Partitioning Structure Learning for Segmented Linear Regression Trees

Xiangyu Zheng
Peking University
zhengxiangyu@pku.edu.cn

Song Xi Chen
Peking University
csx@gsm.pku.edu.cn

Abstract

This paper proposes a partitioning structure learning method for segmented linear regression trees (SLRT), which assigns linear predictors over the terminal nodes. The recursive partitioning process is driven by an adaptive split selection algorithm that maximizes, at each node, a criterion function based on a conditional Kendall's τ statistic that measures the rank dependence between the regressors and the fitted linear residuals. Theoretical analysis shows that the split selection algorithm permits consistent identification and estimation of the unknown segments. A sufficiently large tree is induced by applying the split selection algorithm recursively. Then the minimal cost-complexity tree pruning procedure is applied to attain the right-sized tree, that ensures (i) the nested structure of pruned subtrees and (ii) consistent estimation to the number of segments. Implanting the SLRT as a built-in base predictor, we obtain the ensemble predictors by random forests (RF) and the proposed weighted random forests (WRF). The practical performance of the SLRT and its ensemble versions are evaluated via numerical simulations and empirical studies. The latter shows their advantageous predictive performance over a set of state-of-the-art tree-based models on well-studied public datasets.

1 Introduction

Data partitioning is a fundamental pre-processing method that explores the partitioning structure of the feature space such that the subspaces are more compliant to a simple model [1]. We consider the segmented linear regression (SLR) models, which prescribes linear predictors over the partitions. Partitioning structure learning is the core of SLR, that selects the split variables and levels as well as determines the number of segments.

SLR has been studied in statistics and econometrics [2, 3, 4, 5], but the existing methods tend to assume the split variable is known and univariate, with segments estimated by a costly simultaneous optimization. We propose a tree-based approach for SLR called segmented linear regression trees (SLRT), that does not require the pre-specified information about the split variables. SLRT is completely data-driven and facilitates more efficient computation via recursive partitioning, which is fundamentally based on a split selection algorithm and a tree pruning algorithm.

Split Selection Algorithm At each internal node of the tree, the optimal split variable and level pair is selected to partition the feature space into two halves. Let $\hat{e}$ be the fitted residuals by the ordinary least square regression. Any non-linearity in the underlying regression function is reflected in the dependence between $\hat{e}$ and the regressors. Based on the conditional Kendall's τ rank correlation [6], we propose the following criterion function at a candidate split variable index j and a split level a, $\mathcal{C}(j,a) = \sum_{k=1}^{p} \{|\hat{\tau}(X_k, \hat{e}|X_j \le a)| + |\hat{\tau}(X_k, \hat{e}|X_j > a)|\}$, where $\hat{\tau}$ is the sample version of the Kendall's τ, X is a p-dimensional regressors vector with X_k being its k-th component. The optimal split is selected by maximizing $\mathcal{C}(j,a)$ over the candidate split variables $\{X_j\}$ and levels $\{a\}$ in the

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

observed sample of X_j. Theoretical analysis shows that it leads to the consistent identification and estimation of the most prevailing split variable and level that attains the maximum of $\mathcal{C}(j, a)$.

Tree Pruning Algorithm We define an adaptive cost-complexity measure that combines the accuracy of the linear regression fit at each node with a penalty for a large tree size. The optimally pruned tree is selected from a nested sequence of pruned subtrees by minimizing the cost-complexity measure. Theoretical analysis shows that the pruning method leads to consistent estimation of the underlying number of segments, which promotes a parsimonious partitioning structure.

Leaf Modeling and Ensemble Methods For predictors within segments, we employ the LASSO procedure [7] to select the most influential variables and estimate the linear parameters. Furthermore, by implanting SLRT as the base predictor in the random forests (RF) formulation, we obtain the ensemble predictor that improves the model stability and predictive accuracy. A weighted version of the RF (WRF) is also proposed, which shows an improved performance over the RF by reducing the importance of those under-performing trees in weighting.

As a novel tree-based learning method for segmented linear models, SLRT possesses attractive theoretical properties of the consistent identification and estimation of the partitioning structures, which are confirmed favorably in numerical simulations. Applied on nine benchmark datasets, SLRT had advantageous predictive performance over several state-of-the-art tree-based methods, with further improvement offered by the RF and WRF with SLRT as the base predictor.

The source code of the algorithm is available in the supplementary material.

1.1 Related Work

The proposed segmented linear regression tree is a tree-based approach to segmented linear regression (SLR) models, where the partitions of the feature space is axis-aligned. The existing methods of SLR tend to assume a known split variable, such that the partitioning structure learning is reduced to the change-points detection with respect to a given variable. For instance, [2, 3] considered the case where both the univariate partitioning variable and the number of segments are pre-specified. [4, 5] estimated the number of change-points by minimizing the Bayesian information criteria (BIC). [8] selected the change-points via the sum of squared residuals in conjunction with the permutation test, which also assumed a known split variable. Our approach does not require pre-specified information of the segments, and learns the partitioning structure via a tree induction process.

SLR also belongs to the class of region-specific linear models. [1] proposed a partition-wise linear model, where axis-aligned partitions are pre-specified and an 0-1 activeness function was assigned to each region. With each region-specific linear model being estimated first, the activeness functions are optimized through a global convex loss function. [9] proposed a local supervised learning through space partitioning for classification which allows arbitrary partitions and considered linear classifiers, while [10] employed a Bayesian updating process to partition the feature space to rectangular bounding boxes and assigned a constant estimation over each partition like CART.

Our approach is closely related to the regression tree (regression part of CART, [11]), a well-known region-specific approach that used a constant-valued predictor within each terminal node. There have been tree-based algorithms which assigns linear predictors in terminal nodes, which tend to be heuristic without theoretical analysis. One group of the methods [12, 13, 14] adopted splitting algorithms similar to that of CART, which tend to ignore the correspondence between the evaluation criteria for splits and the models in terminal nodes. Another group [15, 16, 17, 18, 19] employed heuristic criteria designed to make the subsets more compliant for linear models in one step, without considering the properties of the estimated boundaries. Our split selection algorithm is closely related to GUIDE [17, 18] as both utilize the estimated residuals $\hat{e}$ at a node level. However, GUIDE used the signs of $\hat{e}$ that would be less informative than using $\hat{e}$ via the Kenall's τ. Another difference is that GUIDE considered the marginal association between signs of $\hat{e}$ and the regressors instead of conditioning on a split variable and level, which can lead to the mis-identification of the split variable.

2 Segmented Linear Regression Models

This section presents the framework of SLRT, and provides the motivation and the theoretical properties to the computational algorithms in Section 3.

2.1 Framework

Consider the relationship between a univariate response Y and a multivariate explanatory covariate $X = (X_1, \cdots, X_p)^T$. Assume that the mean regression function $m(X) = E(Y|X)$ is partition-wise linear over L_0 unknown partitions $\{D_l\}_{l=1}^{L_0}$ in the domain of X so that

$$Y = \sum_{l=1}^{L_0} (\alpha_l + X'\beta_l)\, \mathbb{1}(X \in D_l) + \varepsilon, \tag{1}$$

where (α_l, β_l) are regression coefficients over domain D_l and ε is the random error satisfying $E(\varepsilon|X) = 0$. In this paper, we consider the case of axis-aligned partitions $\{D_l\}$, which are determined by a collection of split levels $\{X_{j_q} = a_q\}_{q=1}^{Q_0}$. The model may be extended to more general shape of $\{D_l\}$ by undergoing pre-transformations, which will be a topic for a future study.

The determination of $\{D_l\}$ is equivalent to selecting the split variable and level pairs, namely $\mathcal{S} = \{(j_q, a_q)\}_{q=1}^{Q_0}$, where Q_0 is determined by L_0 and the geometry of $\{D_l\}$. The task of partitioning structure learning is to identify the underlying split variables and estimate the split levels consistently. We adopt the computationally efficient regression tree approach by applying the split variable and level selection algorithm recursively, ending with terminal nodes for the desired partitions.

2.2 Statistical Analysis of Criterion Functions

To select the optimal split variable and level at a node, we fit the least square regression over the node and select the optimal split by studying the rank correlation between the estimated residuals and the regressors given a candidate split variable and a split level. This is computationally more efficient than the commonly used cost minimization procedure [11, 17, 20], which would require repeated least square fitting for each candidate split variable and level.

For the ease of presentation, we consider the one-time split selection over the root node t_0 which contains data D_{t_0} of n independent observations $\{X(i), Y(i)\}_{i=1}^n$ generated from Model (1). We are to partition D_{t_0} into two subsets to make the data on each subset more compliant to a linear model. To attain this, let $\hat{Y} = \hat{\alpha}_{t_0} + X'\hat{\beta}_{t_0}$ be the fitted ordinary least square (OLS) regression over D_{t_0}, and $\hat{e} = Y - \hat{Y}$ be the estimated residuals. If the underlying regression function $m(X)$ is nonlinear, the non-linearity will be reflected in the residuals $\hat{e}$ and their dependence with the potential split variables. Indeed, if $m(X)$ is piecewise linear, the estimated residuals $\hat{e}$ is also piecewise linear in X since $\hat{e} = \sum_{l=1}^{L} \left((\alpha_l - \hat{\alpha}_{t_0}) + X'(\beta_l - \hat{\beta}_{t_0})\right) \mathbb{I}(X \in D_l) + \varepsilon$. A regressor X_k and $\hat{e}$ tend to be accordant (discordant) for positive (negative) coefficient within each partition. To capture the dependence, we employ the Kendall's τ coefficient [6] to define the following criterion function:

$$\mathcal{C}(j, a) = \sum_{k=1}^{p} \left\{ \left| \hat{\tau}\left(X_k, \hat{e} | X_j \le a\right) \right| + \left| \hat{\tau}\left(X_k, \hat{e} | X_j > a\right) \right| \right\}, \tag{2}$$

for $1 \le j \le p$, $a \in \{X_j(i)\}_{i=1}^n$ and

$$\hat{\tau}\left(X_k, \hat{e} \middle| X_j \le a\right) = \frac{\sum_{\substack{i<i' \\ i,i' \in I_{t_L}(j,a)}} \operatorname{sgn}\left((X_k(i) - \hat{e}(i))(X_k(i') - \hat{e}(i'))\right)}{N_{t_L}(j,a)(N_{t_L}(j,a) - 1)/2} \tag{3}$$

is the Kendall's τ statistic, where $I_{t_L}(j, a) = \{i \,|X_j(i) \le a, 1 \le i \le n\}$ is the index set for the left partition split by variable X_j at level a and the sample size of $I_{t_L}(j, a)$ is $N_{t_L}(j, a) = |I_{t_L}(j, a)|$. The $\hat{\tau}(X_k, \hat{e}\big|X_j > a)$, $I_{t_R}(j, a)$ and $N_{t_R}(j, a)$ are defined analogously.

Based on $\mathcal{C}(j, a)$, we propose the split selection Algorithm 1 in Section 3.1, which is essentially motivated by the following Lemma 2.1.

Lemma 2.1 *Suppose the regressors are uncorrelated conditional on each partition of $\{D_l\}_{l=1}^{L_0}$. Assume the following technical conditions:* (1) $\mathrm{E}(\varepsilon|X) = \mathrm{median}(\varepsilon|X) = \mathrm{median}(X_j - \mathrm{E}(X_j)|X_k) = 0$ *for $k \neq j$;* (2) *for $(j_k, a_k) \in \mathcal{S}$, $0 < \mathrm{P}(X_{j_k}^{(s)} \le a_k | X_{j'}^{(s)} \le a') < 1$ when $(j', a') \neq (j_k, a_k)$. Let $\bar{\mathcal{C}}(j, a)$ be the probability limit of $\mathcal{C}(j, a)$. Then, for any $(j', a') \notin \mathcal{S}$, $\bar{\mathcal{C}}(j', a') < \bar{\mathcal{C}}(j_q, a_q)$ for any $(j_q, a_q) \in \mathcal{S}$, with $\mathcal{S}$ being the genuine set of split variable and level pairs.*

The proof of Lemma 2.1 includes two phases. We firstly investigate the simple case where $L_0 = 2$, and then generalize the conclusion to $L_0 \geq 2$ using the law of iterated expectation. Please refer to the supplements for the details and a further discussion about the technical conditions. Intuitively speaking, maximizing $\mathcal{C}(j, a)$ is to maximize the sum of rank correlations between the estimated residuals $\hat{e}$ and each element of the regressors X over each of the selected subsets ($\{X_j \leq a\}$ and $\{X_j > a\}$), such that the rank correlation with X contained in $\hat{e}$ could be further distilled by regressing $\hat{e}$ on the regressors conditional on each subset, which leads to a segmented linear regression.

Define the distance $d((\hat{j}, \hat{a}), \mathcal{S}) = \min_q\{|(\hat{j}, \hat{a}) - (j_q, a_q)|\big|(j_q, a_q) \in \mathcal{S}\}$. Then, Lemma 2.1 leads to the following theorem that validates the consistency property of the selected split.

Theorem 2.1 *Let* $(\hat{j}, \hat{a}) = \operatorname{argmax} \mathcal{C}(j, a)$. *Then,* $\mathrm{P}\left(d((\hat{j}, \hat{a}), \mathcal{S}) > \varepsilon\right) \to 0$ *as* $n \to \infty$ *under the assumptions of Lemma 2.1. Specially, when* $\bar{\mathcal{C}}(j, a)$ *has a unique maximum* (j^*, a^*)*, we have* $(j^*, a^*) \in \mathcal{S}$ *and* $(\hat{j}, \hat{a}) \xrightarrow{\mathrm{P}} (j^*, a^*)$ *as* $n \to \infty$.

