boldsymbol{\Omega}_{\text{prev}})$ and $f(\boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{\Omega}_{\text{prev}})$. This is different from James and Sugar (2003). Instead, we use the MCEM of Wei and Tanner (1990) that approximates the conditional expectation using a Monte Carlo approximation and was shown to converge to the maximum likelihood estimate under some general regularity conditions (Chan and Ledolter, 1995). More specifically, we use Gibbs sampling (Geman and Geman, 1984) based on the full conditional distributions $f(\boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\Omega}_{\text{prev}})$ and $f(\boldsymbol{\mathcal{Z}}|\boldsymbol{Y}, \boldsymbol{\gamma}, \boldsymbol{\Omega}_{\text{prev}})$ to simulate the joint conditional distribution. Then $\mathcal{Q}(\cdot)$ in (8) can be estimated using the Monte Carlo average:

$$\widehat{\mathcal{Q}}(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}}) = \frac{1}{\overline{\mathrm{T}}}\sum_{\tau=1}^{\overline{\mathrm{T}}} \ell\left(\boldsymbol{\Omega}; \boldsymbol{Y}, \boldsymbol{\mathcal{Z}}^{(\tau)}, \boldsymbol{\gamma}^{(\tau)}\right), \tag{9}$$

where $\overline{\mathrm{T}}$ is the size of the Monte Carlo samples, and $\boldsymbol{\mathcal{Z}}^{(\tau)}$ and $\boldsymbol{\gamma}^{(\tau)}$ are samples from the conditional distribution $(\boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{\Omega}_{\text{prev}})$ using Gibbs sampling. More details on the E-step

are given in Appendix B.

4.2.2 M-Step

In the M-step, we update the parameter values to the values maximizing the approximated conditional expectation in (9). According to (7), we obtain:

$$\begin{aligned}\widehat{\mathcal{Q}}(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}}) = & -\frac{1}{2\overline{\overline{\mathrm{T}}}}\sum_{\tau=1}^{\overline{\mathrm{T}}}\sum_{i=1}^{n}\sum_{k=1}^{K} I\left(\mathcal{Z}_i^{(\tau)}=k\right)\left\{m_i \log\left(\sigma_\epsilon^2\right)+\left\|\boldsymbol{Y}_i-\boldsymbol{B}^{\mathrm{T}}\left(\boldsymbol{\alpha}_k+\boldsymbol{\Theta}_k\boldsymbol{\gamma}_i^{(\tau)}\right)\right\|^2/\sigma_\epsilon^2\right\} \\ & -\frac{1}{2\overline{\overline{\mathrm{T}}}}\sum_{\tau=1}^{\overline{\mathrm{T}}} I\left(\mathcal{Z}_1^{(\tau)}=k_1,\ldots,\mathcal{Z}_n^{(\tau)}=k_n\right)\left(\log\left|\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}^{(\tau)}}\right|+\boldsymbol{\gamma}^{(\tau)^{\mathrm{T}}}\widetilde{\boldsymbol{\Gamma}}^{-1}_{\boldsymbol{\mathcal{Z}}^{(\tau)}}\boldsymbol{\gamma}^{(\tau)}\right) \\ & +\frac{1}{\overline{\overline{\mathrm{T}}}}\sum_{\tau=1}^{\overline{\mathrm{T}}}\sum_{i=1}^{n}\sum_{k=1}^{K} I\left(\mathcal{Z}_i^{(\tau)}=k\right)\left\{U_{ik}^{(\tau)}(\nu)-\log N_i^{(\tau)}(\nu)\right\} \\ = & \ \widehat{\mathcal{Q}}_1\left(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}}\right)+\widehat{\mathcal{Q}}_2(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}})+\widehat{\mathcal{Q}}_3(\boldsymbol{\Omega}|\boldsymbol{\Omega}_{\text{prev}}). \end{aligned} \tag{10}$$

Because $\widehat{\mathcal{Q}}_1$, $\widehat{\mathcal{Q}}_2$, and $\widehat{\mathcal{Q}}_3$ depend on mutually disjoint collections of parameters in $\boldsymbol{\Omega}$, we can maximize them separately. To be more specific, $(\sigma_\epsilon^2, \boldsymbol{\alpha}_k, \boldsymbol{\Theta}_k)$ are updated by maximizing $\widehat{\mathcal{Q}}_1$, $(\boldsymbol{\phi}_{q,k}, \sigma^2_{\gamma,q,k})$ by $\widehat{\mathcal{Q}}_2$, and ν by $\widehat{\mathcal{Q}}_3$, respectively. The detailed M-step algorithm is provided in Appendix C.

4.3 Tuning Parameter Selection

There are three key tuning parameters in the model: number of clusters (K), number of FPCs (Q), and dimension of the spline basis (p). We develop a data-driven method to select them.

Bayesian Information Criterion (BIC) is one of the most popular methods for model selection. It adds a penalty term for the dimension of the parameter space to the log-likelihood function. Under our modeling framework, exact likelihood calculations using the standard EM algorithm can be quite challenging due to the presence of latent variables ($\mathcal{Z}$ and $\boldsymbol{\gamma}$). For a candidate model $\mathcal{M}$, we propose to approximate its log-likelihood by the Monte Carlo average $\widehat{\mathcal{Q}}_{\mathcal{M}}$ defined in (9), which is computed using Gibbs sampling in the final EM iteration. Note that this approximate likelihood coincides with the integrated likelihood introduced in Fraley and Raftery (2002), $f(\boldsymbol{Y}|\boldsymbol{\Omega}) = \int f\left(\boldsymbol{Y}|\boldsymbol{\mathcal{Z}}, \boldsymbol{\gamma}, \boldsymbol{\Omega}\right) f\left(\boldsymbol{\gamma}, \boldsymbol{\mathcal{Z}}|\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}, \nu\right) d\boldsymbol{\gamma} d\boldsymbol{\mathcal{Z}}$. Through numerical approximation and Monte Carlo samples, Fraley and Raftery (2002)

approximates the integrated likelihood by the BIC. Similarly, we define a Monte Carlo BIC for the model $\mathcal{M}$:

$$\text{Monte Carlo BIC}(\mathcal{M}) = -2\widehat{\mathcal{Q}}_{\mathcal{M}} + c_{\mathcal{M}} \cdot \log(\widetilde{n}), \tag{11}$$

where $c_{\mathcal{M}}$ is the number of parameters in $\mathcal{M}$. The tuning parameters are then selected simultaneously by minimizing (11).

Among all tuning parameters, the number of clusters (K) is the most critical and also the most challenging parameter to estimate. In real-life applications of this method, the BIC score gradually decreases with increasing K (James and Sugar, 2003; Zhou et al., 2010) and achieves its minimum when $K > 30$. However, such a large number of regions is not supported by any scientific evidence, and it becomes impractical to establish and implement a separate policy for each region. As a result, we use the expert knowledge of environmental scientists and do not consider K as a tuning parameter in our regionalization studies. Other implementation issues to expedite the model selection are addressed in the Supplementary Material S.1.

5 Simulation Studies

We carry out two simulation studies to compare the performance of the proposed spatial-functional clustering methodology against other methods in the literature. The first simulation study focuses on the homoscedastic case where different clusters share the same covariance structure, while the second study considers the heteroscedastic case where different clusters have different covariance structures. For each study, we first simulate a synthetic dataset, carry out several clustering methods, and then evaluate their performances. This procedure is repeated 100 times.

Two metrics are adopted to quantify the accuracy of assigning cluster memberships to curves. The first is the adjusted Rand index (ARI) (Hubert and Arabie, 1985), which is an improved version of the Rand index (Rand, 1971) with expected value 0 and bounded by ± 1. It measures the similarity between the true cluster membership and the clustering result obtained from a clustering method. A larger value of the adjusted Rand index implies a more accurate clustering method. The second metric is the standardized Root Mean Squared Error (RMSE), defined as $\text{RMSE} = \sqrt{\frac{||\boldsymbol{\mu}_k(t) - \widehat{\boldsymbol{\mu}}_k(t)||^2}{||\boldsymbol{\mu}_k(t)||^2}}$, measuring the accuracy of the estimated mean pattern $\widehat{\boldsymbol{\mu}}_k$ against the truth $\boldsymbol{\mu}_k$. We also evaluate the accuracy of the parameters and

functional principal components estimated from the model.

5.1 Homoscedastic Case

We simulate $n = 156$ points with coordinates $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{156} \in \mathbb{R}^2$ over a rectangular region (107°E $\sim$ 125°E, 28°N $\sim$ 43°N) in North China. The cluster memberships are simulated by generating a Markov random field using Gibbs sampling with $\nu = 0.5$, $K = 2$, and a multinomial distribution. The neighbors $\boldsymbol{\mathcal{Z}}_{\partial i}$ are chosen using 5 nearest neighbors. The synthetic data are generated from the following functional model:

$$Y_i(t)|(\gamma_i, \mathcal{Z}_i = k) = \mu_k(t) + \sum_{q=1}^{Q} \gamma_{i,q}\psi_q(t) + \epsilon_i(t), \tag{12}$$

for $i \in \{1, \ldots, n\}$, $t \in \{\frac{1}{30}, \frac{2}{30}, \ldots, \frac{29}{30}, 1\}$, $Q = 2$, and $k \in \{1, 2\}$. Following the simulation setup of the mean functions in Jiang and Serban (2012), we let the two cluster-dependent mean functions be $\mu_1(t) = \frac{1}{2}\exp(t)\cos(t)$ and $\mu_2(t) = \cos\left(\frac{5\pi}{2}t\right)$. The two functional principal component functions are orthogonalized by $\psi_1(t) = \sqrt{2}\sin(2\pi t)$ and $\psi_2(t) = b_3(t)$, where $b_3(t)$ is the third basis function of the 4-dimensional cubic spline basis $\boldsymbol{B}(t)$ defined in Section 4.1. The FPC scores, $\boldsymbol{\gamma}_q$'s, are generated with the isotropic exponential covariance structure $\text{cov}(\gamma_{i,q}, \gamma_{i',q}) = \sigma^2_{\gamma,q}\exp(-||s_i - s'_i||/\phi)$ with $\phi = 1$ and $(\sigma^2_{\gamma,1}, \sigma^2_{\gamma,2}) = (7, 2)$. The error term ϵ_i is a white-noise process with variance $\sigma^2_\epsilon = 0.4$. For each simulated dataset, we use the proposed Monte Carlo BIC to determine K, and the true $K = 2$ is correctly selected for 79% of the time. We closely monitor the convergence of the algorithm and provide additional trace plots for the MCEM iterations in the online Supplementary Material S.2.