When the regressors are not conditional uncorrelated as required by Lemma 2.1, we conduct a linear transformation when calculating the conditional Kendall's τ coefficients with $\hat{e}$. Specifically, let $\mathbf{X} = (X(1), \cdots, X(n))'$ be the data matrix for n observations of X. Given a split variable and level $X_j = a$, define $\mathbf{X}_L = \mathbf{X}\,\mathrm{diag}\{\mathbb{1}(X_j(1) \leq a), \cdots, \mathbb{1}(X_j(n) \leq a)\}$ and $\mathbf{X}_R = \mathbf{X}\,\mathrm{diag}\{\mathbb{1}(X_j(1) > a), \cdots, \mathbb{1}(X_j(n) > a)\}$. Then, there exists a non-singular matrix $P^{(j,a)}$ such that $\mathbf{Z}'\mathbf{Z}$ is diagonal for $\mathbf{Z} = \mathbf{X}P^{(j,a)}$, $\mathbf{X}_L P^{(j,a)}$ and $\mathbf{X}_R P^{(j,a)}$, which is facilitated by the simultaneous diagonalization of positive definite matrices (see supplements for detailed calculation procedures of $P^{(j,a)}$ that are based on the spectral decomposition). Let $Z^{(\bar{j},\bar{a})} = XP^{(\bar{j},\bar{a})}$ be the transformed regressors with $Z_k^{(\bar{j},\bar{a})}$ being the k-th element. Define the modified criterion function with index $(\bar{j}, \bar{a})$,

$$\mathcal{C}_{(\bar{j},\bar{a})}(j, a) = \sum_{k=1}^{p} \left\{ \left| \hat{\tau}\left(Z_k^{(\bar{j},\bar{a})}, \hat{e} | X_j \leq a \right) \right| + \left| \hat{\tau}\left(Z_k^{(\bar{j},\bar{a})}, \hat{e} | X_j > a \right) \right| \right\}, \tag{4}$$

where $Z^{(\bar{j},\bar{a})}$ replaces X in (2) and $\hat{e}$ is still the residuals of OLS regression at the node t_0 without transformation. The following Lemma 2.2 shows that $\mathcal{C}_{(\bar{j},\bar{a})}$ possesses properties similar to that of $\mathcal{C}$ introduced in Lemma 2.1, while Lemma 2.2 does not require X to be conditional uncorrelated. This motivates the Algorithm 2 in Section 3.1 and leads to the convergence result in Theorem 2.2.

Lemma 2.2 *Let* $\bar{\mathcal{C}}_{(\bar{j},\bar{a})}(j, a)$ *be the probability limit of* $\mathcal{C}_{(\bar{j},\bar{a})}(j, a)$*, Then, under the same technical conditions (1) and (2) as required in Lemma 2.1,* $\operatorname{argmax}_{(j,a)} \bar{\mathcal{C}}_{(\bar{j},\bar{a})}(j, a) = (\bar{j}, \bar{a})$ *when* $(\bar{j}, \bar{a}) \in \mathcal{S}$*, with* $\mathcal{S} = \{(j_q, a_q)\}_{q=1}^{Q_0}$ *being the genuine set of split variable and level pairs.*

Theorem 2.2 *Let* $(\hat{j},\ \hat{a})_{(\bar{j},\ \bar{a})} = \operatorname{argmax}_{(j,a)} \mathcal{C}_{(\bar{j},\bar{a})}(j, a)$. *Under the technical conditions in Lemma 2.2, if* $(\bar{j}, \bar{a}) \in \mathcal{S}$*, then* $(\hat{j},\ \hat{a})_{(\bar{j},\ \bar{a})} \xrightarrow{\mathrm{P}} (\bar{j}, \bar{a})$ *as* $n \to \infty$.

Theorem 2.2 implies that the convergence of $\left(\hat{j},\ \hat{a}\right)_{(\bar{j},\bar{a})}$ to $(\bar{j}, \bar{a})$ is a necessary condition for $(\bar{j}, \bar{a}) \in \mathcal{S}$. This motivates the distance minimization procedure in Line 7 of Algorithm 2 .

3 Partitioning Structure Learning

We first use the recursive partitioning procedures to generate the initial partitions. Then, we employ the cost-complexity tree pruning procedure to obtain a parsimonious partitions structure.

3.1 Initial Partitions by Recursive Partitioning

The recursive partitioning needs a split selection algorithm at the node level, and a stopping rule for the termination of the partitioning process, the latter is based on two tuning parameters: $N_{\min}$ that controls the sample size in any leaf node and $\mathrm{Dep}_{\max}$ that limits the depth of the tree.

The split selection is the core of the recursive partitioning. In the following Algorithm 1 of recursive partitioning, the split is selected by maximizing $\mathcal{C}(j, a)$. This is directly motivated by Lemma 2.1,

that shows the maximum of $\bar{\mathcal{C}}(j,a)$ is within the genuine set of split variable and level pairs $\mathcal{S}$. Besides, according to Theorem 2.1, the selected split level $X_{\hat{j}} = \hat{a}$ determined by the maximum of the criterion function $\mathcal{C}(j,a)$ is consistent to one of the underlying genuine partitioning boundary provided the regressors are uncorrelated conditional on each partition.

Algorithm 1 Recursive Partitioning for Conditional Uncorrelated Regressors

Input: Training data $D_{t_0} = \{(X_i, Y_i)\}_{i=1}^n$.
Output: Data partitions $\mathcal{D} = \{D_i\}_{i=1}^L$.
1: Initialize: No pre-specified partitions, $\mathcal{D} = \varnothing$; the depth of the root node $\text{Dep}(t_0) = 0$.
2: **if** $|D_{t_0}| < 2N_{\min}$ or $\text{Dep}(t_0) > \text{Dep}_{\max}$ **then**
3: **return** $\mathcal{D} = \mathcal{D} \cup \{D_{t_0}\}$
4: **else**
5: Fit a least square linear regression of Y on X over D_{t_0} and get the estimated residuals $\hat{e}$.
6: Calculate the criterion function $\mathcal{C}(j,a)$ for (j,a) in the set of candidate split pairs $C_{t_0} = \{(j,a)|j = 1, \cdots, p, a \in \{X_j(i)|(X(i),Y(i)) \in D_{t_0}\}, N_{\min} \le |\{i|X_j(i) > a\}| < |D_{t_0}| - N_{\min}\}$.
7: $(\hat{j}, \hat{a}) = \text{argmax}_{(j,a)} \mathcal{C}(j,a)$
8: Let t_L and t_R be the left and right child-nodes of t_0, $\text{Dep}(t_L) = \text{Dep}(t_R) = \text{Dep}(t_0) + 1$, with $D_{t_L} = \left\{(X(i),Y(i)) \,\middle|\, X_{\hat{j}}(i) \le \hat{a}\right\} \cap D_{t_0}$ and $D_{t_R} = \left\{(X(i),Y(i)) \,\middle|\, X_{\hat{j}}(i) > \hat{a}\right\} \cap D_{t_0}$.
9: $t_0 \Leftarrow t_L$ and execute step $2-11$.
10: $t_0 \Leftarrow t_R$ and execute step $2-11$.
11: **end if**

When taking the correlation between regressors into consideration, we apply Algorithm 2 to select the optimal split over the original untransformed variables, which retains easy interpretability of the partitions but requires higher computation cost as is analyzed in the following. Since Algorithm 1 has outlined the recursive partitioning process, Algorithm 2 will concentrate on the split selection at the node level, which corresponds to Line $5-7$ of Algorithm 1.

Enlightened by Theorem 2.2, we select the optimal split by minimizing the distance between $(\hat{j}, \hat{a})_{(\bar{j},\bar{a})}$ and $(\bar{j}, \bar{a})$ in Algorithm 2, where the standardized distance $d(\hat{a}_{(\bar{j},\bar{a})}, \bar{a}) = |\hat{a}_{(\bar{j},\bar{a})} - \bar{a}|/\hat{\sigma}(X_{\bar{j}})$ is used for $\bar{j} = \hat{j}_{(\bar{j},\bar{a})}$, with $\hat{\sigma}(X_{\bar{j}})$ being the sample standard deviation of $X_{\bar{j}}$.

Algorithm 2 Split Selection for Correlated Regressors

Input: Training data $D_{t_0} = \{(X_i, Y_i)\}_{i=1}^n$, $|D_{t_0}| > 2N_{\min}$.
Output: The optimal split variable and level pair $(\hat{j}, \hat{a})$; or no splits and t_0 is a terminal node.
1: Fit a least square linear regression of Y on X over D_{t_0} and get the estimated residuals $\hat{e}$.
2: **for** each $(\bar{j}, \bar{a}) \in C_{t_0}$ **do**
3: calculate the criterion function $\mathcal{C}_{(\bar{j},\bar{a})}(j,a)$ for each $(j,a) \in C_{t_0}$;
4: $(\hat{j}, \hat{a})_{(\bar{j},\bar{a})} = \text{argmax}_{(j,a)} \mathcal{C}_{\bar{j},\bar{a}}(j,a)$, the 'local' optimal split under $(\bar{j}, \bar{a})$,
5: **end for**
6: **if** $\{(\hat{j}, \hat{a})_{(\bar{j},\bar{a})} | \bar{j} = \hat{j}_{(\bar{j},\bar{a})}\} \neq \varnothing$ **then**
7: **return** the optimal split $(\hat{j}, \hat{a}) = \text{argmin}_{(\bar{j},\bar{a})}\{d(\hat{a}_{(\bar{j},\bar{a})}, \bar{a}) \big| (\bar{j}, \bar{a}) \in C_{t_0},\ \bar{j} = \hat{j}_{(\bar{j},\bar{a})}\}$.
8: **else**
9: **return** no suitable splits, t_0 is a terminal node.
10: **end if**

As for the computation complexity of Algorithm 2, suppose there are M_t candidate splits in a node t, then it involves M_t^2 times of calculations of the criterion functions. Since the calculation of Kendall's τ in $\mathcal{C}_{(\bar{j},\bar{a})}(\cdot,\cdot)$ is of complexity $O(N_t \log(N_t))$, with N_t being the sample size of node t. Hence the complexity of Algorithm 2 is $M_t^2 O(N_t \log(N_t))$, which is costly compared to Algorithm 1 that is only $M_t O(N_t \log(N_t))$. Therefore, we may adopt a stopping strategy that terminates the split process when the $\min d(\hat{a}_{(\bar{j},\bar{a})}, \bar{a})$ in Line 7 in Algorithm 2 is larger than a given threshold.

Applying either Algorithm 1 or Algorithm 2 recursively for split selections leads to an initial tree $T_{\max}$, which determines the initial partitions. We outline the pruning of $T_{\max}$ in the following.

3.2 Minimal Cost-complexity Tree Pruning

We adopt the minimal cost-complexity pruning procedure in CART [11], but with a newly defined cost-complexity measure $I_\alpha(T)$ for the regression tree T with linear regression models on leaves.

Define the accuracy measure at a node t in a tree T as $I(t) = \sum_{(X(i),Y(i))\in t}(Y(i) - \hat{m}_T(X_i))^2$, where $\hat{m}_T(\cdot)$ is the segmented liner regression function determined by T. The accuracy of T is $I(T) = \sum_{t\in\widetilde{T}} I(t)$, where $\widetilde{T}$ denotes the set of leaf nodes in T and n is the sample size of the training data. The model complexity of T is measured by the number of leaf nodes $|\widetilde{T}|$.

Taking both the accuracy measure and model complexity into consideration, the cost-complexity measure $I_\alpha(T) = I(T)/n + \alpha|\widetilde{T}|$, where α is a positive penalizing parameter. The optimally pruned tree $T(\alpha)$ is defined as the smallest subtree of $T_{\max}$ that minimizes $I_\alpha(T)$, same as the Definition 3.6 in [11]. Proposition 3.1 verifies the existence and the uniqueness of $T(\alpha)$ and the nested structure of $\{T(\alpha), \alpha > 0\}$ as α varies, which is essential for an efficient programming. The proof is by induction where the key is an inequality satisfied by $I_\alpha(T)$. Please refer to the supplementary for details.

Proposition 3.1 *Let $T_{\max}$ be the initial tree, then*

(i) given an α, there exists one optimally pruned subtree $T(\alpha)$ of $T_{\max}$;

(ii) if $\alpha_2 > \alpha_1$, then $T(\alpha_2)$ is a subtree of or equal to $T(\alpha_1)$.

To obtain the optimally pruned tree, the optimal complexity parameter α^* should be selected. Although α runs through a continuum of values, there are finite number of subtrees $T(\alpha)$, say K subtrees of $T_{\max}$. Then by Proposition 3.1, there exists an increasing sequence of $\{\alpha_k | k = 1, \cdots, K\}$ such that $T(\alpha_{k+1}) \subset T(\alpha_k)$, and for $\alpha \in [\alpha_k, \alpha_{k+1})$, $T(\alpha) = T(\alpha_k)$. In fact, $\{\alpha_k\}_{k=1}^K$ can be exactly calculated from $T_{\max}$. Specifically, let T_t be the subbranch of a tree T with node t being its root, then $\alpha_k = \min_t\{\frac{I(t)-I(T_t)}{|\widetilde{T}_t|-1}\big| t \in T \text{ and } t \notin \widetilde{T}\}$ for $T = T_{\max}$ when $k = 1$ and $T = T(\alpha_{k-1})$ when $k > 1$.

Therefore, the optimization of α^* is reduced to selecting an optimal k^* from $\{1, \cdots, K\}$. let $\bar{\alpha}_k = \sqrt{\alpha_k \alpha_{k+1}}$ for $1 \le k \le (K-1)$ and $\bar{\alpha}_K = \alpha_K$. The optimal complexity parameter α^* is selected from $\{\bar{\alpha}_k\}_{k=1}^K$ by the ten-fold cross-validation to optimize the average predictive accuracy measured by the sum of squared residuals. Then, the optimally pruned subtree is $T(\alpha^*)$. Let $\widehat{L}$ be the number of terminal nodes in $T(\alpha^*)$, under certain general conditions for the distribution of ε and given appropriate α^*, it can be proved that $\widehat{L}$ converges to the genuine number of segments L_0 in probability.

4 Leaf Modeling and Ensemble Methods

4.1 LASSO Linear Regression on Leaf Nodes

Let $\{\hat{D}_l\}_{l=1}^{\widehat{L}}$ be the partitions structure determined by $T(\alpha^*)$. Confined on each partition, the regression coefficients (α_l, β_l) can be estimated by the ordinary least square. However, as X is the overall regressors, not each of them necessarily owns a non-zero coefficient over $\hat{D}_l$, and the significant variables set within each $\hat{D}_l$ may vary. Thus, we consider the variables selection within each leaf node. Besides, as the partitioning process decreases the sample size for estimation in each node, we would like to determine a smaller set of variables that exhibit the strongest effects on each $\hat{D}_l$.

To this purpose, the LASSO method [7] is employed for the variables selection, where the regression coefficients (α_l, β_l) is estimated by

$$\hat{\alpha}_l^{\text{lasso}} = \bar{Y}_{\hat{D}_l};\ \hat{\beta}_l^{\text{lasso}} = \underset{\beta_l}{\operatorname{argmin}} \sum_{\{i|X^{(s)}(i)\in\hat{D}_l\}} \{(Y(i) - \bar{Y} - X^{(r)}(i)'\beta_l)^2 + \lambda_l \sum_{j=1}^{p} |\beta_l(j)|\},$$

where $\bar{Y}_{\hat{D}_l}$ is the sample average over $\hat{D}_l$ and λ_l is the shrinkage parameter selected by cross-validation.

Then, the final prediction function is $\hat{m}^{\text{lasso}}_{T(\alpha^*)}(X) = \sum_{l=1}^{\widehat{L}} \left(\hat{\alpha}_l^{\text{lasso}} + X^{(r)\prime}\hat{\beta}_l^{\text{lasso}}\right) \mathbb{1}(X^{(s)} \in \hat{D}_l)$.

4.2 Weighted Random Forests

We can implant the Kendall's τ based partitioning learning algorithm in random forests (RF, [21]) to create the ensemble predictor. Here we propose the weighted random forests (WRF), that considers the accuracy of each tree and puts the final predictor as a weighted average to improve the predictions.

Suppose $\{T_b\}_{b=1}^B$ are the B regression trees induced from the bootstrap training sets. The RF takes the simple average over the predictions of all regression trees, that is, $\hat{m}_{\text{RF}}(X) = \frac{1}{B}\sum_{b=1}^B \hat{m}_{T_b}(X)$. Note given the training set and X, $\{\hat{m}_{T_1}(X), \cdots, \hat{m}_{T_B}(X)\}$ are independent random variables. The variance of $\hat{m}_{T_b}(X)$ reflects the predictive accuracy of regression tree T_b, that can be estimated by $I(t_{T_b}(X))$ for $t_{T_b}(X)$ is the leaf node in T_b containing X. According to Proposition 4.1, taking the variances of single predictors into account will improve the accuracy of the ensemble predictor.

Proposition 4.1 *Let $\{Z_b\}_{b=1}^B$ be independently distributed random variables from a population $P \in \mathcal{P}$ having a common mean μ and $\text{Var}(Z_b) = \sigma_b^2$. Let $\mathcal{A}$ be the class of all unbiased linear estimations for μ. Then, the optimal estimation that minimizes $\text{E}(A-\mu)^2 (A \in \mathcal{A})$ is $\sum_{b=1}^B \frac{1/\sigma_b^2}{\sum_{j=1}^B 1/\sigma_j^2} Z_b$.*

Motivated by Proposition 4.1, we propose the weighted random forests (WRF), which shows improved predictive accuracy over the RF on the benchmark datasets in Section 5.2.

$$\hat{m}_{\text{WRF}}(X) = \sum_{b=1}^B \frac{1/I(t_{T_b}(X))}{\sum_{j=1}^B 1/I(t_{T_j}(X))} \hat{m}_{T_b}(X). \tag{5}$$

5 Experimental Results

5.1 Simulation Study

In this part, we would illustrate SLRT with two examples. One is segmented linear regression function to investigate the performance of partitions structure learning, the other is a general continuous regression function considered in [22], to illustrate how our method works under a general setting.

First, we consider a regression function $m(X)$ that is segmented linear with 12 segments determined by 4 binary splits at $X_1 = 10, X_2 = 10, X_2 = 15, X_4 \in \{a, b\}$ or $\{c\}$:

$$\begin{aligned} m(X) = {} & 3X_1\mathbb{I}\{X_2 > 15\} - 3X_1\mathbb{I}\{X_2 \le 15\} - 3X_2\mathbb{I}\{X_2 > 10\} - 5X_2\mathbb{I}\{X_2 \le 10\} \\ & + X_3\mathbb{I}\{X_1 > 10\} - X_3\mathbb{I}\{X_1 \le 10\} + X_3\mathbb{I}\{X_4 \in \{\text{'a', 'b'}\}\} - 3X_3\mathbb{I}\{X_4 \in \{\text{'c'}\}\}. \end{aligned} \tag{6}$$

Training data of size 1500 was generated from $Y = m(X) + \varepsilon$ with $\varepsilon \sim N(0,1)$ and independent regressors $X_1 \sim U(0,20)$, $X_2 \sim U(0,25)$, $X_3 \sim U(0,10)$ and X_4 took values in {'a', 'b', 'c'} with equal probabilities (see Supplementary for the case of dependent regressors). Figure 1 shows that the estimated partitions in the terminal nodes of $T(\alpha^*)$ are quite close to the space partitions in $m(X)$.

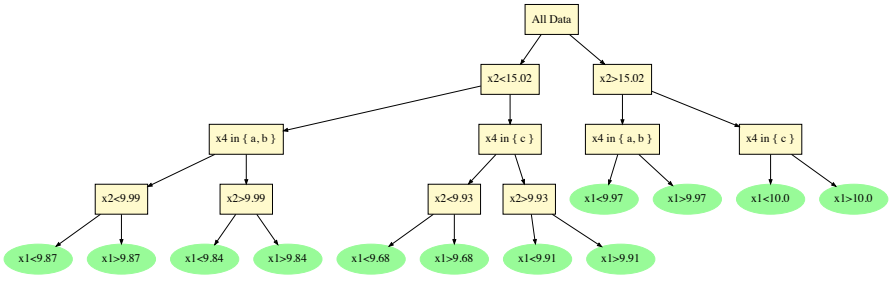


Figure 1: The optimally pruned tree $T(\alpha^*)$ of $\alpha^* = 0.0957$, with $|\widetilde{T}(\alpha^*)| = 12$ and $I(T(\alpha^*)) = 1.13$.

Furthermore, 100 repetitions of the simulation from (6) were made. Figure 2 provides the histograms of the estimated split levels on X_1, X_2 and X_4, collected from the 100 optimally pruned trees.

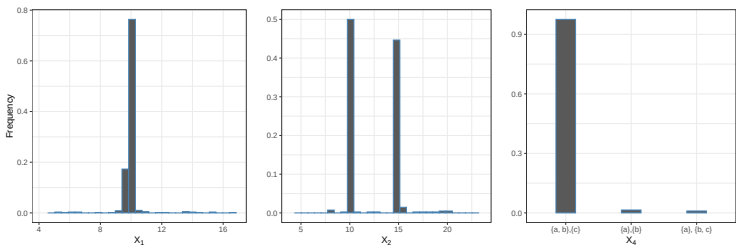


Figure 2: The histogram of split levels in 100 pruned trees

The selected splits concentrated around the genuine split levels with high probability. Specifically, 95% of the splits on X_1 were within $(9, 11)$ (the true split at $X_1 = 10$), 96% of the splits on X_2 were within $(9, 11) \cup (14, 16)$ (the true splits at $X_2 = 10$ and 15), and nearly 98% of the splits on X_4 were in the form of $\{\{a, b\}, \{c\}\}$, This strongly supported the consistency results of the split selection procedure and validated the pruning procedures could effectively remove the redundant splits.

The second example is a regression function that does not conform the segmented linear form,

$$m(X) = \max\{e^{-10X_1^2}, e^{-50X_2^2} 1.25 e^{-5(X_1^2+X_2^2)}\},$$

which was also considered in [22]. Figure 3 demonstrates the surface of $m(X)$ within the domain of $[-1, 1]^2$. We generated the training data of 1000 records $Y = m(X) + \varepsilon$, for $X_1, X_2 \overset{i.i.d}{\sim} U(-1, 1)$ and $\varepsilon \sim N(0, 0.01)$. With the same stopping parameter of $N_{\min} = 10, \text{Dep}_{\max} = 10$, we applied SLRT and CART respectively, obtaining the approximated surface in Figure 4 and 5.

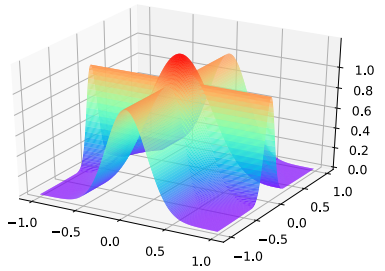


Figure 3: The true surface defined by $m(X)$

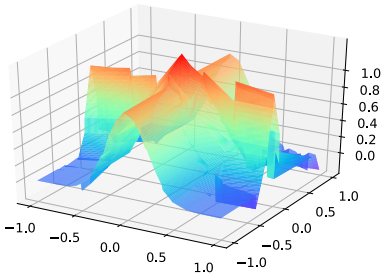


Figure 4: The approximated surface by SLRT

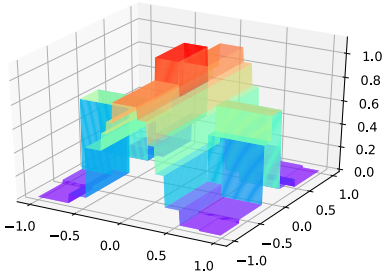


Figure 5: The approximated surface by CART

Then we calculated the root mean squared prediction errors (RMSPE) on an independent testing sample of size 500. The RMSPE of SLRT is 0.047, and that of CART is higher at 0.073. Under this situation, SLRT obtains a locally linear approximation with a large tree structure, which tend to outperforms CART since it is locally constant. In practice, since the model complexity (the tree size) is adaptive to the nature of data, it depends on data whether the estimated regression function is a interpretable segmented linear approximation or a locally linear approximation.

5.2 Comparisons on Benchmark Datasets

The predictive performance is tested on 9 benchmark datasets from the StatLib library [23] and the UCI Machine Learning Repository [24], where the sample sizes range from 74(Pyrimidine) to 39644(News Popularity) and 5 datasets include categorical variables. Detailed information about the covariates and sample sizes are reported in the supplementary materials.

The proposed SLRT with the least square estimation (SLRT_{LS}) and the LASSO ($\text{SLRT}_{\text{LASSO}}$) were compared with three tree-based methods: CART [11], GUIDE [17, 25] and MARS [26], with the

same $N_{\min}$ and $\text{Dep}_{\max}$ for all the methods. Ensemble predictors RF_{SLRT} and WRF_{SLRT} are random forests (RF) and the newly proposed weighted random forests (WRF) equipped with SLRT as the base predictor. The conventional RF based on CART (RF_{CART}) as well as WRF with CART (WRF_{CART}) were also implemented to serve as benchmarks. To make the results of RF and WRF comparable, their predictions were based on the same ensembles of 50 trees and only different in the way of aggregation. To evaluate the predictive performance, we divided each dataset into 10 subsets and implemented each method for 10 times using each subset as the testing set and the rest as the training set, where all methods shared the same training and testing sets. Table 1 summarizes the average RMSPEs from the 10-fold cross validation, where the integers in parentheses indicate the ranks within the single and the ensemble predictors, respectively.

Table 1: RMSPE (rank) of 10-fold cross-validation on 9 data sets: data name(sample size). The best performance is marked in ***bold italic***, within each group of single predictors and ensemble predictors.

Datasets	Single Predictors					Ensemble Predictors			
	SLRT_{LS}	$\text{SLRT}_{\text{LASSO}}$	GUIDE	CART	MARS	RF_{SLRT}	WRF_{SLRT}	RF_{CART}	WRF_{CART}
Boston Housing(506)	0.174(2)	***0.170***(1)	0.187(4)	0.262(5)	0.179(3)	0.162(2)	***0.158***(1)	0.218(4)	0.200(3)
ComputerHardware(209)	47.89(2)	***47.40***(1)	48.06(3)	62.60(5)	54.17(4)	38.49(2)	***36.77***(1)	64.91(4)	39.08(3)
Auto-mpg(392)	2.831(2)	***2.791***(1)	3.545(4)	3.680(5)	2.942(3)	2.633(2)	***2.614***(1)	3.273(3)	3.240(4)
Auto-mobile(159)	0.154(2)	***0.140***(1)	0.231(5)	0.192(4)	0.184(3)	0.143(2)	***0.142***(1)	0.165(4)	0.162(3)
Kinematics(8192)	0.139(2)	***0.138***(1)	0.140(3)	0.257(5)	0.198(4)	0.117(2)	***0.115***(1)	0.249(4)	0.247(3)
Abalone(4176)	2.162(3)	***2.143***(1)	2.151(2)	2.497(5)	2.161(3)	2.116(2)	***2.113***(1)	2.458(4)	2.456(3)
Parkinson(5875)	9.374(3)	9.327(2)	***9.300***(1)	10.534(5)	9.660(4)	8.691(2)	***8.679***(1)	10.326(4)	10.317(3)
Pyrimidine(74)	0.088(2)	0.093(3)	0.096(5)	0.094(4)	***0.074***(1)	0.078(4)	0.076(3)	0.073(2)	***0.049***(1)
NewsPopularity(39644)	0.877(4)	0.872(2)	0.873(3)	0.903(5)	***0.865***(1)	***0.868***(1)	***0.868***(1)	0.901(3)	0.901(3)

Among the five single tree predictors, $\text{SLRT}_{\text{LASSO}}$ attained the best prediction in six dataset, MARS in two, and SLRT_{LS} and GUIDE in one, respectively. This demonstrates the advantages of the proposed SLRT. Directly comparing SLRT with GUIDE, SLRT_{LS} had better prediction in 7 out of the 9 datasets and $\text{SLRT}_{\text{LASSO}}$ in 8 out of 9 datasets. SLRT also compared favorably to MARS, having better performance in 6 out of the 9 datasets. CART appeared to be the worst predictor in seven datasets, while GUIDE ranked the last on the other two datasets. The better performance of $\text{SLRT}_{\text{LASSO}}$ over SLRT_{LS} shows the benefits of conducting the variables selection on the leaf nodes.