We compare our proposed spatial-functional mixture model under a Markov random field (SFMM-MRF) with the following clustering methods:

(a) **k-means** clustering,

(b) **James**' method (James and Sugar, 2003), a classical functional mixture model assuming the same covariance structure for the random effects across all clusters,

(c) **Jiang**'s method (Jiang and Serban, 2012), spatial clustering method assuming a locally dependent Markov random field model for memberships,

(d) a functional mixture model assuming independence in both the random effects and cluster memberships (**FMM**),

Table 1: Means and standard deviations (in parentheses) of the adjusted Rand index (larger is better) and the RMSE (smaller is better) using different clustering methods, based on 100 simulations for the homoscedastic case.

	k-means	Jiang	James	FMM	FMM-MRF	SFMM	SFMM-MRF
ARI	0.600	0.453	0.673	0.889	0.858	0.903	0.909
	(0.307)	(0.304)	(0.378)	(0.299)	(0.325)	(0.282)	(0.278)
RMSE	0.639	0.959	0.905	0.418	0.433	0.393	0.392
	(0.324)	(0.254)	(0.380)	(0.323)	(0.325)	(0.265)	(0.268)

Table 2: Means and standard deviations of parameter estimates of 100 simulations using the proposed method (SFMM-MRF) for the homoscedastic case. The first row shows parameters and their true values.

Parameter	$\phi = 1$	$\nu = 0.5$	$\sigma^2_{\gamma,1} = 7$	$\sigma^2_{\gamma,2} = 2$
Mean	0.893	0.442	6.443	2.385
Standard Deviation	0.150	0.075	1.142	1.468

(e) a functional mixture model assuming independence in the random effects and Markov random fields for the cluster memberships (**FMM-MRF**), and

(f) a functional mixture model with spatially dependent random effects but independent cluster memberships (**SFMM**).

It is worth noting that the last three methods, namely FMM, FMM-MRF, and SFMM, are special cases of the proposed SFMM-MRF. The FMM approach can also be seen as an extension of James' method by using functional principal component analysis.

Table 1 summarizes the means and standard deviation of the adjusted Rand index and RMSE of all clustering methods based on 100 simulations. Our proposed model, SFMM-MRF, has the largest adjusted Rand index and the smallest RMSE, which shows that the proposed method outperforms the others. Table 2 demonstrates that our parameter estimation outlined in Section 4 performs reasonably well.

5.2 Heteroscedastic Case

With the same setup for spatial locations and cluster memberships, we generate another set of synthetic data from a heteroscedastic functional model:

$$Y_i(t)|(\mathcal{Z}_i = k) = \mu_k(t) + \sum_{q=1}^{Q} \gamma_{i,q,k}\psi_{q,k}(t) + \epsilon_i(t), \tag{13}$$

where $\sigma_\epsilon^2 = 0.4$, $\mu_1(t) = \frac{1}{2}\exp(t)\cos(t)$, and $\mu_2(t) = \cos(\frac{5\pi}{2}t)$, same as in Section 5.1. The heteroscedastic model is different from the homoscedastic model in that the random effects and FPCs of the former depend on the cluster membership, whereas those of the latter do not. To include the heteroscedasticity with reasonable complexity, we let $\psi_{1,k=1}(t) = b_2(t)$ and $\psi_{2,k=1}(t) = b_3(t)$ for Cluster 1, and $\psi_{1,k=2}(t) = b_4(t)$ and $\psi_{2,k=2}(t) = b_1(t)$ for Cluster 2. Here, $\{b_1(t), b_2(t), b_3(t), b_4(t)\}^{\mathrm{T}}$ forms a 4-dimensional cubic spline basis as defined in Section 4.1. The spatial covariance functions of the FPC scores are $\mathrm{cov}(\gamma_{i,q,k}, \gamma_{i',q',k}) = \sigma^2_{\gamma,q,k}\exp(-||s_i - s_i'||/\phi)$ where $\phi = 1$, $(\sigma^2_{\gamma,1,1}, \sigma^2_{\gamma,2,1}) = (4, 1)$, and $(\sigma^2_{\gamma,1,2}, \sigma^2_{\gamma,2,2}) = (2, 0.5)$.

For the heteroscedastic case, we compare the heteroscedastic spatial-functional clustering under a Markov random field (HSFMM-MRF) with others. One special case of HSFMM-MRF is the heteroscedastic functional spatial clustering with spatial dependence in the random effects and independence of cluster memberships (HSFMM). For each of the 100 simulations, we compare five of the seven clustering methods described in the previous subsection and substitute the remaining two methods using their corresponding heteroscedastic counterparts. The adjusted Rand index and the RMSE are summarized in Table 3. On average, our proposed method HSFMM-MRF produces the largest adjusted Rand index and the lowest RMSE compared to other methods, demonstrating its superiority over the rest. Moreover, the last three columns in Table 3 are quite similar in values, indicating that the framework of spatial-functional mixture model is generally robust to theassumptions of spatial structures. In other words, even when the spatial structure of data is misspecified, we can still obtain relatively good parameter estimates and clustering results. In real data analy-sis, this flexibility allows us to choose different methods for different purposes - either the more complicated one for feature specification or the simpler one for computational advan-tages. Table 4 consists of a summary of the parameter estimates obtained from the proposed method. Again, they perform reasonably well. We also display both the true and the estimated mean functions and functional principal components in Figure 3, which shows the estimated functions are very close to the true ones.

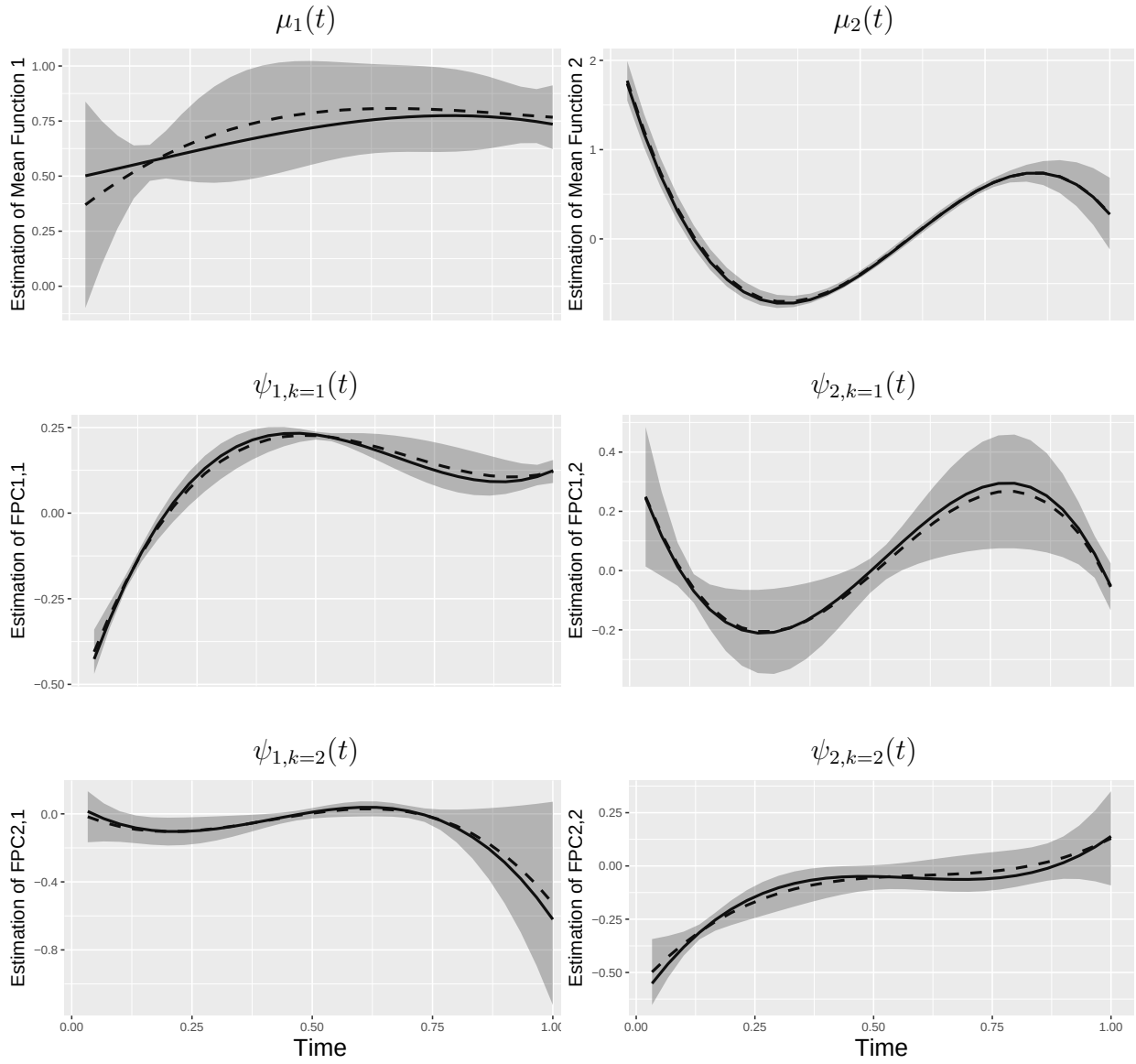

Figure 3: True (solid lines) and estimated (dashed lines) functions of two mean functions (top panels) and four functional principal components (middle and bottom panels) for the heteroscedastic case. The shaded areas are bound by the 5th and 95th percentiles of the estimated functions.