The ensemble predictors RF_{SLRT} and WRF_{SLRT} showed better performance than the conventional RF with CART in 8 out of 9 datasets. Meanwhile, the ensemble predictors tended to outperform the single predictors, which suggests the effects of the bagging operation. The proposed WRF also showed improved predictions over the RF, which benefit from the weighting procedure that reduces the importance of those under-performing trees.

6 Conclusion

We propose a tree based approach called segmented linear regression trees (SLRT), which is based on two consecutive algorithms for partitioning structure learning: one for the split selection at each internal node based on a cumulative Kendall's τ statistic; the other for the parsimonious partitioning structure by tree pruning through an adaptive cost-complexity measure. Theoretical analysis shows that the split selection algorithm leads to the consistent identification and estimation of both the genuine split variables and the split levels, and the pruning procedure ensures the consistent estimation of the genuine number of segments. We implant the SLRT as the base predictor in RF and WRF to create two breeds of ensemble predictors. The proposed procedures are evaluated by numerical simulations and case studies, which shows advantageous predictive accuracy over other tree-based methods, and in creating more powerful breeds of ensemble predictors.

Acknowledgments

This research is funded by China's National Key Research Special Program Grant 2016YFC0207701, National Key Basic Research Program Grant 2015CB856000 and National Natural Science Foundation of China grant 71532001.

References

[1] H. Oiwa and R. Fujimaki, "Partition-wise linear models," in *NeurIPS*, 2014.

[2] J. Kim and H. J. Kim, "Asymptotic results in segmented multiple regression," *Journal of Multivariate Analysis*, vol. 99, no. 9, pp. 2016–2038, 2008.

[3] P. Perron and Z. Qu, "Estimating restricted structural change models," *Journal of Econometrics*, vol. 134, pp. 373–399, 2006.

[4] J. Liu, S. Wu, and J. V. Zidek, "On segmented multivariate regression," *Statistica Sinica*, vol. 7, pp. 497–525, 1997.

[5] J. Gonzaloa and J.-Y. Pitarakisb, "Estimation and model selection based inference in single and multiple threshold models," *Journal of Econometrics*, vol. 110, no. 2, pp. 319–352, 2002.

[6] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[7] H. Trevor, T. Robert, and F. JH, *The Elements of Statistical Learning: data mining, inference, and prediction*. New York: Springer series in statistics, 2009.

[8] H.-J. Kim, B. Yu, and E. J. Feuer, "Selecting the number of change-points in segmented line regression," *Statistica Sinica*, vol. 19, no. 2, p. 597, 2009.

[9] J. Wang and V. Saligrama, "Local supervised learning through space partitioning," in *NeurIPS*, 2012.

[10] X. Fan, B. Li, and S. SIsson, "Rectangular bounding process," in *NeurIPS*, 2018.

[11] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*. New York: Wadsworth International Group, 1984.

[12] J. R. Quinlan, "Combining instance-based and model-based learning," in *ICML*, 1993.

[13] L. Torgo, "Partial linear trees," in *ICML*, 2000.

[14] Y. Wang and I. H. Witten, "Inducing model trees for continuous classes," in *ECML*, 1997.

[15] W. P. Alexander and S. D. Grimshaw, "Treed regression," *Journal of Computational and Graphical Statistics*, vol. 5, no. 2, pp. 156–175, 1996.

[16] K.-C. Li, H.-H. Lue, and C.-H. Chen, "Interactive tree-structured regression via principal hessian directions," *Journal of the American Statistical Association*, vol. 95, pp. 547–560, 2000.

[17] W.-Y. Loh, "Regression trees with unbiased variable selection and interaction detection," *Statistica Sinica*, vol. 12, pp. 361–386, 2002.

[18] W.-Y. Loh and W. Zheng, "Regression trees for longitudinal and multi-response data," *The Annals of Applied Statistics*, vol. 7, no. 1, pp. 495–522, 2013.

[19] D. Malerba, F. Esposito, M. Ceci, and A. Appice, "Top-down induction of model trees with regression and splitting nodes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 612–625, 2004.

[20] A. Karalič, "Linear regression in regression tree leaves," in *ECAI*, 1992.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] D. Potts and C. Sammut, "Incremental learning of linear model trees," *Machine Learning*, vol. 61, no. 1-3, pp. 5–48, 2005.

[23] "Statlib library," http://lib.stat.cmu.edu/datasets/.

[24] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," http://archive.ics.uci.edu/ml, 2017.

[25] W.-Y. Loh, "Guide classification and regression trees and forests version 29.7," https://www.stat.wisc.edu/~loh/guide.html, 2018.

[26] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.

中国科学: 数学 2020 年 第 50 卷 第 4 期: 527 ~ 558
SCIENTIA SINICA Mathematica
论 文

气象调整下的区域空气质量评估

张澍一[1], 陈松蹊[1,2*], 郭斌[3,4], 王恒放[5], 林伟[2,6]

1. 北京大学光华管理学院, 北京 100871;
2. 北京大学统计科学中心, 北京 100871;
3. 西南财经大学统计研究中心, 成都 611130;
4. 西南财经大学统计学院, 成都 611130;
5. Department of Statistics, Iowa State University, Ames, IA 50011, USA;
6. 北京大学数学科学学院, 北京 100871

E-mail: shuyizhang@fas.harvard.edu, csx@gsm.pku.edu.cn, guobin@swufe.edu.cn, hengfang@iastate.edu, weilin@math.pku.edu.cn

收稿日期: 2019-06-11; 接受日期: 2019-12-18; 网络出版日期: 2020-04-01; * 通信作者
科技部国家重点研发计划 (批准号: 2016YFC0207701, 2016YFC0207702 和 2016YFC0207703)、国家重点基础研究发展计划 (批准号: 2015CB856000)、国家自然科学基金 (批准号: 71532001, 71371016, 11971390 和 11671018)、北京市自然科学基金 (批准号: Z190001) 和中央高校基本科研业务费专项资金 (批准号: JBK1806002) 资助项目

摘要 虽然空气污染是由污染物排放到大气中造成的, 但是由于气象条件会影响污染物的扩散, 因而实际观测到的污染水平会受到气象条件的影响. 因此, 有效的空气质量管理要求污染评估指标和统计方法不受气象因素的干扰, 并能准确客观地反映污染物浓度的变化. 为了评估北京地区潜在污染物排放的变化, 本文提出一种消除气象干扰的时空调整方法. 通过控制气象条件, 调整后的污染物时空平均浓度可以捕捉到潜在排放量的变化. 本文提出具体调整均值的方法, 并进行理论和数值分析, 将此方法应用于北京地区的空气质量评估, 揭示一些有趣的模式和趋势, 这些结果可以用于空气质量评估和管理.

关键词 空气质量评估 气象混杂 非参数回归 时空调整

MSC (2010) 主题分类 62-07, 62P12, 62G08

1 引言

近 20 年来, 中国的快速工业化也带来了严重的空气污染. 北京周边地区受影响最大. 中国城市的主要空气污染物是颗粒物 $PM_{2.5}$ 和 PM_{10} (参见文献 [1, 2]), 它们是空气动力学当量直径分别小于 2.5 和 10 μm 的空气颗粒. 近年来, 我国地面臭氧 (O_3) 也呈上升趋势 (参见文献 [3]).

改善空气质量的关键是减少排放, 这需要及时和准确地计算排放量. 排放源清单是用于排放测量的常用工具, 可收集工业数据并将这些数据调整至更高的分辨率 (参见文献 [4]). 此清单通常处于每年

英文引用格式: Zhang S Y, Chen S X, Guo B, et al. Regional air-quality assessment that adjusts for meteorological confounding (in Chinese). Sci Sin Math, 2020, 50: 527–558, doi: 10.1360/SCM-2019-0368

或更短的时间频率, 并且容易出现测量和报告错误. 在中国, 尽管有一些排放源清单, 但它们通常要滞后 3 到 4 年.

本文提出使用小时的空气质量数据进行排放水平的量化. 这里最直接的一个挑战是观测到的污染水平会被气象条件所干扰, 例如, 在 Liang 等 [5] 和 Finazzi 等 [6] 的研究中, 风向、风速和相对湿度都会对污染水平产生一定的影响. 气象因素对空气污染造成的影响与观测研究 (参见文献 [7,8]) 相似. 一般观测研究在评估治疗效果时, 需要调整由于协变量差异而引起的偏差. 但是, 我们的设置与协变量遵循相同基准分布的观察性研究有所不同 (参见文献 [8,9]). 在我们的研究中, 需要构建合适的气象基准来描述气象的变化. 本研究与现有有关治疗效果评估文献的另一个主要区别是, 这里的气象条件是固定的, 无法进行随机分配, 因此现有的基于倾向得分 (propensity score) 的方法是不能直接使用的.

为了得到在时间和空间上可比的均值和分位数, 我们提出一种在时间和空间两个维度对观测到的浓度中气象因素进行调整的新方法. 调整后的均值可以在不同年份之间进行比较, 从而可以获得有关排放量是否减少的信息. Thompson 等 [10] 提出的趋势分析也可以进行气象调整. 趋势分析使用线性回归, 它可以作为特殊情形包含在我们新提出的调整框架中. 另一种方法是美国环境保护署 (EPA) 提倡的 3 年移动平均法, 该方法的局限已经在文献 [3] 中给出. 我们新提出的调整方法的一个优点是它考虑的是一般的回归模型. 此外, 它在时间调整中也考虑了空间的变化.

本文余下内容的结构如下. 第 2 节描述研究区域、数据和模型. 第 3 节概述时间和空间调整方法及其在测量潜在排放量方面的优势. 调整后的区域空气质量估计量及其理论性质将在第 4 节中给出. 新的估计量的方差估计和假设检验在第 5 节给出. 第 6 节利用新提出的方法评估北京周边的空气质量. 我们将一些假设条件、理论结果的证明、模拟研究以及其他实际结果归纳到附录中.

2 研究区域、数据和模型

中国于 2013 年 1 月在 74 个城市建立了 496 个国控监测点的空气质量监测网络, 并于 2015 年 1 月在 338 个城市扩展到 1438 个监测点. Liang 等 [11] 交叉对比了来自中国 5 个城市的美国使领馆监测的 $PM_{2.5}$ 数据与邻近国控站点的 $PM_{2.5}$ 数据, 说明了国控监测站点数据的准确性且具有很高的质量. 从 2008 年 4 月起, 美国驻北京大使馆开始报告每小时的 $PM_{2.5}$ 浓度. 作为国家网络的一部分, 北京市环境监测中心 (BMEMC) 管理着一个监测网络, 该网络由 35 个空气质量监测站点组成, 每小时监测 $PM_{2.5}$、PM_{10}、二氧化硫 (SO_2)、二氧化氮 (NO_2)、一氧化碳 (CO) 和臭氧 (O_3) 共 6 种污染物的浓度, 而美国驻北京大使馆仅监测 $PM_{2.5}$ 的浓度. 在我们的研究中, 考虑的不是通常的日历年, 而是使用从每年的三月到次年的二月的季节年度, 该季节年度涵盖了从春季到冬季 4 个季节 (春季: 3–5 月, 夏季: 6–8 月, 秋季: 9–11 月, 冬季: 12 月到次年 2 月).

如图 1 所示, 我们关注的研究区域是整个华北平原 (NCP) 的一部分, 其经度从 116.0°E 到 116.8°E, 纬度从 39.5°N 到 40.2°N, 共占地 5180 平方公里. 我们的研究范围覆盖了北京的核心区域, 该区域由六环以及六环与河北省的边界之间的南部地区构成. 该研究区域包括美国大使馆观测站点在内, 共计 28 个监测点. 我们将研究区域划分成两部分: “中心区域” 和 “南部区域”, 其分别包含了 25 和 3 个监测站点. 由于与北京南部相接壤的河北拥有大量的高排放行业 (如钢铁冶炼), 因此其空气质量较差. 这里设置的 “南部区域” 有助于帮助我们理解河北对北京污染的传播.

为了调整气象干扰因素, 我们使用中央气象局 (CMA) 的 11 个气象站 (图 1 中蓝色三角形) 的气象数据. 气象变量包含每小时的气温、气压、相对湿度、露点温度、风向、累积风速和累积降水量的

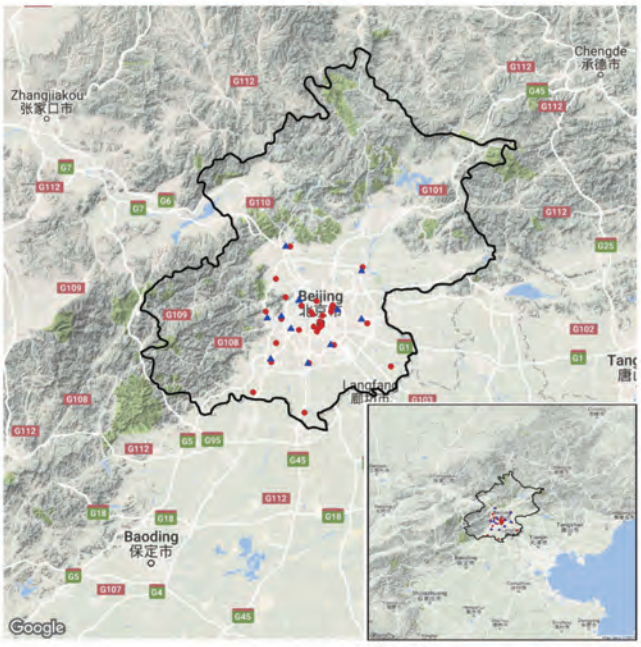


图 1 (网络版彩图) 北京地区空气质量监测站 (红色圆点) 和气象站 (蓝色三角形) 的位置. 嵌入图: 华北平原内的研究区域以及西部和北部的山脉

测量值. 风向是分为 5 类的无序离散变量: 西北风 (northwest, NW)、东北风 (northeast, NE)、东南风 (southeast, SE)、西南风 (southwest, SW)、静风和无固定方向 (calm and variable, CV). 根据 Magnus 公式 [12], 露点温度可以在数学上表示为相对湿度和温度的一个已知的非线性函数. 因此, 为了减少协变量的个数, 我们决定省略一个变量. 由于相对湿度是一个值在 $[0,1]$ 之间的有界变量, 故在使用非参数方法进行处理时会存在不可忽略的边界偏差 (参见文献 [13, 第 202 页]). 所以, 在这 3 个气象变量中, 我们将选用大气温度和露点温度, 而未考虑湿度变量.

假设在研究区域内共有 L 个大气污染物监测站点, 所有污染物监测站点构成了整个研究区域 $\mathcal{R}$. 气象站点有 S 个, 所有的气象站点的空间位置构成了集合 $\mathcal{W}$. 在空气质量监测点 $\boldsymbol{s}$, 记 $Y_{ijt}(\boldsymbol{s})$ 为第 i 年第 j 季度的 t 时刻污染物的浓度, 其中 j 取 1、2、3 和 4 分别是春季、夏季、秋季和冬季, $\boldsymbol{X}_{ijt}(\boldsymbol{s})$ 是 6 维气象变量, 包括气压、气温、露点温度、风向、该风向下的累积风速和累积降水. 这些气象变量来自于最接近空气质量监测点 $\boldsymbol{s}$ 的气象观测站.

记 $\boldsymbol{U}_{ijt}(\boldsymbol{s})$ 为排放水平, 在实际数据中, $\boldsymbol{U}_{ijt}(\boldsymbol{s})$ 可以是与能源消耗和社会经济活动有关的变量. 但是, 这些变量往往观测频率比较大, 多数为月度和年度数据, 且存在一定的时间滞后, 这也妨碍了排放源清单的及时产生. 这里考虑一个描述 $Y_{ijt}(\boldsymbol{s})$ 与 $\{\boldsymbol{X}_{ijt}(\boldsymbol{s})^{\mathrm{T}},\boldsymbol{U}_{ijt}(\boldsymbol{s})\}^{\mathrm{T}}$ 之间关系的基本模型:

$$Y_{ijt}(\boldsymbol{s})=\tilde{m}_j\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{U}_{ijt}(\boldsymbol{s})\}+\tilde{\epsilon}_{ijt}(\boldsymbol{s}),\tag{2.1}$$

其中 $t=1,\ldots,n_{ij}$, $\tilde{m}_j\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{U}_{ijt}(\boldsymbol{s})\}=\mathrm{E}\{Y_{ijt}(\boldsymbol{s})\mid\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{U}_{ijt}(\boldsymbol{s})\}$, $\tilde{\epsilon}_{ijt}(\boldsymbol{s})$ 是残差, n_{ij} 是第 i 年 j 季度的小时观测数.

由于 $\boldsymbol{U}_{ijt}(\boldsymbol{s})$ 是观测不到的潜在变量, 在 (2.1) 两边同时取对 $\boldsymbol{X}_{ijt}(\boldsymbol{s})$ 的条件期望, 于是得到

$$Y_{ijt}(\boldsymbol{s})=m_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{s}\}+\sigma_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{s}\}e_{ijt}(\boldsymbol{s}),\tag{2.2}$$

其中 $m_{ij}(\boldsymbol{x},\boldsymbol{s})=\mathrm{E}\{Y_{ijt}(\boldsymbol{s})\mid \boldsymbol{X}_{ijt}(\boldsymbol{s})=\boldsymbol{x}\}$, $\sigma_{ij}^2(\boldsymbol{x},\boldsymbol{s})=\mathrm{Var}\{Y_{ijt}(\boldsymbol{s})\mid \boldsymbol{X}_{ijt}(\boldsymbol{s})=x\}$, $e_{ijt}(\boldsymbol{s})$ 是标准化残差. 此外, 通过计算给定 $\boldsymbol{X}_{ijt}(\boldsymbol{s})$ 下 (2.2) 右侧的条件期望, 我们可以得到 $m_{ij}(\boldsymbol{x},\boldsymbol{s})$ 的另一个表达式 $m_{ij}(\boldsymbol{x},\boldsymbol{s})=\mathrm{E}[\tilde{m}_j\{\boldsymbol{x},\boldsymbol{U}_{ijt}(\boldsymbol{s})\}\mid \boldsymbol{X}_{ijt}(\boldsymbol{s})=\boldsymbol{x}]$, 这本质上是由给定 $\boldsymbol{X}_{ijt}(\boldsymbol{s})$ 下 $\boldsymbol{U}_{ijt}(\boldsymbol{s})$ 的条件分布决定的. 我们用模型 (2.2) 来进行统计推断. 需要注意的是, 由于 $\boldsymbol{U}_{ijt}(\boldsymbol{s})$ 的分布可能会随年份和空间而变化, 因此在回归方程 $m_{ij}(\boldsymbol{x},\boldsymbol{s})$ 里包括了年份 i 和站点位置 $\boldsymbol{s}$.

记 $\mathbb{X}_{ijt}=\{\boldsymbol{X}_{ijt}(\boldsymbol{s}_1)^{\mathrm{T}},\ldots,\boldsymbol{X}_{ijt}(\boldsymbol{s}_L)^{\mathrm{T}}\}^{\mathrm{T}}$, $\boldsymbol{e}_{ijt}=\{e_{ijt}(\boldsymbol{s}_1),\ldots,e_{ijt}(\boldsymbol{s}_L)\}^{\mathrm{T}}$, L 是空气质量监测站点的数量. 那么 $\mathbb{X}_{ijt}$ 和 $\boldsymbol{e}_{ijt}$ 就分别是 $6L$ 维和 L 维数据, 代表该地区所有站点的气象变量和标准化残差数据. 假设多元时间序列 $\{\mathbb{X}_{ijt}\}_{t=1}^{n_{ij}}$ 和 $\{\boldsymbol{e}_{ijt}\}_{t=1}^{n_{ij}}$ 在时间上是平稳的并且满足 α 混合过程.

为了更好地讨论排放物和气象的综合影响, 我们可以假设 $\tilde{m}_j\{\boldsymbol{X}(\boldsymbol{s}),\boldsymbol{U}(\boldsymbol{s})\}$ 的结构是

$$\tilde{m}_j\{\boldsymbol{X}(\boldsymbol{s}),\boldsymbol{U}(\boldsymbol{s})\}=\tilde{m}_{j,1}\{\boldsymbol{X}(\boldsymbol{s})\}+\tilde{m}_{j,2}\{\boldsymbol{U}(\boldsymbol{s})\}+\tilde{m}_{j,3}\{\boldsymbol{X}(\boldsymbol{s}),\boldsymbol{U}(\boldsymbol{s})\},\tag{2.3}$$

其中, 第一项关于气象向量的单独效应是不依赖于年份和空间点而只依赖于季节指标的. 观测数据的对应版本为

$$m_{ij}(\boldsymbol{x},\boldsymbol{s})=\tilde{m}_{j,1}(\boldsymbol{x})+\mathrm{E}[\tilde{m}_{j,2}\{\boldsymbol{U}_{ijt}(\boldsymbol{s})\}\mid \boldsymbol{X}_{ijt}(\boldsymbol{s})=\boldsymbol{x}]+\mathrm{E}[\tilde{m}_{j,3}\{\boldsymbol{x},\boldsymbol{U}_{ijt}(\boldsymbol{s})\}\mid \boldsymbol{X}_{ijt}(\boldsymbol{s})=\boldsymbol{x}].\tag{2.4}$$

这里 (2.3) 和 (2.4) 都是为了描述下一节的排放效应.

还需要注意的是, 因为本文考虑的是空气质量评估而非预测问题, 因此无需构建 (2.2) 的参数模型, 也没有必要包括时间上的滞后和空间邻近项. 要达到评估的目的, 非参数模型 (2.2) 是足够的.

3 时空调整方法

在模型 (2.2) 的基础上, 我们提出了在时间和空间两个维度上进行气象调整的调整均值和分位数, 用以控制气象这一混杂因素. 在中国和其他一些国家, 常采用一段时间内的平均污染物浓度作为空气质量的度量, 但这会受到气象混杂因素的影响而不能反映真实的排放水平. Thompson 等[10] 考虑了线性回归模型下地面臭氧污染的趋势分析. 我们也将说明, 趋势分析是我们新提出的方法的特例.

一般地, $Y_{ijt}(\boldsymbol{s})$ 的平均值可以表示为 $\mathrm{E}\{Y_{ijt}(\boldsymbol{s})\}=\mathrm{E}[\mathrm{E}\{Y_{ijt}(\boldsymbol{s})\mid \boldsymbol{X}_{ijt}(\boldsymbol{s})\}]=\mathrm{E}[m_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{s}\}]$. 这里的关键问题是在最后的期望中使用哪种概率密度. 均值的一般形式是

$$\mathrm{E}\{Y_{ijt}(\boldsymbol{s})\}=\int m_{ij}(\boldsymbol{x},\boldsymbol{s})f_j(\boldsymbol{x},\boldsymbol{s})d\boldsymbol{x},\tag{3.1}$$

其中 $f_j(\boldsymbol{x},\boldsymbol{s})$ 表示 $\boldsymbol{X}_{ijt}(\boldsymbol{s})$ 的某种边缘概率密度函数, $f_j(\boldsymbol{x},\boldsymbol{s})$ 的不同形式产生以下介绍的不同度量.

记 $f_{ij}(\boldsymbol{x},\boldsymbol{s})$ 为站点 $\boldsymbol{s}$ 第 i 年第 j 季的 $\boldsymbol{X}_{ijt}(\boldsymbol{s})$ 的密度. 如果选择 $f_j(\boldsymbol{x},\boldsymbol{s})=f_{ij}(\boldsymbol{x},\boldsymbol{s})$, 则得到总体均值, 记为 $\mu_{ij}^0(\boldsymbol{s})$. 常用的空气质量度量是其简单平均 $\bar{Y}_{ij}(\boldsymbol{s})=n_{ij}^{-1}\sum_{t=1}^{n_{ij}}Y_{ijt}(\boldsymbol{s})$. 根据弱相关过程的大数定律可知, 当 $n_{ij}\to\infty$ 时, $\bar{Y}_{ij}(\boldsymbol{s})\xrightarrow{p}\mu_{ij}^0(\boldsymbol{s})$. 尽管 $\mu_{ij}^0(\boldsymbol{s})$ 和 $\bar{Y}_{ij}(\boldsymbol{s})$ 可以作为用于健康评估的空气暴露指标, 但是它受不同年份和空间点的气象分布差异的影响, 不能反映由于排放变化所导致的大气污染物浓度的变化.

考虑 $f_j(\boldsymbol{x},\boldsymbol{s})$ 的另一种形式是所有 A_j 年 j 季度 $\{f_{aj}(\boldsymbol{x},\boldsymbol{s})\}_{a=1}^{A_j}$ 的均值, 即

$$f_{\cdot j}(\boldsymbol{x},\boldsymbol{s})=A_j^{-1}\sum_{a=1}^{A_j}f_{aj}(\boldsymbol{x},\boldsymbol{s}),$$

我们称其为气象变量在时间上均衡的边缘概率密度函数. 在 (3.1) 中令 $f_j(\boldsymbol{x},\boldsymbol{s})=f_{\cdot j}(\boldsymbol{x},\boldsymbol{s})$, 得出

$$\tilde{\mu}_{ij}(\boldsymbol{s})=\int m_{ij}(\boldsymbol{x},\boldsymbol{s})f_{\cdot j}(\boldsymbol{x},\boldsymbol{s})d\boldsymbol{x}=A_j^{-1}\sum_{a=1}^{A_j}\int m_{ij}(\boldsymbol{x},\boldsymbol{s})f_{aj}(\boldsymbol{x},\boldsymbol{s})d\boldsymbol{x}. \tag{3.2}$$

在 (3.2) 中, 当 $a\neq i$ 时, $\int m_{ij}(\boldsymbol{x},\boldsymbol{s})f_{aj}(\boldsymbol{x},\boldsymbol{s})d\boldsymbol{x}$ 称为反事实 (counter factuals) [14], 它表示了在年份 i 的排放条件下, 如果气象条件采用的是年份 a 的分布, 所得到污染物浓度的均值. 我们称 $\tilde{\mu}_{ij}(\boldsymbol{s})$ 为在时间上进行气象调整后的调整均值.