6 Data Analysis

6.1 China

For the city-level daily spatio-temporal $PM_{2.5}$ data of 313 cities from 2015 to 2016 (refer to Section 2.1), we apply the proposed method, more specifically the SFMM-MRF, to carry out a regionalization analysis. To model the mean functions and eigenfunctions, we use cubic

Table 3: Means and standard deviations (in parentheses) of the adjusted Rand index and the RMSE using different clustering methods based on 100 simulations for the heteroscedastic case.

	k-means	James	Jiang	FMM	FMM-MRF	HSFMM	HSFMM-MRF
ARI	0.824	0.824	0.853	0.896	0.917	0.931	0.933
	(0.103)	(0.105)	(0.152)	(0.122)	(0.100)	(0.054)	(0.056)
RMSE	0.321	0.629	0.630	0.303	0.298	0.283	0.284
	(0.115)	(0.097)	(0.114)	(0.116)	(0.108)	(0.098)	(0.097)

Table 4: Means and standard deviations of parameter estimates of 100 simulations using the proposed method (HSFMM-MRF) for the heteroscedastic case.

Parameter	$\phi = 1$	$\nu = 0.5$	$\sigma^2_{\gamma,1,1} = 4$	$\sigma^2_{\gamma,2,1} = 1$	$\sigma^2_{\gamma,1,2} = 2$	$\sigma^2_{\gamma,2,2} = 0.5$
Mean	0.729	0.437	3.440	0.875	1.844	0.509
Standard Deviation	0.267	0.071	0.774	0.244	0.468	0.170

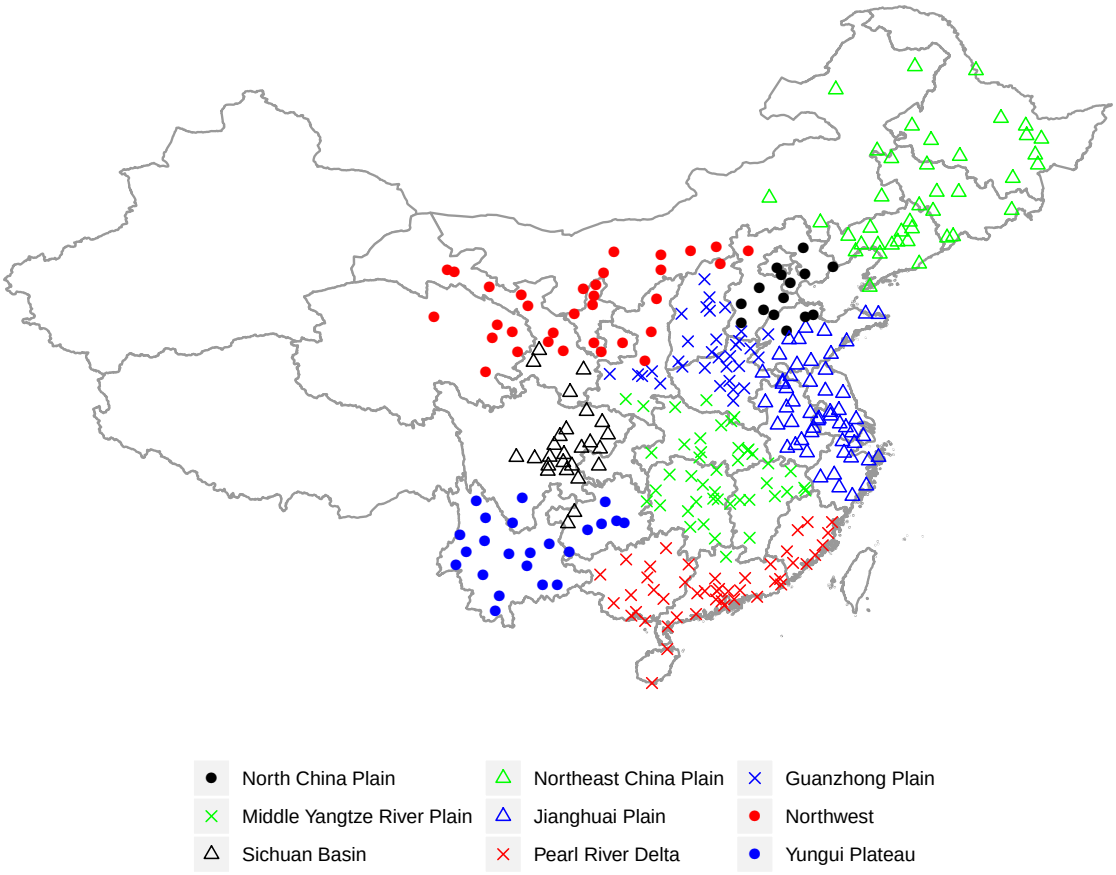

Figure 4: Regionalization map of China's $PM_{2.5}$ using daily average $PM_{2.5}$ concentrations of 313 cities from January 2015 to December 2016. Each colored symbol represents a different cluster. The total number of clusters is 9.

B-spline with 16 equally spaced interior knots. To model the cluster membership, for any given city, we consider all cities within a 500 km radius as its neighbors since the correlation of air pollution patterns at sites more than 500 km apart is generally weak (Gao et al., 2011).

China has a topographically diverse landscape, including the highest mountains and the largest plateau on Earth, which affects the movement of many air pollutants from one place to another (Bryan and Adams, 2002). Two adjacent monitoring sites separated by a high mountain may have distinctive pollution patterns and thus belong to two clusters. This "mountain effect" coupled with other geographical factors can be easily incorporated into our model by modifying the standard Euclidean distance with a spatial deformation to obtain a "geographical distance." For example, the distance between two sites that are separated by a high mountain can be set to be much greater than their Euclidean distance while other geometric properties of the Euclidean distance are also retained. The change in the definition of the distance metric may lead to an alteration of neighbors. This method has been implemented in earlier works, including Sampson and Guttorp (1992), and Anderes and Stein (2008), among others.

Alternatively, in our study, we extend the energy function $U_{ik}(\nu)$ defined in Section 3.2 by introducing a function $g(\cdot,\cdot)$ to model the geographical covariates between a site and its neighbors. More specifically, we define $\widetilde{U}_{ik}(\nu) = \nu \sum_{i' \in \partial i} g(\boldsymbol{s}_i, \boldsymbol{s}_{i'}) I(\mathcal{Z}_{i'} = k)$ and $\widetilde{N}_i(\nu) = \sum_{k=1}^{K} \exp\{\widetilde{U}_{ik}(\nu)\}$. The function $g(\boldsymbol{s}_i, \boldsymbol{s}_{i'})$ captures the geographical information between site $\boldsymbol{s}_i$ and its neighboring site $\boldsymbol{s}_{i'}$. For instance, consider a simple case where we set d as the altitude threshold between $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i'}$. If the largest altitude between $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i'}$ is greater than d implying the presence of an extremely high mountain between them, then $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i'}$ should not be in the same cluster. For this scenario, $g(\cdot,\cdot)$ can be written as

$$g(\boldsymbol{s}_i, \boldsymbol{s}_{i'}) = \begin{cases} 0, & \text{if the largest altitude between } \boldsymbol{s}_i \text{ and } \boldsymbol{s}_{i'} \text{ is greater than } d, \\ 1, & \text{otherwise.} \end{cases}$$

Then, the conditional probability for the cluster membership becomes

$$P(\mathcal{Z}_i = k | \boldsymbol{\mathcal{Z}}_{\partial i}) = \frac{\exp\left\{\widetilde{U}_{ik}(\nu)\right\}}{\widetilde{N}_i(\nu)} = \frac{\exp\left\{\nu \sum_{i' \in \partial i} g(\boldsymbol{s}_i, \boldsymbol{s}_{i'}) I(\mathcal{Z}_{i'} = k)\right\}}{\sum_{k=1}^{K} \exp\left\{\nu \sum_{i' \in \partial i} g(\boldsymbol{s}_i, \boldsymbol{s}_{i'}) I(\mathcal{Z}_{i'} = k)\right\}}.$$

In our analysis, d is set to be 1 km.

The formation of $PM_{2.5}$ is very complex with many important contributing factors in-

cluding meteorological conditions, population, local industry, traffic, instantaneous energy consumption, secondary chemical reactions in the atmosphere, and others. $PM_{2.5}$ comprises a list of primary and secondary components that can also contribute to the study of emission sources and patterns, but unfortunately their concentrations are not collected in the monitoring network. In this study, the only accessible data are the $PM_{2.5}$ concentrations obtained from the monitoring stations, and they provide partial information about local emission characteristics. We try to cluster cities based on the spatial-temporal trends observed from the $PM_{2.5}$ concentrations so that more effective regional policies and strategies than the current practices may be established and implemented.

Figure 4 displays the clustering results when the number of clusters (K) is set to 9. This choice of K follows the recommendation of environmental scientists (Wang et al., 2015). We observe a clear spatial clustering with several distinct geographical regions. For example, North China Plain, Yangtze River Delta, Pearl River Delta, and Sichuan Basin are all classified into separate clusters. These regions coincide with the list published by CNEMC where air pollution is severe, and prevention and control strategies are needed (China's State Council, 2013). Another study done by Wang et al. (2015) also reports similar regions, but our method defines regions with clearer boundaries. In addition, our method successfully combines sites that are geographically far apart yet showing similar $PM_{2.5}$ patterns into one cluster. For example, the Northwest cluster includes cities across five provinces: northern Shanxi, middle Inner Mongolia, Ningxia, northern Gansu, and eastern Qinghai. These cities are mostly resource-based, and their winter weather conditions are largely influenced by the northwest monsoons. We also include the clustering results from other methods in the Supplementary Material S.4 online for comparison. These methods appear to have much less clear boundaries in their regionalization maps. This "clear spatial boundary" effect is also one of the merits of the proposed method and can be mainly attributed to the Markov random field employed in this approach.