由于 $\{\tilde{\mu}_{aj}(\boldsymbol{s})\}_{a=1}^{A_j}$ 由时间均衡的概率密度函数 $f_{\cdot j}(\boldsymbol{x},\boldsymbol{s})$ 计算, 其在不同年份是可比的. 特别地,

$$\mu_{ij}^0(\boldsymbol{s})-\tilde{\mu}_{ij}(\boldsymbol{s})=\int m_{ij}(\boldsymbol{x},\boldsymbol{s})\{f_{ij}(\boldsymbol{x},\boldsymbol{s})-f_{\cdot j}(\boldsymbol{x},\boldsymbol{s})\}d\boldsymbol{x}$$

可以用来度量在年份 i 观测到的污染物均值浓度与在时间上进行气象调整的调整均值的差值. 这个差值是由于不同年的气象混杂因素所导致的, 并且

$$\tilde{\mu}_{ij}(\boldsymbol{s})-\tilde{\mu}_{kj}(\boldsymbol{s})=\int\{m_{ij}(\boldsymbol{x},\boldsymbol{s})-m_{kj}(\boldsymbol{x},\boldsymbol{s})\}f_{\cdot j}(\boldsymbol{x},\boldsymbol{s})d\boldsymbol{x}$$

可以用来度量年份 i 和 k 由于排放不同所导致的污染物浓度的差异, 这个比较是通过使用在时间上均衡的气象分布而实现的.

Thompson 等 [10] 考虑通过在回归模型下进行趋势分析, 从而实现气象调整. 趋势分析可以包含在我们提出的框架中. 为了说明这一点, 假设回归函数 $m_{ij}(\boldsymbol{x},\boldsymbol{s})$ 是线性的, 即

$$Y_{ijt}(\boldsymbol{s})=\alpha_{ij}(\boldsymbol{s})+\boldsymbol{\beta}_{ij}^{\mathrm{T}}(\boldsymbol{s})\tilde{\boldsymbol{X}}_{ijt}(\boldsymbol{s})+\epsilon_{ijt}(\boldsymbol{s}), \tag{3.3}$$

其中 $\tilde{\boldsymbol{X}}_{ijt}(\boldsymbol{s})$ 是通过所有 A_j 年的气象均值进行中心化的气象向量. 由于协变量的时间中心化, 我们可以得到 $\tilde{\mu}_{ij}(\boldsymbol{s})=\alpha_{ij}(\boldsymbol{s})$, 这一点 Thompson 等 [10] 没有明确指出. 此外, 由于 $\alpha_{ij}(\boldsymbol{s})=\mu_{ij}(\boldsymbol{s})$, $\alpha_{ij}(\boldsymbol{s})$ 也可以在时间和空间上进行比较, 其中时空调整均值 $\mu_{ij}(\boldsymbol{s})$ 在 (3.5) 中定义. 需要注意到的是, 我们提出的通过 $\tilde{\mu}_{ij}(\boldsymbol{s})$ 和 $\mu_{ij}(\boldsymbol{s})$ 的调整方法允许更具普遍性的回归模型, 趋势分析中采用的线性模型只是一种特殊情形.

在两个不同的站点 $\boldsymbol{s}_1$ 和 $\boldsymbol{s}_2$, 时间调整后的均值 $\{\tilde{\mu}_{aj}(\boldsymbol{s}_1)\}_{a=1}^{A_j}$ 和 $\{\tilde{\mu}_{aj}(\boldsymbol{s}_2)\}_{a=1}^{A_j}$ 在空间上是不可比的, 因为这两个站点的气象分布不同. 在下文中, 我们将空间变量纳入时间调整中. 如前所述, 研究的区域 $\mathcal{R}$ 具有 S 个气象站点, 将包含所有气象站空间位置的集合表示为 $\mathcal{W}$. 我们可以将空间和时间的均衡气象分布定义为 $f_{\cdot j}(\boldsymbol{x},\boldsymbol{s})$ 的加权形式. 具体来说, 设 $p(\boldsymbol{s})$ 表示研究区域上的概率密度函数, 我们可以构建一个加权的时空均衡气象:

$$f_{\cdot j}^p(\boldsymbol{x})=\int_{\boldsymbol{s}\in\mathcal{R}}f_{\cdot j}(\boldsymbol{x},\boldsymbol{s})p(\boldsymbol{s})d\boldsymbol{s}.$$

在空间上的固定设计采样下, 如果将 $p(\boldsymbol{s})$ 选取为各气象站点的均匀分布密度函数, 则可以得到

$$f_{\cdot j}(\boldsymbol{x})=S^{-1}\sum_{\boldsymbol{s}'\in\mathcal{W}}f_{\cdot j}(\boldsymbol{x},\boldsymbol{s}'). \tag{3.4}$$

我们称 $f_{\cdot j}(\boldsymbol{x})$ 为区域 $\mathcal{R}$ 上季节 j 的时空均衡气象分布. 为了简化分析, 我们将采用这个非加权的气象基准 $f_{\cdot j}(\boldsymbol{x})$. 通过 (3.4) 中的 $f_{\cdot j}(\boldsymbol{x})$, 我们得到空间上和时间上的调整均值:

$$\mu_{ij}(\boldsymbol{s})=\int m_{ij}(\boldsymbol{x},\boldsymbol{s})f_{\cdot j}(\boldsymbol{x})d\boldsymbol{x}=S^{-1}A_j^{-1}\sum_{\boldsymbol{s}'\in\mathcal{W}}\sum_{a=1}^{A_j}\int m_{ij}(\boldsymbol{x},\boldsymbol{s})f_{aj}(\boldsymbol{x},\boldsymbol{s}')d\boldsymbol{x}, \tag{3.5}$$

其中带有 $a \neq i$ 或者 $s' \neq s$ 的项是关于时间和空间两个维度的反事实.

在加性模型 (2.3) 和 (2.4) 下, 定义

$$\mu_j^M = \int \tilde{m}_{j,1}(\boldsymbol{x}) f_{\cdot j}(\boldsymbol{x}) d\boldsymbol{x},$$
$$\mu_{ij}^E(\boldsymbol{s}) = \int \mathrm{E}[\tilde{m}_{j,2}\{\boldsymbol{U}_{ijt}(\boldsymbol{s})\} \mid \boldsymbol{X}_{ijt}(\boldsymbol{s}) = \boldsymbol{x}] f_{\cdot j}(\boldsymbol{x}) d\boldsymbol{x},$$
$$\mu_{ij}^{ME}(\boldsymbol{s}) = \int \mathrm{E}[\tilde{m}_{j,3}\{\boldsymbol{x}, \boldsymbol{U}_{ijt}(\boldsymbol{s})\} \mid \boldsymbol{X}_{ijt}(\boldsymbol{s}) = \boldsymbol{x}] f_{\cdot j}(\boldsymbol{x}) d\boldsymbol{x}.$$

在 (3.5) 中, $\mu_{ij}(\boldsymbol{s}) = \mu_j^M + \mu_{ij}^E(\boldsymbol{s}) + \mu_{ij}^{ME}(\boldsymbol{s})$. 注意到对于气象效应项 μ_j^M, 由于时间空间的调整, 其只与季节 j 有关, 而与年份 i 和空间点 $\boldsymbol{s}$ 无关. 然而, 排放效应 $\mu_{ij}^E(\boldsymbol{s})$ 和交互效应 $\mu_{ij}^{ME}(\boldsymbol{s})$ 在不同年份和空间点可以不相等, 这是因为对于给定的季节 j, 排放变量 $\boldsymbol{U}_{ijt}(\boldsymbol{s})$ 在不同年份和空间点往往具有不同的分布.

因此, 连续两年 j 季节的年度差异为

$$\mu_{ij}(\boldsymbol{s}) - \mu_{i-1,j}(\boldsymbol{s}) = \mu_{ij}^E(\boldsymbol{s}) - \mu_{i-1,j}^E(\boldsymbol{s}) + \mu_{ij}^{ME}(\boldsymbol{s}) - \mu_{i-1,j}^{ME}(\boldsymbol{s}).$$

考虑与排放有关的年度变化

$$\mu_{ij}^E(\boldsymbol{s}) - \mu_{i-1,j}^E(\boldsymbol{s}) = \int\int \tilde{m}_{j,2}(\boldsymbol{u})\{g_{ij}(\boldsymbol{u}, \boldsymbol{s} \mid \boldsymbol{x}) - g_{i-1,j}(\boldsymbol{u}, \boldsymbol{s} \mid \boldsymbol{x})\} f_{\cdot j}(\boldsymbol{x}) d\boldsymbol{u} d\boldsymbol{x}, \tag{3.6}$$

其中 $g_{ij}(\boldsymbol{u}, \boldsymbol{s} \mid \boldsymbol{x})$ 是给定 $\boldsymbol{X}_{ijt}(\boldsymbol{s}) = \boldsymbol{x}$ 下 $\boldsymbol{U}_{ijt}(\boldsymbol{s})$ 的条件密度. 同样, 交互效应的年度变化为

$$\mu_{ij}^{ME}(\boldsymbol{s}) - \mu_{i-1,j}^{ME}(\boldsymbol{s}) = \int\int \tilde{m}_{j,3}(\boldsymbol{x}, \boldsymbol{u})\{g_{ij}(\boldsymbol{u}, \boldsymbol{s} \mid \boldsymbol{x}) - g_{i-1,j}(\boldsymbol{u}, \boldsymbol{s} \mid \boldsymbol{x})\} f_{\cdot j}(\boldsymbol{x}) d\boldsymbol{u} d\boldsymbol{x}. \tag{3.7}$$

注意到, (3.6) 和 (3.7) 所描述的年际变化均由条件分布的差异 $g_{ij}(\boldsymbol{u}, \boldsymbol{s} \mid \boldsymbol{x}) - g_{i-1,j}(\boldsymbol{u}, \boldsymbol{s} \mid \boldsymbol{x})$ 决定. 这只有通过使用时空均衡的气象分布 $f_{\cdot j}(\boldsymbol{x})$ 才能实现. 否则, 所定义的均值浓度的差异不能由排放的年际差异所解释, 因为此时仍然存在着气象的混杂因素.

同理, 如果比较调整均值 $\mu_{ij}(\boldsymbol{s}_1)$ 和 $\mu_{ij}(\boldsymbol{s}_2)$ 时, 可以进行相同的分析, 并且由于气象变量已在空间上进行了均衡化, 因此, 我们可以将差异归因于两个位置的排放量的差异.

通过时空调整的均值 $\mu_{ij}(\boldsymbol{s})$, 我们可以构建区域 $\mathcal{A}$ 中的平均污染物浓度为

$$\mu_{ij}(\mathcal{A}) = |\mathcal{A}|^{-1} \sum_{\boldsymbol{s} \in \mathcal{A}} \mu_{ij}(\boldsymbol{s}), \tag{3.8}$$

其中 $|\mathcal{A}|$ 表示 $\mathcal{A}$ 中空气质量监测点的数量. 这个地区空气质量度量的表达式 $\mu_{ij}(\mathcal{A})$ 是 $\mu_{ij}(\boldsymbol{s})$ 在 $\mathcal{A}$ 地区空气质量监测站的简单平均, 这与中国空气质量管理方式是一致的. 这一点类似于固定抽样调查, 这里考虑的是给定了这些监测站点的地理位置. 实际上, 监测站点的空间分布可能不是均匀的, 在一些区域 (如北京北部) 密度高于其他地区 (北京南部). 比起监测站较多的地区, 监测站点较少的地区的结果会有更高的变异性, 但是其他方面是一样的.

为了适应站点的不均匀分布, 我们引入加权函数 $w_{\mathcal{A}}(\boldsymbol{s})$ 来得到该地区均值的加权表达式

$$\mu_{ij}^w(\mathcal{A}) = |\mathcal{A}|^{-1} \sum_{\boldsymbol{s} \in \mathcal{A}} \mu_{ij}(\boldsymbol{s}) w_{\mathcal{A}}(\boldsymbol{s}),$$

其中 $w_{\mathcal{A}}(\boldsymbol{s})$ 可以重新分配权重以获得空间上的均衡. 为了方便起见, 我们在理论和实际分析中均采用 (3.8) 来计算 $\mu_{ij}(\mathcal{A})$.

4 估计和理论性质

要得到 (3.5) 中的 $\mu_{ij}(\boldsymbol{s})$ 和 (3.8) 中的 $\mu_{ij}^{w}(\mathcal{A})$, 关键是估计 $m_{ij}(\boldsymbol{x},\boldsymbol{s})$. 本文采用非参数核方法[13,15]估计回归函数 $m_{ij}(\boldsymbol{x},\boldsymbol{s})$.

记 $\boldsymbol{X}_{ijt}(\boldsymbol{s})=\{\boldsymbol{Z}_{ijt}(\boldsymbol{s})^{\mathrm{T}},W_{ijt}(\boldsymbol{s})\}^{\mathrm{T}}$, 其中 $W_{ijt}(\boldsymbol{s})$ 是类别风向, $\boldsymbol{Z}_{ijt}(\boldsymbol{s})$ 是 d 维的剩余连续协变量. 记 $K(\cdot)$ 为 d 维对称核函数 (有关详细信息, 参见附录). 定义

$$K_{\boldsymbol{H}}(\boldsymbol{z})=(h_1h_2\cdots h_d)^{-1}K\bigg(\frac{z_1}{h_1},\frac{z_2}{h_2},\ldots,\frac{z_d}{h_d}\bigg),$$

其中 $\boldsymbol{z}=(z_1,z_2,\ldots,z_d)^{\mathrm{T}}$, $\boldsymbol{H}=(h_1,h_2,\ldots,h_d)^{\mathrm{T}}$ 是窗宽向量. 利用风向 w 下 $\boldsymbol{s}$ 站点上年份 i 的季节 j 的数据, 可以得到 $m_{ij}(\boldsymbol{x},\boldsymbol{s})$ 的核估计量 (参见文献 [13])

$$\hat{m}_{ij}(\boldsymbol{z},w;\boldsymbol{s})=\frac{\sum_{t=1}^{n_{ij}}K_{\boldsymbol{H}}\{\boldsymbol{z}-\boldsymbol{Z}_{ijt}(\boldsymbol{s})\}Y_{ijt}(\boldsymbol{s})I\{W_{ijt}(\boldsymbol{s})=w\}}{\sum_{t=1}^{n_{ij}}K_{\boldsymbol{H}}\{\boldsymbol{z}-\boldsymbol{Z}_{ijt}(\boldsymbol{s})\}I\{W_{ijt}(\boldsymbol{s})=w\}},\tag{4.1}$$

其中 n_{ij} 是样本量, $I(\cdot)$ 是示性函数, $W_{ijt}(\boldsymbol{s})$ 等于 1、2、3、4 和 5 对应于风向 CV、NE、NW、SE 和 SW.