The estimated mean functions of all clusters are displayed in Figure 5. Despite the temporal trends varying across regions, there is a consistent "W" shape in almost all regions. $PM_{2.5}$ concentrations are generally higher in winter than in summer, mostly from coal burning in many parts of China. This phenomenon becomes less obvious in Pearl River Delta and Yungui Plateau of southern China where winters are relatively warm. Moreover, the estimated four leading functional principal components are shown in Figure 6. A spike in

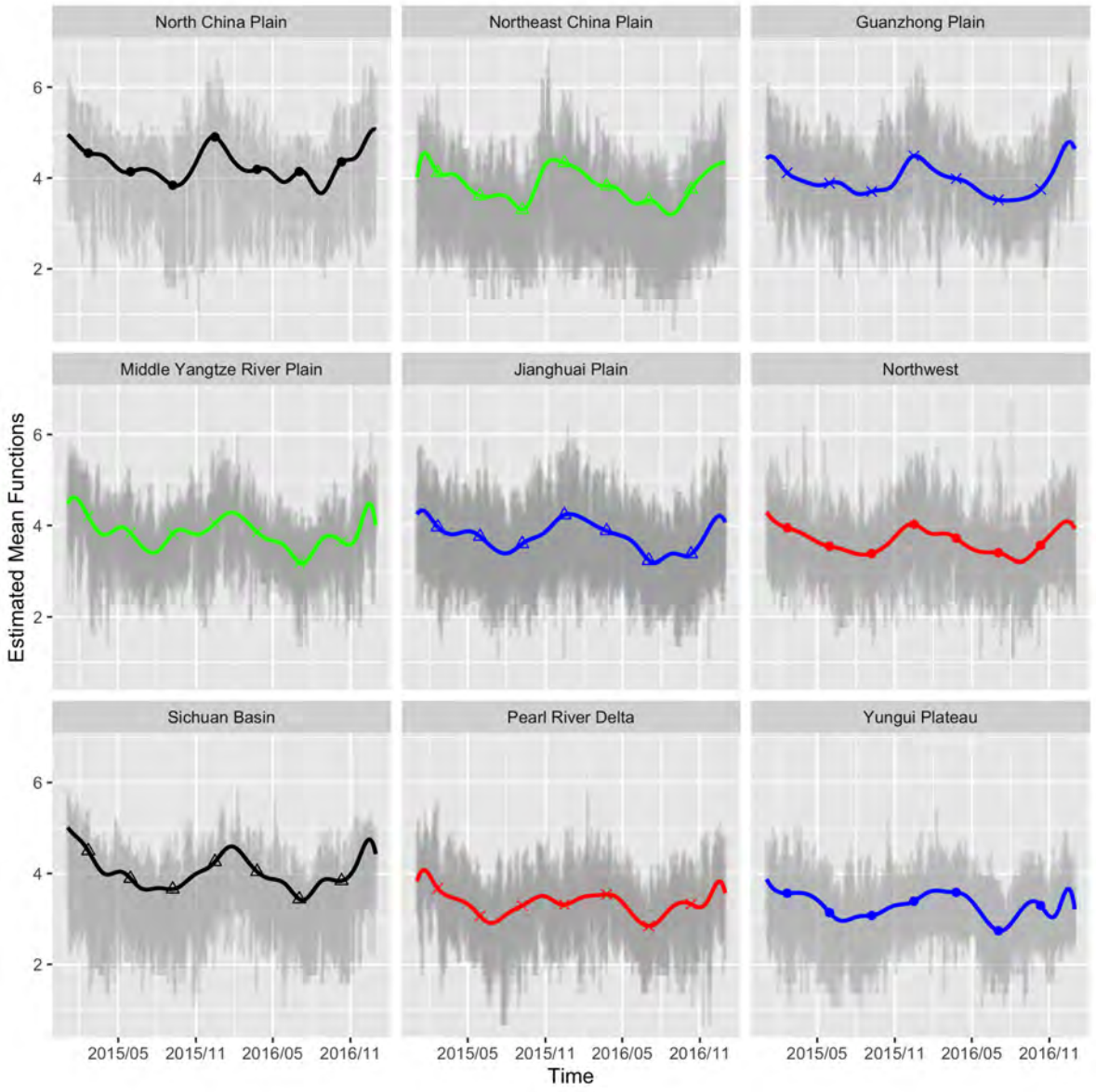

Figure 5: Estimated mean functions of nine clusters using China's daily-averaged $PM_{2.5}$ concentrations of 313 cities from 2015 to 2016. Same legend as in Figure 4. The observed $PM_{2.5}$ concentrations are marked in grey.

the first component corresponds to an increased $PM_{2.5}$ level in the winter of 2016 across all regions that is not completely captured by the mean functions. The remaining components also show some seasonality providing further evidence that $PM_{2.5}$ pollution is milder in summer and worse in winter.

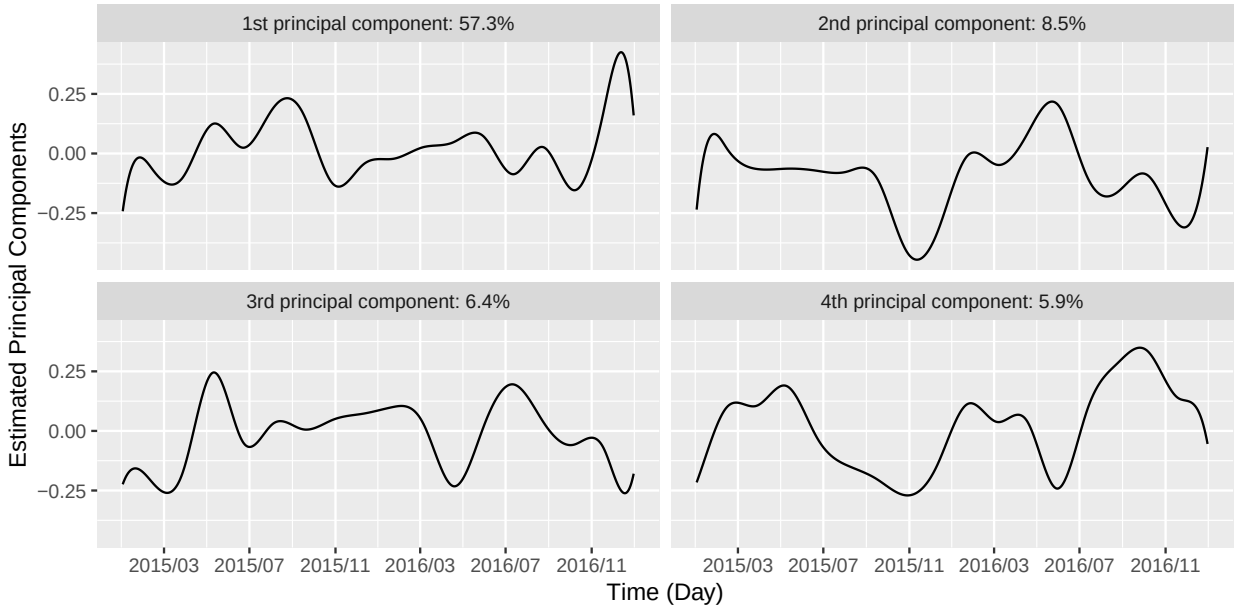

Figure 6: Estimated four leading functional principal components for China's daily $PM_{2.5}$ data. The first four eigenvalues explain a total of 78.1% of variation (5: 82.1%; 6: 85%; 7: 87.7%; 8: 90.0%; 9: 92.2%; 10: 93.8%).

6.2 Beijing-Tianjin-Hebei

We apply our method to analyze the monthly-averaged $PM_{2.5}$ concentrations of the 73 stations in the BTH region from June 2013 to December 2016; refer to Section 2.2 for more details on the data. A cubic spline with eight equally spaced interior knots is considered. Due to the small area and flat terrain of the BTH region, we use the five nearest neighbors to model the Markov random fields.

We divide all stations into three regions and present the results in Figure 7. The number of clusters $K = 3$ follows another study of the BTH region (Chen et al., 2018). Our results show that stations with the lowest $PM_{2.5}$ are clustered in the north – these are the stations from Zhangjiakou, Chengde, and Qinhuangdao. The most severely polluted stations are in the southern BTH area, including Baoding, Shijiazhuang, Hengshui, Xingtai, and Handan that are frequently on the list of most polluted cities in China. Stations with moderate pollution, including those from Beijing, Langfang, Tianjin, Tangshan, and Cangzhou, are clustered together. These three regions are in agreement with Chen et al. (2018). The northern mountainous stations are separated from the rest, demonstrating the significance of the "mountain effect." The estimated mean functions of the three regions are plotted in Figure 8 where the most polluted southern region has the worst pollution in winters. This is

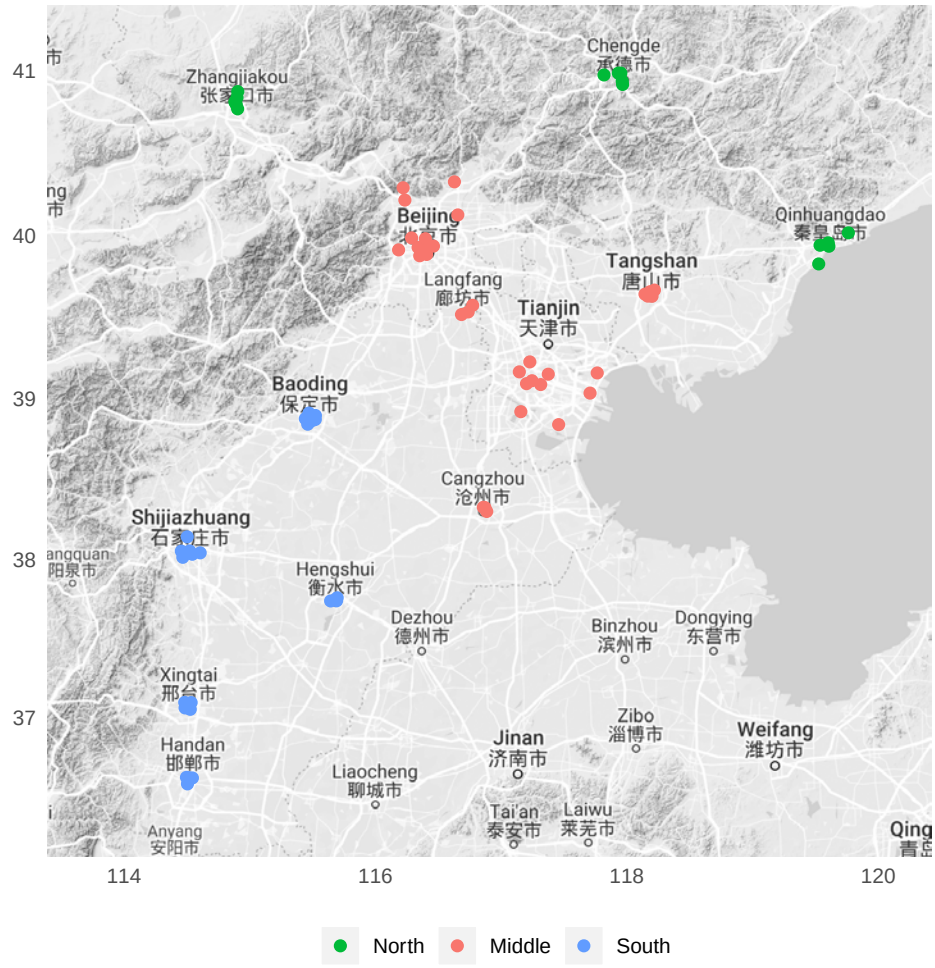

Figure 7: Regionalization map of the BTH region using monthly-averaged $PM_{2.5}$ concentrations of 73 stations from June 2013 to December 2016. Each color represents a different cluster.

in accord with many other researchers' findings. Because the three mean functions are well separated from each other, we recommend separate pollution control strategies in different regions. All regions show a slowly descending trend over time indicating the positive effects of China's pollution-control efforts in recent years.

6.3 Clustering Results for Demeaned Data

It is worth noting that functional clustering methods do not only group observations based on the scales of the data (e.g., the three mean functions in Figure 8 have distinct scales), but more importantly, they cluster time-dependent data according to the shapes of the temporal patterns. Following the recommendation of one reviewer, we also demean the data to bring

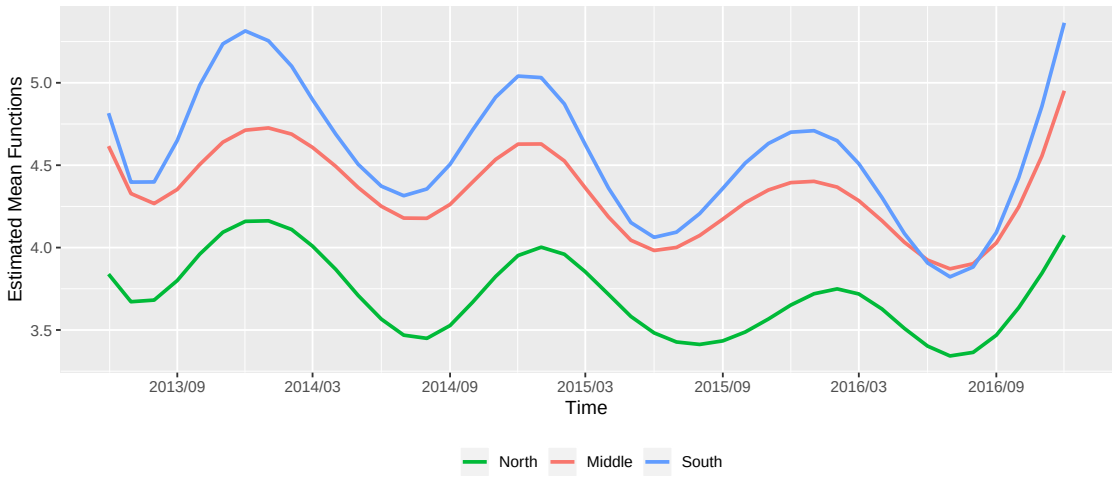

Figure 8: Estimated mean functions of three clusters using monthly-averaged $PM_{2.5}$ concentrations of 73 stations from the BTH region between June 2013 and December 2016. Each color represents a different cluster.

all stations to the same scale and then apply the proposed methodology to identify clusters of stations according to their temporal patterns. Both the global mean function $\mu_0(t)$ and the regional mean functions $\mu_k(t)$ representing the regional deviations from the global mean function are also estimated.

Figure 9 displays the results using China's demeaned data. The estimated global mean function shows a clear "W" shape, similar to that in Figure 5, while the regional mean functions demonstrate significant differences in their scales and patterns despite some fluctuations. For example, the scales of North China Plain, Sichuan Basin, and Guanzhong Plain are all higher than the average, yet they are clustered into separate regions because their temporal patterns show distinct trends, with peaks and troughs at different times points.

The global and regional mean functions of BTH data are presented in Figure 10. There is a clear seasonal pattern in the global mean function: the $PM_{2.5}$ measurements are generally higher in winter than in summer. The overall trend decreases gradually over time suggesting air pollution control in the BTH region has been more effective in 2016 than a few years ago. The stations in the South cluster have the worst pollution in the BTH region and show a persistently positive deviation from the global mean. The regional mean function of the Middle cluster is mostly 0 while that of the North cluster is always negative.

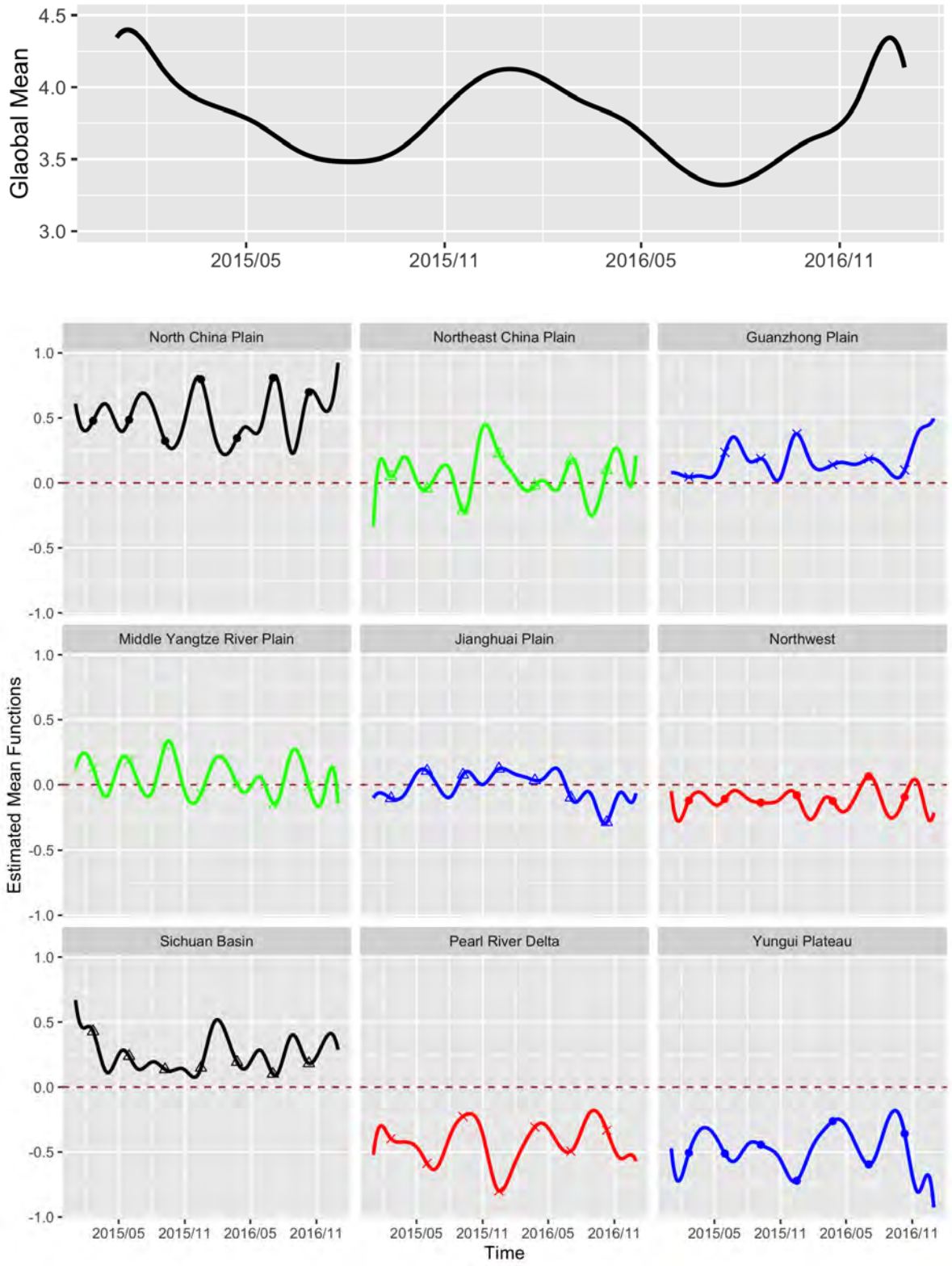

Figure 9: *Top Panels*: Estimated global mean function $\hat{\mu}_0(t)$ using China's $PM_{2.5}$ concentrations of 313 cities from 2015 to 2016. *Bottom Panels*: Estimated regional mean functions of 9 clusters $\hat{\mu}_k(t)$ for $k = 1, 2, \ldots, 9$ after removing the estimated global mean. The dashed lines represent the zero lines.