对于给定的风向, 平滑窗宽是通过交叉验证的方法 (参见文献 [13,15]) 决定的. 在某些季节, 当某风向下的样本量较少时, 会将该风向与对污染有类似影响的另一个风向的数据合并. 例如, 3 个污染增强风向 SW、CV 和 SE 可以组合, 两个减少污染的方向 NW 和 NE 也可以组合. 本文的目的是给出关于 $\mu_{ij}(\boldsymbol{s})$ 和 $\mu_{ij}(\mathcal{A})$ 的理论性质, 而它们是 $\hat{m}_{ij}(\boldsymbol{x},\boldsymbol{s})$ 的积分形式. 根据定理 4.1 和 4.2 可知, $\mu_{ij}(\boldsymbol{s})$ 和 $\mu_{ij}(\mathcal{A})$ 的估计量是 $\sqrt{n}$ 收敛的. 这意味着它们对平滑窗宽的敏感性不如估计值 $\hat{m}_{ij}(\boldsymbol{x},\boldsymbol{s})$.

对于 $\boldsymbol{x}=(\boldsymbol{z}^{\mathrm{T}},w)^{\mathrm{T}}$, 记 $\hat{F}_{\cdot j}(\boldsymbol{x})$ 为与以 $f_{\cdot j}(\boldsymbol{x})$ 为密度的分布 $F_{\cdot j}(\boldsymbol{x})$ 相对应的经验分布函数. 它可以通过所有站点 A_j 年的第 j 季节的数据进行估计. 根据大数定律, 我们没有必要构造 $\hat{F}_{\cdot j}(\boldsymbol{x})$ 的具体形式, 因此, 可以得到 $\mu_{ij}(\boldsymbol{s})$ 的估计是

$$\begin{aligned}\hat{\mu}_{ij}(\boldsymbol{s})&=\int\hat{m}_{ij}(\boldsymbol{x},\boldsymbol{s})d\hat{F}_{\cdot j}(\boldsymbol{x})\\&=S^{-1}\bigg(\sum_{a=1}^{A_j}n_{aj}\bigg)^{-1}\sum_{w=1}^{5}\sum_{\boldsymbol{s}'\in\mathcal{W}}\sum_{a=1}^{A_j}\sum_{t=1}^{n_{aj}}\hat{m}_{ij}\{\boldsymbol{Z}_{ajt}(\boldsymbol{s}'),w,\boldsymbol{s}\}I\{W_{ajt}(\boldsymbol{s}')=w\},\end{aligned}\tag{4.2}$$

其中 (4.1) 中给出了 $\hat{m}_{ij}(\boldsymbol{x},\boldsymbol{s})$ 的具体形式. 由此, 区域平均值 $\mu_{ij}(\mathcal{A})$ 的估计为

$$\hat{\mu}_{ij}(\mathcal{A})=|\mathcal{A}|^{-1}\sum_{\boldsymbol{s}\in\mathcal{A}}\hat{\mu}_{ij}(\boldsymbol{s}).\tag{4.3}$$

我们可以将上述框架扩展到对污染物的分布进行气象调整, 由此, 可以得到调整后的分位数, 以提供有关极端污染浓度水平的信息. 类似于 (3.5) 中的调整均值, 我们定义站点 $\boldsymbol{s}$ 中 i 年份的 j 季节调整分布为

$$G_{ij}(y,\boldsymbol{s})=\sum_{w=1}^{5}\int F_{ij}(y,\boldsymbol{s}\mid\boldsymbol{z},w)f_{\cdot j}(\boldsymbol{z},w)d\boldsymbol{z},$$

其中 $F_{ij}(y,\boldsymbol{s}\mid\boldsymbol{z},w)=\mathrm{P}\{Y_{ijt}(\boldsymbol{s})\leqslant y\mid\boldsymbol{Z}_{ijt}(\boldsymbol{s})=\boldsymbol{z},W_{ijt}(\boldsymbol{s})=w\}$ 是条件分布. 与 (4.2) 类似, $G_{ij}(y,\boldsymbol{s})$ 的估计量为

$$\hat{G}_{ij}(y,\boldsymbol{s})=S^{-1}\bigg(\sum_{a=1}^{A_j}n_{aj}\bigg)^{-1}\sum_{w=1}^{5}\sum_{\boldsymbol{s}'\in\mathcal{W}}\sum_{a=1}^{A_j}\sum_{t=1}^{n_{aj}}\hat{F}_{ij}\{y,\boldsymbol{s}\mid\boldsymbol{Z}_{ajt}(\boldsymbol{s}'),w\}I\{W_{ajt}(\boldsymbol{s}')=w\},\tag{4.4}$$

其中 $\hat{F}_{ij}(y,\boldsymbol{s}\mid\boldsymbol{z},w)$ 是 $F_{ij}(y,\boldsymbol{s}\mid\boldsymbol{z},w)$ 的核估计量:

$$\hat{F}_{ij}(y,\boldsymbol{s}\mid\boldsymbol{z},w)=\frac{\sum_{t=1}^{n_{ij}}K_{\boldsymbol{H}}\{\boldsymbol{z}-\boldsymbol{Z}_{ijt}(\boldsymbol{s})\}R_{h_0}\{Y_{ijt}(\boldsymbol{s})-y\}I\{W_{ijt}(\boldsymbol{s})=w\}}{\sum_{t=1}^{n_{ij}}K_{\boldsymbol{H}}\{\boldsymbol{z}-\boldsymbol{Z}_{ijt}(\boldsymbol{s})\}I\{W_{ijt}(\boldsymbol{s})=w\}},\tag{4.5}$$

这里 $R_{h_0}(y)=\int_0^{y/h_0}k(u)du$ 是单变量核 $k(\cdot)$ 的积分, 而 h_0 是相应的窗宽. 对于任意 $q\in(0,1)$, 调整后的 q 百分位数可以由调整后分布的逆 (即 $\hat{G}_{ij}^{-1}(q,\boldsymbol{s})$) 得到.

在本节的剩余部分, 我们将给出估计量 $\hat{\mu}_{ij}(\boldsymbol{s})$ 和 $\hat{\mu}_{ij}(\mathcal{A})$ 的渐近性质. 为了简化表示, 本文只考虑 $\boldsymbol{X}_{ijt}(\boldsymbol{s})$ 的协变量是连续的情形, 这也归结到对每个风向下的调整均值进行研究. 而汇总的结果 (4.2) 和 (4.3) 则可以通过综合各风向的结果来得到.

注意到, 本文前面已经给出 $\mathbb{X}_{ijt}=\{\boldsymbol{X}_{ijt}(\boldsymbol{s}_1)^{\mathrm{T}},\ldots,\boldsymbol{X}_{ijt}(\boldsymbol{s}_L)^{\mathrm{T}}\}^{\mathrm{T}}$ 和 $\boldsymbol{e}_{ijt}=\{e_{ijt}(\boldsymbol{s}_1),\ldots,e_{ijt}(\boldsymbol{s}_L)\}^{\mathrm{T}}$. 类似地, 定义 $\mathbb{U}_{ijt}=\{\boldsymbol{U}_{ijt}(\boldsymbol{s}_1)^{\mathrm{T}},\ldots,\boldsymbol{U}_{ijt}(\boldsymbol{s}_L)^{\mathrm{T}}\}^{\mathrm{T}}$. 在假设 $\mathbb{X}_{ijt}$ 和 $\boldsymbol{e}_{ijt}$ 时间平稳的条件下 (详见附录), 定义 $q_{ij,t-t'}(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{s},\boldsymbol{s}')$ 为 $\boldsymbol{X}_{ijt}(\boldsymbol{s})$ 和 $\boldsymbol{X}_{ijt'}(\boldsymbol{s}')$ 的联合密度, $\rho(|t-t'|;\boldsymbol{s},\boldsymbol{s}')=\mathrm{E}\{e_{ijt}(\boldsymbol{s})e_{ijt'}(\boldsymbol{s}')\mid\mathcal{F}_{ij}\}$ 并且

$$C^a_{i_1i_2,j,t_1-t_2}(\boldsymbol{s}_1',\boldsymbol{s}_2';\boldsymbol{s}_1,\boldsymbol{s}_2)=\mathrm{Cov}[m_{i_1j}\{\boldsymbol{X}_{ajt_1}(\boldsymbol{s}_1'),\boldsymbol{s}_1\},m_{i_2j}\{\boldsymbol{X}_{ajt_2}(\boldsymbol{s}_2'),\boldsymbol{s}_2\}].$$

本节中定理所需要的假设以及它们的证明都将在附录中给出, 其中的主要假设是在给定 i 和 j 时, (i) 排放量 $\{\mathbb{U}_{ijt}\}_{t=1}^{n_{ij}}$ 是同分布的; (ii) 天气变量 $\{\mathbb{X}_{ijt}\}_{t=1}^{n_{ij}}$ 和标准化残差 $\{\boldsymbol{e}_{ijt}\}_{t=1}^{n_{ij}}$ 在时间上都是严平稳且 α 混合的, 但在空间上不一定是平稳的, 以便处理更一般的空间相关性. 在这些正则条件下, 定义

$$\gamma_{ij}(\boldsymbol{s}_1,\boldsymbol{s}_2)=\sum_{k=-\infty}^{\infty}\rho(|k|;\boldsymbol{s}_1,\boldsymbol{s}_2)\iint\sigma_{ij}(\boldsymbol{x}_1,\boldsymbol{s}_1)\sigma_{ij}(\boldsymbol{x}_2,\boldsymbol{s}_2)\frac{q_{ij,k}(\boldsymbol{x}_1,\boldsymbol{x}_2;\boldsymbol{s}_1,\boldsymbol{s}_2)}{f_{ij}(\boldsymbol{x}_1,\boldsymbol{s}_1)f_{ij}(\boldsymbol{x}_2,\boldsymbol{s}_2)}dF_{\cdot j}(\boldsymbol{x}_1)dF_{\cdot j}(\boldsymbol{x}_2),$$

$$\lambda_{i_1i_2,j}(\boldsymbol{s}_1,\boldsymbol{s}_2)=S^{-2}A_j^{-2}\sum_{a=1}^{A_j}\sum_{\boldsymbol{s}_1',\boldsymbol{s}_2'\in\mathcal{W}}\sum_{k=-\infty}^{\infty}C^a_{i_1i_2,j,k}(\boldsymbol{s}_1',\boldsymbol{s}_2';\boldsymbol{s}_1,\boldsymbol{s}_2),$$

以及它们对应的区域形式

$$\gamma_{ij}(\mathcal{A},\mathcal{B})=|\mathcal{A}|^{-1}|\mathcal{B}|^{-1}\sum_{\boldsymbol{s}_1\in\mathcal{A},\boldsymbol{s}_2\in\mathcal{B}}\gamma_{ij}(\boldsymbol{s}_1,\boldsymbol{s}_2)\quad\text{和}\quad\lambda_{i_1i_2,j}(\mathcal{A},\mathcal{B})=|\mathcal{A}|^{-1}|\mathcal{B}|^{-1}\sum_{\boldsymbol{s}_1\in\mathcal{A},\boldsymbol{s}_2\in\mathcal{B}}\lambda_{i_1i_2,j}(\boldsymbol{s}_1,\boldsymbol{s}_2).$$

定理 4.1 在附录给出的假设 A.1–A.9 下, 当 $n_{ij}\to\infty$ 时,

$$\sqrt{n_{ij}}\{\hat{\mu}_{ij}(\boldsymbol{s})-\mu_{ij}(\boldsymbol{s})\}\xrightarrow{d}N(0,\tilde{\sigma}_{ij}^2(\boldsymbol{s},\boldsymbol{s}))\quad\text{和}\quad\sqrt{n_{ij}}\{\hat{\mu}_{ij}(\mathcal{A})-\mu_{ij}(\mathcal{A})\}\xrightarrow{d}N(0,\tilde{\sigma}_{ij}^2(\mathcal{A},\mathcal{A})),$$

其中 $\tilde{\sigma}_{ij}^2(\boldsymbol{s},\boldsymbol{s})=\gamma_{ij}(\boldsymbol{s},\boldsymbol{s})+\lambda_{ii,j}(\boldsymbol{s},\boldsymbol{s})$ 和 $\tilde{\sigma}_{ij}^2(\mathcal{A},\mathcal{A})=\gamma_{ij}(\mathcal{A},\mathcal{A})+\lambda_{ii,j}(\mathcal{A},\mathcal{A})$.