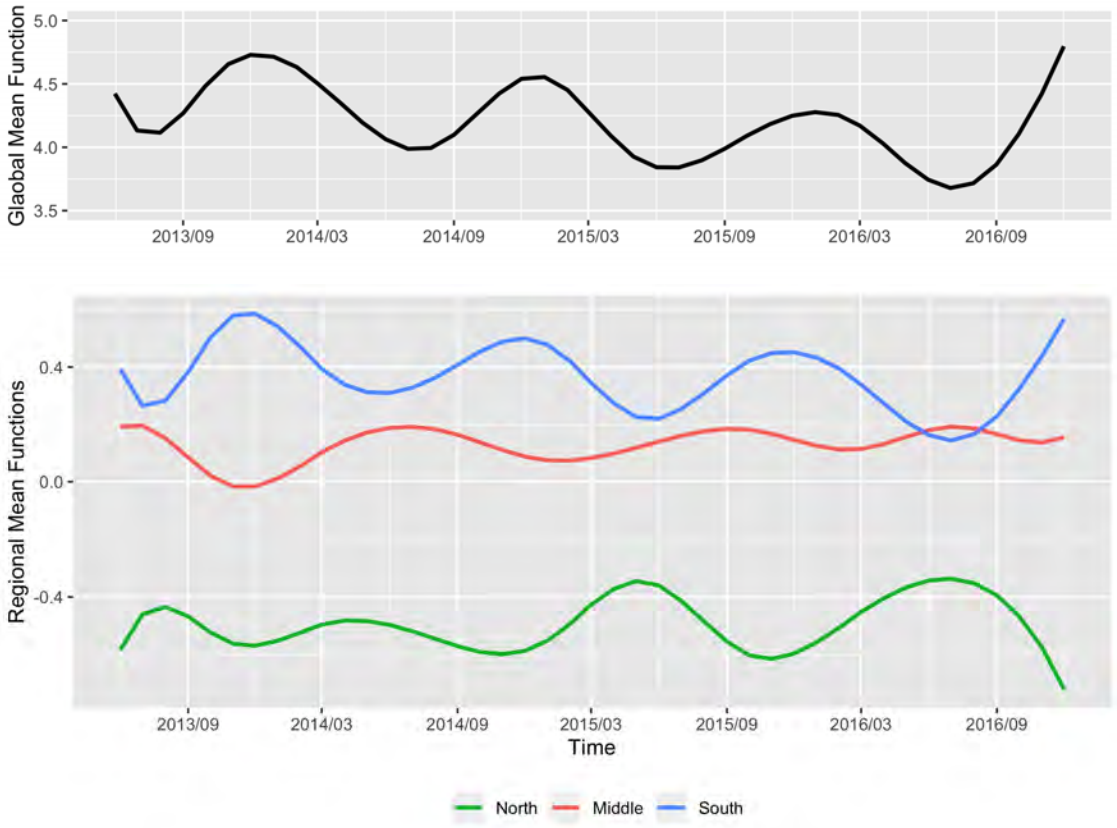

Figure 10: *Top Panel*: Estimated global mean function $\hat{\mu}_0(t)$ using the $PM_{2.5}$ data from the BTH region. *Bottom Panel*: Estimated regional mean functions of three clusters $\hat{\mu}_k(t)$ for $k = 1, 2, 3$ after removing the estimated global mean.

7 Concluding Remarks

In this study, we propose a novel approach to jointly model and cluster spatial dependent functional data with applications to $PM_{2.5}$ concentrations collected from China and the BTH region. Our model allows data from different clusters to have different mean functions and covariance structures, and is able to incorporate spatial dependence through the FPC scores. Markov random fields are assumed for the cluster memberships to define spatial boundaries between regions. Our model respects the spatio-temporal characteristics of the data. It serves as a tool not only for data clustering but also for uncertainty quantification and results interpretation. We use a spline basis system and a data-driven FPC analysis approach to strike a balance between model complexity and flexibility. An efficient MCEM algorithm is used to estimate model parameters, mean functions, and eigenfunctions. The extensive simulation studies show that the proposed method is superior to other methods in

terms of cluster membership prediction and model parameter estimation.

In the analysis of the $PM_{2.5}$ data, our regionalization results not only are in accordance with the findings in the literature (Wang et al., 2015; Chen et al., 2018) but also show much more clear region boundaries that would be helpful for policy making. In addition, the estimated mean functions and FPC functions present distinct and interpretable time-varying patterns, reveal important underlying emission features, and would be useful for pollution prevention and control. As a result, for more efficient control strategies, we recommend identifying and implementing separate interventions (e.g., adopting different control measures and pollution reduction strategies) in the nine regions of China and three regions of the BTH. Although the focus of our data analysis is on air pollution data, our clustering method can be easily expanded to other environmental science or meteorological datasets with a similar structure.

As pointed out by one reviewer, $PM_{2.5}$ is a complex mixture with many constituents. Including $PM_{2.5}$ constituents in the study may provide a better understanding of the local emission patterns than using the particulate matter alone. Currently, the concentrations of $PM_{2.5}$ constituents are not being recorded in the monitoring network. In the future, environmental organizations may consider collecting the constituents and using them as supplementary assessment indicators for more effective air pollution source control. The methodology described in this work can be implemented for the $PM_{2.5}$ constituents as well.

Our approach also opens up some new research questions. For instance, one important question is how to model and cluster multivariate pollutants from multiple sources simultaneously. For the application studies in Section 6, apart from the $PM_{2.5}$ measurements, we also have the concentrations of other air pollutants (e.g., PM_{10}, O_3, and SO_2) at the monitoring stations. Though the particulate matters ($PM_{2.5}$ and PM_{10}) are the most important air pollutants in terms of the proportion of variation explained, it is of scientific importance to have a systematic approach to combine multiple measurements in a framework of joint modeling and clustering. Another question of interest is how to assess the uncertainty of cluster assignments. The memberships of cities or stations are determined using the "posterior" mean of a random variable, thus we may not have a great deal of confidence in the assignments of some "transition zones" where $PM_{2.5}$ patterns are highly variable. These questions and extensions call for future research.

Acknowledgment

The research was partially supported by the National Natural Science Foundation of China (Grant No. 11871485) and China's National Key Research Special Program (Grant No. 2016YFC0207702). The authors are also grateful for the detailed and constructive comments from an Associate Editor and three referees.

SUPPLEMENTARY MATERIAL

Technical Details: Implementation issues related to model selection, additional results for the Monte Carlo EM algorithm, model diagnostic results for the data analysis, and clustering results from other methods (PDF file).

Code: R code for simulation studies and data analysis (zip file).

Dataset: City-level daily $PM_{2.5}$ concentrations of China's entirety from January 2015 to December 2016, and station-level monthly $PM_{2.5}$ concentrations from 73 stations in the BTH region from June 2013 to December 2016 (CSV file), the topographic information including the longitude and latitude for corresponding cities and stations (CSV file), and China's elevation with 1km resolution (TIF file).

Appendix A Computational Details for Spatial Random Effects

For the convenience of computation, we first re-group the elements of the spatial random effects according to the principal components, denoted as $\boldsymbol{\gamma}_1$, and then conditional on $\boldsymbol{\mathcal{Z}} = (\mathcal{Z}_1, \ldots, \mathcal{Z}_n)$, $\boldsymbol{\gamma}_1$ is re-grouped again into $\boldsymbol{\gamma}^*_{\boldsymbol{\mathcal{Z}}}$ based on the cluster membership. Let $\mathbf{O}_1$ and $\mathbf{O}_{2,\boldsymbol{\mathcal{Z}}}$ be two $nQ \times nQ$ permutation matrices. We use the following expressions:

$$\begin{aligned}
\boldsymbol{\gamma} &= (\gamma_{11}, \ldots, \gamma_{1Q}, \ldots, \gamma_{n1}, \ldots, \gamma_{nQ})^{\mathrm{T}} = \mathbf{O}_1 \boldsymbol{\gamma}_{\mathbf{1}}, \\
\boldsymbol{\gamma}_{\mathbf{1}} &= (\gamma_{11}, \ldots, \gamma_{n1}, \ldots, \gamma_{1Q}, \ldots, \gamma_{nQ})^{\mathrm{T}} = \mathbf{O}_{2,\boldsymbol{\mathcal{Z}}} \boldsymbol{\gamma}^*_{\boldsymbol{\mathcal{Z}}}, \\
\boldsymbol{\gamma}^*_{\boldsymbol{\mathcal{Z}}} &= (\boldsymbol{\gamma}_{\cdot 1,1}, \ldots, \boldsymbol{\gamma}_{\cdot 1,K}, \ldots, \boldsymbol{\gamma}_{\cdot Q,1}, \ldots, \boldsymbol{\gamma}_{\cdot Q,K})^{\mathrm{T}},
\end{aligned}$$

where for $q = 1, \ldots, Q$ and $k = 1, \ldots, K$, $\boldsymbol{\gamma}_{\cdot q,k}$ collects all γ_{iq}'s that belong to the same cluster; in other words, $\boldsymbol{\gamma}_{\cdot q,k} = (\gamma_{k_1 q}, \ldots, \gamma_{k_{n_k} q})^{\mathrm{T}}$, where n_k is the number of locations belonging to

cluster k, and $k_1 < \cdots < k_{n_k}$ are the corresponding indices, i.e, $\mathcal{Z}_{k_j} = k$, for $j = k_1, \ldots, k_{n_k}$. Use $\boldsymbol{\Gamma}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\Gamma}^*_{\boldsymbol{\gamma}}$ to represent the covariance matrices of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*_{\boldsymbol{\mathcal{Z}}}$, respectively.

Let $\boldsymbol{\Gamma}_{\cdot q,k}$ be the covariance matrix of $\boldsymbol{\gamma}_{\cdot q,k}$, then we have $\boldsymbol{\Gamma}_{\cdot q,k} = \sigma^2_{\gamma,q,k}\boldsymbol{R}_k(\boldsymbol{\phi}_{q,k})$ under the assumptions in Section 3.1, where $\boldsymbol{R}_k(\boldsymbol{\phi}_{q,k})$ is an $n_k \times n_k$ matrix with elements $\boldsymbol{R}_{k,ii'}(\boldsymbol{\phi}_{q,k}) = \rho(\|s_i - s_{i'}\|; \boldsymbol{\phi}_{q,k})$, $i, i' = 1, \ldots, n_k$. For simplicity, we replace $\boldsymbol{\phi}_{q,k}$ by a set of common parameters $\boldsymbol{\phi}$ as follows. Since $\operatorname{cov}(\gamma_{iq,k}, \gamma_{i'q',k'}) = 0$ when $q \neq q'$ or $k \neq k'$, the covariance matrix of $\boldsymbol{\gamma}^*_{\boldsymbol{\mathcal{Z}}}$ is block diagonal $\boldsymbol{\Gamma}^*_{\boldsymbol{\gamma}} = \operatorname{diag}(\boldsymbol{\Gamma}_{\cdot 1}, \ldots, \boldsymbol{\Gamma}_{\cdot Q})$, where $\boldsymbol{\Gamma}_{\cdot q} = \operatorname{diag}(\boldsymbol{\Gamma}_{\cdot q,1}, \ldots, \boldsymbol{\Gamma}_{\cdot q,K})$ for $q = 1, \ldots, Q$. Note that $\boldsymbol{\gamma}$ is just a reordering of $\boldsymbol{\gamma}^*_{\boldsymbol{\mathcal{Z}}}$, i.e., $\boldsymbol{\gamma} = \mathbf{O}_{\boldsymbol{\mathcal{Z}}}\boldsymbol{\gamma}^*_{\boldsymbol{\mathcal{Z}}}$, where $\mathbf{O}_{\boldsymbol{\mathcal{Z}}} = \mathbf{O}_1\,\mathbf{O}_{2,\boldsymbol{\mathcal{Z}}}$. It follows that the covariance matrix of $\boldsymbol{\gamma}$ is $\boldsymbol{\Gamma}_{\boldsymbol{\gamma}} = \mathbf{O}_{\boldsymbol{\mathcal{Z}}}\,\boldsymbol{\Gamma}^*_{\boldsymbol{\gamma}}\,\mathbf{O}_{\boldsymbol{\mathcal{Z}}}{}^{\mathrm{T}}$.

Appendix B Technical Details for the E-Step

In this section, we provide more details on the Gibbs sampling procedure used in the E-step. The posterior distribution of $(\boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{\mathcal{Z}})$ can be derived from the following joint Gaussian distribution:

$$\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{\gamma} \end{pmatrix} \Bigg| \boldsymbol{\mathcal{Z}} \sim \text{Normal}\left(\begin{pmatrix} \widetilde{\boldsymbol{B}}^{\mathrm{T}}\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\mathcal{Z}}} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \widetilde{\boldsymbol{B}}^{\mathrm{T}}\widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{\Theta}}^{\mathrm{T}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{B}} + \sigma^2_\epsilon \boldsymbol{I} & \widetilde{\boldsymbol{B}}^{\mathrm{T}}\widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}} \\ \widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{\Theta}}^{\mathrm{T}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{B}} & \widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}} \end{pmatrix} \right).$$

Therefore, we obtain that $\left(\boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\Omega}_{\text{prev}}\right) \sim \text{Normal}\,(e, v)$, where

$$\begin{aligned} e =& \mathrm{E}(\boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{\mathcal{Z}}) = \widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{\Theta}}^{\mathrm{T}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{B}}\,\operatorname{Var}(\boldsymbol{Y}|\boldsymbol{\mathcal{Z}})^{-1}(\boldsymbol{Y} - \widetilde{\boldsymbol{B}}^{\mathrm{T}}\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\mathcal{Z}}}), \\ v =& \operatorname{Var}(\boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{\mathcal{Z}}) = \widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}} - \widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{\Theta}}^{\mathrm{T}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{B}}\,\operatorname{Var}(\boldsymbol{Y}|\boldsymbol{\mathcal{Z}})^{-1}\widetilde{\boldsymbol{B}}^{\mathrm{T}}\widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}, \end{aligned}$$

and $\left(\mathcal{Z}_i|\boldsymbol{\gamma}_i, \boldsymbol{Y}_i, \boldsymbol{\Omega}_{\text{prev}}\right) \sim \text{Multinomial}\,(p_{i1}, \ldots, p_{iK})$, where

$$p_{ik} = \frac{f(\boldsymbol{Y}_i|\boldsymbol{\gamma}_i, \mathcal{Z}_i = k)\pi_k}{\sum_{k=1}^K f(\boldsymbol{Y}_i|\boldsymbol{\gamma}_i, \mathcal{Z}_i = k)\pi_k}.$$

Assume we have $(\boldsymbol{\mathcal{Z}}^{(\tau-1)}, \boldsymbol{\gamma}^{(\tau-1)})$ at the $(\tau-1)$th step. Using the above marginal results, we first generate $\boldsymbol{\gamma}^{(\tau)}$ from $(\boldsymbol{\gamma}^{(\tau)}|\boldsymbol{Y}, \boldsymbol{\mathcal{Z}}^{(\tau-1)}, \boldsymbol{\Omega}_{\text{prev}})$ and then $\mathcal{Z}_i^{(\tau)}$ from $(\mathcal{Z}_i^{(\tau)}|\boldsymbol{Y}_i, \boldsymbol{\gamma}_i^{(\tau-1)}, \boldsymbol{\Omega}_{\text{prev}})$. At each E-step, we repeat these two steps of Gibbs sampling $\overline{\mathrm{T}}_0+\overline{\mathrm{T}}$ times and omit the first $\overline{\mathrm{T}}_0$ samples. In the simulation studies, we use $\overline{\mathrm{T}}_0 = 50$ and $\overline{\mathrm{T}} = 100$. The *Sherman-Woodbury-Morrison* formula is also applied to invert the high-dimensional conditional variance of $\boldsymbol{Y}$

given $\boldsymbol{\mathcal{Z}}$ appeared in e and v, that

$$\left(\widetilde{\boldsymbol{B}}^{\mathrm{T}}\widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}\widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}}^{\mathrm{T}}\widetilde{\boldsymbol{B}}+\sigma_\epsilon^2\boldsymbol{I}\right)^{-1}=\sigma_\epsilon^{-2}\boldsymbol{I}-\widetilde{\boldsymbol{B}}^{\mathrm{T}}\widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}}\left(\widetilde{\boldsymbol{\Gamma}}_{\boldsymbol{\mathcal{Z}}}^{-1}-\sigma_\epsilon^{-2}\widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}}^{\mathrm{T}}\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^{\mathrm{T}}\widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}}\right)\widetilde{\boldsymbol{\Theta}}_{\boldsymbol{\mathcal{Z}}}^{\mathrm{T}}\widetilde{\boldsymbol{B}}.$$

Appendix C Technical Details for the M-Step

Following is the complete procedure for the parameter updates in the M-step.

1. Estimation of $\boldsymbol{\alpha}_k$ and σ_ϵ^2, for $k=1,\ldots,K$. By maximizing $\hat{Q}_1$ in (10), we are able to update $\boldsymbol{\alpha}_k$ and σ_ϵ^2:

$$\widehat{\boldsymbol{\alpha}}_k=\left\{\sum_{\tau=1}^{\overline{\mathrm{T}}}\left(\sum_{i=1}^{n}Z_{ik}^{(\tau)}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right)\right\}^{-1}\cdot\sum_{\tau=1}^{\overline{\mathrm{T}}}\left\{\sum_{i=1}^{n}Z_{ik}^{(\tau)}\boldsymbol{B}\left(\boldsymbol{Y}_i-\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Theta}_k\boldsymbol{\gamma}_i^{(\tau)}\right)\right\},$$

$$\widehat{\sigma}_\epsilon^2=\frac{1}{\overline{\mathrm{T}}\widetilde{n}}\sum_{\tau=1}^{\overline{\mathrm{T}}}\left\{\sum_{i=1}^{n}\sum_{k=1}^{K}Z_{ik}^{(\tau)}\left\|\boldsymbol{Y}_i-\ \boldsymbol{B}^{\mathrm{T}}\left(\boldsymbol{\alpha}_k+\boldsymbol{\Theta}_k\boldsymbol{\gamma}_i^{(\tau)}\right)\right\|^2\right\},$$

   where we use the notation $Z_{ik}^{(\tau)}=I\left(\mathcal{Z}_i^{(\tau)}=k\right)$ for simplicity.

2. Update each column of $\boldsymbol{\Theta}_k$ sequentially. For $q=1,\ldots,Q$,

$$\widehat{\boldsymbol{\Theta}}_{k,q}=\left\{\sum_{\tau=1}^{\overline{\mathrm{T}}}\left(\sum_{i=1}^{n}Z_{ik}^{(\tau)}\gamma_{iq}^{2^{(\tau)}}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right)\right\}^{-1}\boldsymbol{B}\sum_{\tau=1}^{\overline{\mathrm{T}}}\left\{\sum_{i=1}^{n}Z_{ik}^{(\tau)}\gamma_{iq}^{(\tau)}\left(\boldsymbol{Y}_i-\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\alpha}_k-\sum_{l\neq q}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Theta}_{k,l}\gamma_{il}^{(\tau)}\right)\right\}.$$

   Then, orthogonalize $\widehat{\boldsymbol{\Theta}}_k$ using a QR decomposition.

3. Update $\sigma_{\gamma,q,k}^2$ by maximizing $\widehat{Q}_2$. To simplify the computation of partial derivatives, we use the expressions of $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}_{\boldsymbol{\mathcal{Z}}}^*$, and the block diagonal matrix $\boldsymbol{\Gamma}_{\boldsymbol{\gamma}}^*$ in Appendix A. The updating formula is:

$$\widehat{\sigma}_{\gamma,q,k}^2=\frac{1}{\overline{\mathrm{T}}}\sum_{\tau=1}^{\overline{\mathrm{T}}}\left(\frac{1}{n_k}\boldsymbol{\gamma}_{\cdot q,k}^{(\tau)^{\mathrm{T}}}\boldsymbol{R}_k^{-1}(\boldsymbol{\phi})\boldsymbol{\gamma}_{\cdot q,k}^{(\tau)}\right),\quad\text{for } q=1,\ldots,Q.$$

4. Estimation of $\boldsymbol{\phi}$. Denote the components of $\boldsymbol{\phi}$ as ϕ_r's. Given the current estimates of other parameters, we minimize $\widehat{Q}_2$ and obtain the gradient with respect to $\boldsymbol{\phi}$, which is

a vector with elements

$$\frac{1}{\overline{\overline{\mathrm{T}}}}\sum_{\tau=1}^{\overline{\mathrm{T}}}\left\{\sum_{q=1}^{Q}\sum_{k=1}^{K}\mathrm{tr}\left(\boldsymbol{R}_k^{-1}\frac{\partial \boldsymbol{R}_k}{\partial \phi_r}\right)-\frac{1}{\sigma^2_{\gamma,q,k}}\mathrm{tr}\left(\boldsymbol{R}_k^{-1}\frac{\partial \boldsymbol{R}_k}{\partial \phi_r}\boldsymbol{R}_k^{-1}\boldsymbol{\gamma}_{\cdot q,k}^{(\tau)}\boldsymbol{\gamma}_{\cdot q,k}^{(\tau)^{\mathrm{T}}}\right)\right\}.$$

Due to the lack of analytic solutions, we use the Newton-Raphson method to find the solution as the updated estimate $\widehat{\boldsymbol{\phi}}$.

5. Update ν by maximizing $\widehat{Q}_3$. The gradient with respect to ν is

$$\sum_{i=1}^{n}\sum_{k=1}^{K}Z_{ik}^{(\tau)}\left[\sum_{i'\in\partial i}Z_{i'k}^{(\tau)}-\frac{\sum_{k=1}^{K}\left\{\sum_{i'\in\partial i}Z_{i'k}^{(\tau)}\right\}\exp\left\{\nu\sum_{i'\in\partial i}Z_{i'k}^{(\tau)}\right\}}{\sum_{k=1}^{K}\exp\left\{\nu\sum_{i'\in\partial i}Z_{i'k}^{(\tau)}\right\}}\right].$$

The Newton-Raphson method is also used to obtain $\widehat{\nu}$.

In the updating formulas given above, we fix the parameters on the right-hand side of each equation at their current estimates obtained from the last EM iteration.

References

Anderes, E. B. and Stein, M. L. (2008). Estimating deformations of isotropic Gaussian random fields on the plane. *The Annals of Statistics*, 36(2):719–741.

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195.

Bouveyron, C. and Jacques, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, 5(4):281–300.

Bryan, B. and Adams, J. (2002). Three-dimensional neurointerpolation of annual mean precipitation and temperature surfaces for China. *Geographical Analysis*, 34(2):93–111.

Chan, K. S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association*, 90(429):242–252.

Chen, L., Guo, B., Huang, J., He, J., Wang, H., Zhang, S., and Chen, S. X. (2018). Assessing air-quality in Beijing-Tianjin-Hebei region: The method and mixed tales of $PM_{2.5}$ and O_3. *Atmospheric Environment*, 193:290–301.

China's State Council (2013). The action plan for air pollution prevention and control. http://www.gov.cn/zwgk/2013-09/12/content_2486773.htm. In Chinese.

Chiou, J.-M. and Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):679–699.

Clifford, P. (1990). Markov random fields in statistics. In Grimmett, G. and Welsh, D. J., editors, *Disorder in Physical Systems*, Clarendon. Oxford.

Cohen, A. J., Brauer, M., Burnett, R., et al. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, 389(10082):1907–1918.

de Boor, C. (2001). *A practical guide to splines*. Springer-Verlag, New York.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.

Gao, H., Chen, J., Wang, B., Tan, S.-C., Lee, C. M., Yao, X., Yan, H., and Shi, J. (2011). A study of air pollution of city clusters. *Atmospheric Environment*, 45(18):3069–3077.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

Giacofci, M., Lambert-Lacroix, S., Marot, G., and Picard, F. (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics*, 69(1):31–40.

Giraldo, R., Delicado, P., and Mateu, J. (2012). Hierarchical clustering of spatially correlated functional data: Clustering of spatial functional data. *Statistica Neerlandica*, 66(4):403–421.

Hoek, G., Krishnan, R. M., Beelen, R., et al. (2013). Long-term air pollution exposure and cardio- respiratory mortality: A review. *Environmental Health*, 12(1):43.

Huang, R.-J., Zhang, Y., Bozzetti, C., et al. (2014). High secondary aerosol contribution to particulate pollution during haze events in China. *Nature*, 514(7521):218–222.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Jacques, J. and Preda, C. (2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing*, 112:164–171.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408.

Jiang, H. and Serban, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics*, 54(2):108–119.

Kindermann, R. and Snell, J. L. (1980). *Markov random fields and their applications*. Contemporary Mathematics. American Mathematical Society, Providence, RI.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., and Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525:367–371.

Li, K. (2015). Report on the Work of the Government (2015). `http://english.www.gov.cn/archive/publications/2015/03/05/content_281475066179954.htm`. Delivered at Third Session of the 12th National People's Congress on March 5, 2015.

Li, S.-T., Chou, S.-W., and Pan, J.-J. (2000). Multi-resolution spatio-temporal data mining for the study of air pollutant regionalization. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, pages 1–7.

Li, X., Zhou, W., and Chen, Y. D. (2015). Assessment of regional drought trend and risk over China: A drought climate division perspective. *Journal of Climate*, 28(18):7025–7037.

Li, Y., Wang, N., and Carroll, R. J. (2013). Selecting the number of principal components in functional data. *Journal of the American Statistical Association*, 108(504):1284–1294.

Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S. X. (2016). $PM_{2.5}$ data reliability, consistency, and air quality assessment in five Chinese cities. *Journal of Geophysical Research: Atmospheres*, 121(17):10220–10236.

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., and Chen, S. X. (2015). Assessing Beijing's $PM_{2.5}$ pollution: Severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A*, 471(2182):20150257.

Lin, W., Xu, X., Zhang, X., and Tang, J. (2008). Contributions of pollutants from North China Plain to surface ozone at the shangdianzi gaw station. *Atmospheric Chemistry and Physics*, 8(19):5889–5898.

Matérn, B. (1960). *Spatial Variation*. Springer-Verlag, Berlin.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(May):1145–1164.

Peng, J. and Müller, H.-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *The Annals of Applied Statistics*, 2(3):1056–1077.

Pope, C. A. I., Burnett, R. T., Thun, M. J., and et al (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Journal of American Medical Association*, 287(9):1132–1141.

Porcu, E., Bevilacqua, M., and Genton, M. G. (2016). Spatio-temporal covariance and cross-covariance functions of the great circle distance on a sphere. *Journal of the American Statistical Association*, 111(514):888–898.

Qian, W., Tang, X., and Quan, L. (2004). Regional characteristics of dust storms in China. *Atmospheric Environment*, 38(29):4895–4907.

Ramsay, J. and Bernard, S. (2005). *Functional data analysis*. Springer series in statistics. Springer.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Romano, E., Balzanella, A., and Verde, R. (2013). *A Regionalization Method for Spatial Functional Data Based on Variogram Models: An Application on Environmental Data*, pages 99–108. Springer, Berlin, Heidelberg.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119.

van Donkelaar, A., Martin, R. V., Brauer, M., et al. (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. *Environmental Health Perspectives*, 118(6):847–855.

Wang, S., Li, G., Gong, Z., et al. (2015). Spatial distribution, seasonal variation and regionalization of PM2.5 concentrations in China. *Science China Chemistry*, 58(9):1435–1443.

Wang, Y., Hao, J., McElroy, M. B., et al. (2009). Ozone air quality during the 2008 Beijing Olympics: Effectiveness of emission restrictions. *Atmospheric Chemistry and Physics*, 9(14):5237–5251.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Xu, W. Y., Zhao, C. S., Ran, L., et al. (2011). Characteristics of pollutants and their correlation to meteorological conditions at a suburban site in the North China Plain. *Atmospheric Chemistry and Physics*, 11(9):4353–4369.

Zhang, H., Zhu, Z., and Yin, S. (2016). Identifying precipitation regimes in China using model-based clustering of spatial functional data. In *Proceedings of the Sixth International Workshop on Climate Informatics*, pages 117–120.

Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., and Chen, S. X. (2017). Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A*, 473(2205):20170457.

Zhang, X. Y., Wang, Y. Q., Niu, T., et al. (2012). Atmospheric aerosol compositions in China: spatial/temporal variability, chemical signature, regional haze distribution and comparisons with global aerosols. *Atmospheric Chemistry and Physics*, 12(2):779–799.

Zhou, L., Huang, J. Z., and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, 95(3):601–619.

Zhou, L., Huang, J. Z., Martinez, J. G., Maity, A., Baladandayuthapani, V., and Carroll, R. J. (2010). Reduced rank mixed effects models for spatially correlated hierarchical functional data. *Journal of the American Statistical Association*, 105(489):390–400.