注意到, 定理 4.1 中不存在通常与核估计相关的偏差, 这是由于附录中的假设 A.8 中采用了欠平滑. 为了比较 i_1 和 i_2 两年间或 $\mathcal{A}$ 和 $\mathcal{B}$ 两个不同地区间的调整均值, 需要分别得到 $\hat{\mu}_{i_2j}(\mathcal{A})-\hat{\mu}_{i_1j}(\mathcal{A})$ 和 $\hat{\mu}_{i_2j}(\mathcal{B})-\hat{\mu}_{i_1j}(\mathcal{B})$ 的渐近分布. 下面的定理给出了所需的结果. 定义

$$\phi_{i_1i_2,j}(\boldsymbol{s}_1,\boldsymbol{s}_2)=\lambda_{i_1i_1,j}(\boldsymbol{s}_1,\boldsymbol{s}_2)+\lambda_{i_2i_2,j}(\boldsymbol{s}_1,\boldsymbol{s}_2)-\lambda_{i_1i_2,j}(\boldsymbol{s}_1,\boldsymbol{s}_2)-\lambda_{i_2i_1,j}(\boldsymbol{s}_1,\boldsymbol{s}_2),$$
$$\phi_{i_1i_2,j}(\mathcal{A},\mathcal{A})=|\mathcal{A}|^{-2}\sum_{\boldsymbol{s}_1,\boldsymbol{s}_2\in\mathcal{A}}\phi_{i_1i_2,j}(\boldsymbol{s}_1,\boldsymbol{s}_2).$$

定理 4.2 在附录的假设 A.1–A.9 下, (i) 对于 $i_1\neq i_2$, 当 $n_{i_1j},n_{i_2j}\to\infty$ 时,

$$\sqrt{n_{i_1j}}[\{\hat{\mu}_{i_2j}(\mathcal{A})-\hat{\mu}_{i_1j}(\mathcal{A})\}-\{\mu_{i_2j}(\mathcal{A})-\mu_{i_1j}(\mathcal{A})\}]\xrightarrow{d}N(0,\tilde{\sigma}^2_{i_2i_1,j}(\mathcal{A})),$$

其中 $\tilde{\sigma}^2_{i_2i_1,j}(\mathcal{A}) = \sum_{p=1}^{2}\gamma_{i_pj}(\mathcal{A},\mathcal{A}) + \phi_{i_1i_2,j}(\mathcal{A},\mathcal{A})$; 并且 (ii) 当 $n_{ij}\to\infty$ 时,

$$\sqrt{n_{ij}}[\{\hat{\mu}_{ij}(\mathcal{A}) - \hat{\mu}_{ij}(\mathcal{B})\} - \{\mu_{ij}(\mathcal{A}) - \mu_{ij}(\mathcal{B})\}] \xrightarrow{d} N(0, \tilde{\sigma}^2_{ij}(\mathcal{A}-\mathcal{B})),$$

其中 $\tilde{\sigma}^2_{ij}(\mathcal{A}-\mathcal{B}) = \tilde{\sigma}^2_{ij}(\mathcal{A},\mathcal{A}) - 2\tilde{\sigma}^2_{ij}(\mathcal{A},\mathcal{B}) + \tilde{\sigma}^2_{ij}(\mathcal{B},\mathcal{B})$, $\tilde{\sigma}^2_{ij}(\mathcal{A},\mathcal{B}) = \gamma_{ij}(\mathcal{A},\mathcal{B}) + \lambda_{ii,j}(\mathcal{A},\mathcal{B})$.

利用定理 4.2 中的渐近正态性, 在第 6 节实证研究中, 我们可以得到估计的调整均值在时空上差异的显著性.

5 方差估计和假设检验

注意到, 定理 4.2 中的渐近方差比较复杂, 本文通过自助法 (bootstrap) 来得到它们的估计. 首先, 可以证明当 $n_{ij}\to\infty$ 时,

$$\hat{\mu}_{ij}(\boldsymbol{s}) - \mu_{ij}(\boldsymbol{s}) = T_{ij,1}(\boldsymbol{s}) + T_{ij,2}(\boldsymbol{s}) + o_p(n_{ij}^{-1/2}),$$

其中决定渐近方差的两个主要项是

$$\begin{aligned}
T_{ij,1}(\boldsymbol{s}) &= \int \{\hat{m}_{ij}(\boldsymbol{x},\boldsymbol{s}) - m_{ij}(\boldsymbol{x},\boldsymbol{s})\} dF_{\cdot j}(\boldsymbol{x}),\\
T_{ij,2}(\boldsymbol{s}) &= \int m_{ij}(\boldsymbol{x},\boldsymbol{s}) d\{\hat{F}_{\cdot j}(\boldsymbol{x}) - F_{\cdot j}(\boldsymbol{x})\}\\
&= S^{-1}\sum_{\boldsymbol{s}'\in\mathcal{W}}\sum_{a=1}^{n_{\cdot j}} n_{aj}^{-1}\sum_{t=1}^{n_{aj}}\left[n_{aj}\left(\sum_{a=1}^{A_j} n_{aj}\right)^{-1} m_{ij}\{\boldsymbol{X}_{ajt}(\boldsymbol{s}'),\boldsymbol{s}\} - A_j^{-1}\int m_{ij}(\boldsymbol{x},\boldsymbol{s}) dF_{aj}(\boldsymbol{x},\boldsymbol{s}')\right],
\end{aligned}$$

这里 $F_{\cdot j}(\boldsymbol{x})$ 和 $F_{aj}(\boldsymbol{x},\boldsymbol{s}')$ 是 $f_{\cdot j}(\boldsymbol{x})$ 和 $f_{aj}(\boldsymbol{x},\boldsymbol{s}')$ 分别对应的分布.

$T_{ij,1}(\boldsymbol{s})$ 和 $T_{ij,2}(\boldsymbol{s})$ 的形式启发我们可以分别对气象向量 $\boldsymbol{X}_{ijt}(\boldsymbol{s})$ 采用在时间上的分块自助法 (block bootstrap) [16] 和对 (2.2) 中的标准化残差序列 $e_{ijt}(\boldsymbol{s})$ 采用广义自助法 (wild bootstrap) [17,18], 以获得抽样数据. 为了保持残差的空间相关性, 我们对残差向量 $\{\boldsymbol{e}_{ijt}\}_{t=1}^{n_{ij}}$ 进行了重抽样. 这里将时间和空间自助法分离的原因是, 由于非参数核估计量的白化 (whitening) 作用, 时间相关性对 $T_{ij,1}(\boldsymbol{s})$ 的方差贡献可以忽略不计 (参见文献 [19]). 下面将具体介绍抽样的过程.

对于 $T_{ij,2}(\boldsymbol{s})$ 所刻画的时间相关性, 我们用时间上的分块自助法来进行抽样. 为此, 我们将所有站点第 i 年第 j 季的气象数据收集到一起, 记为 $\boldsymbol{\psi}_{ijt} = \{\boldsymbol{X}_{ijt}(\boldsymbol{s}_1)^{\mathrm{T}},\ldots,\boldsymbol{X}_{ijt}(\boldsymbol{s}_L)^{\mathrm{T}}\}^{\mathrm{T}}$, $t=1,\ldots,n_{ij}$. 定义 $\boldsymbol{B}_1 = (\boldsymbol{\psi}_{ij1}^{\mathrm{T}},\ldots,\boldsymbol{\psi}_{ijl}^{\mathrm{T}})^{\mathrm{T}},\ldots,\boldsymbol{B}_{n_{ij}-l+1} = (\boldsymbol{\psi}_{ij,n_{ij}-l+1}^{\mathrm{T}},\ldots,\boldsymbol{\psi}_{ij,n_{ij}}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{B}_{n_{ij}-l+2} = (\boldsymbol{\psi}_{ij,n_{ij}-l+2}^{\mathrm{T}},\ldots,\boldsymbol{\psi}_{ij,n_{ij}}^{\mathrm{T}}, \boldsymbol{\psi}_{ij,1}^{\mathrm{T}})^{\mathrm{T}},\ldots,\boldsymbol{B}_{n_{ij}} = (\boldsymbol{\psi}_{ij,n_{ij}}^{\mathrm{T}},\boldsymbol{\psi}_{ij,1}^{\mathrm{T}},\ldots,\boldsymbol{\psi}_{ij,l-1}^{\mathrm{T}})^{\mathrm{T}}$ 作为一系列长度为 l 的头尾相连的移动分块 [20], 这里根据以往经验, 选择 $l=12$ (小时). 我们从 $\{\boldsymbol{B}_t\}_{t=1}^{n_{ij}}$ 中随机抽取 n_{ij}/l 块并将它们组合起来, 从而得到第 i 年第 j 季的第 b 次重抽样气象序列 $\{\boldsymbol{X}_{ijt}^{*b}(\boldsymbol{s}_1)^{\mathrm{T}},\ldots,\boldsymbol{X}_{ijt}^{*b}(\boldsymbol{s}_L)^{\mathrm{T}}\}^{\mathrm{T}}$, $t=1,\ldots,n_{ij}$.

为了生成响应变量 $Y_{ijt}^{*}(\boldsymbol{s})$ 的重抽样样本, 还需要对标准化残差进行重抽样, 其核心思想是广义自助法 [17] 或回归自助法 (regression bootstrap) [19]. 给定 (4.1) 中估计的回归函数 $\hat{m}_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{s}\}$, 我们可以得到 $\hat{\epsilon}^2_{ijt}(\boldsymbol{s}) = [Y_{ijt}(\boldsymbol{s}) - \hat{m}_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{s}\}]^2$, 从而对于 $\boldsymbol{x} = (\boldsymbol{z}^{\mathrm{T}},w)^{\mathrm{T}}$, 我们可以利用核平滑估计其条件方差 $\sigma^2_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{s}\}$,

$$\hat{\sigma}^2_{ij}(\boldsymbol{x},\boldsymbol{s}) = \frac{\sum_{t=1}^{n_{ij}} K_{\tilde{\boldsymbol{H}}}\{\boldsymbol{z} - \boldsymbol{Z}_{ijt}(\boldsymbol{s})\}\hat{\epsilon}^2_{ijt}(\boldsymbol{s}) I\{W_{ijt}(\boldsymbol{s}) = w\}}{\sum_{t=1}^{n_{ij}} K_{\tilde{\boldsymbol{H}}}\{\boldsymbol{z} - \boldsymbol{Z}_{ijt}(\boldsymbol{s})\} I\{W_{ijt}(\boldsymbol{s}) = w\}}. \tag{5.1}$$

与之前一样, 这里的窗宽也是通过交叉验证得到的. 由此, 我们得出标准化残差的估计量为

$$\hat{e}_{ijt}(\boldsymbol{s}) = \frac{\hat{\epsilon}_{ijt}(\boldsymbol{s})}{\hat{\sigma}_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}), \boldsymbol{s}\}}, \tag{5.2}$$

其中 $\hat{\boldsymbol{e}}_{ijt} = \{\hat{e}_{ijt}(\boldsymbol{s}_1), \ldots, \hat{e}_{ijt}(\boldsymbol{s}_L)\}^{\mathrm{T}}$. 令

$$\hat{\boldsymbol{\Sigma}}_{ij} = n_{ij}^{-1} \sum_{t=1}^{n_{ij}} \hat{\boldsymbol{e}}_{ijt} \hat{\boldsymbol{e}}_{ijt}^{\mathrm{T}} - \left(n_{ij}^{-1} \sum_{t=1}^{n_{ij}} \hat{\boldsymbol{e}}_{ijt} \right) \left(n_{ij}^{-1} \sum_{t=1}^{n_{ij}} \hat{\boldsymbol{e}}_{ijt} \right)^{\mathrm{T}}.$$

我们通过 $\hat{\boldsymbol{e}}_{ijt}^{*b} \overset{\text{iid}}{\sim} N_L(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{ij})$ 生成标准化残差的重抽样, 再结合重抽样的天气过程, 得到重抽样的响应变量, 即当 $t = 1, \ldots, n_{ij}$ 时,

$$Y_{ijt}^{*b}(\boldsymbol{s}) = \hat{m}_{ij}\{\boldsymbol{X}_{ijt}^{*b}(\boldsymbol{s}), \boldsymbol{s}\} + \hat{\sigma}_{ij}\{\boldsymbol{X}_{ijt}^{*b}(\boldsymbol{s}), \boldsymbol{s}\} \hat{\boldsymbol{e}}_{ijt}^{*b}(\boldsymbol{s}). \tag{5.3}$$

由此, 我们重新计算每个自助法重抽样本的调整均值

$$\hat{\mu}_{ij}^{*b}(\boldsymbol{s}) = S^{-1} \left(\sum_{a=1}^{A_j} n_{aj} \right)^{-1} \sum_{w=1}^{5} \sum_{\boldsymbol{s}' \in \mathcal{W}} \sum_{a=1}^{A_j} \sum_{t=1}^{n_{aj}} \hat{m}_{ij}^{b}\{\boldsymbol{Z}_{ajt}^{*b}(\boldsymbol{s}'), w, \boldsymbol{s}\} I\{W_{ajt}^{*b}(\boldsymbol{s}') = w\},$$

$$\hat{\mu}_{ij}^{*b}(\mathcal{A}) = |\mathcal{A}|^{-1} \sum_{\boldsymbol{s} \in \mathcal{A}} \hat{\mu}_{ij}^{*b}(\boldsymbol{s}).$$

通过上述自助法抽样, 可以得到 $\hat{\mu}_{ij}(\boldsymbol{s})$、$\hat{\mu}_{ij}(\mathcal{A})$、$\hat{\mu}_{i_2j}(\mathcal{A}) - \hat{\mu}_{i_1j}(\mathcal{A})$ 和 $\hat{\mu}_{ij}(\mathcal{A}) - \hat{\mu}_{ij}(\mathcal{B})$ 的标准差, 分别记为 $\hat{\tilde{\sigma}}_{ij}(\boldsymbol{s}, \boldsymbol{s})$、$\hat{\tilde{\sigma}}_{ij}(\mathcal{A}, \mathcal{A})$、$\hat{\tilde{\sigma}}_{i_2i_1,j}(\mathcal{A})$ 和 $\hat{\tilde{\sigma}}_{ij}(\mathcal{A} - \mathcal{B})$. 利用这些标准差及调整均值在时空上的差异, 我们可以评估区域内空气质量的变化.

为了评估年际间差异的显著性, 考虑原假设 $H_0 : \mu_{i_2j}(\mathcal{A}) = \mu_{i_1j}(\mathcal{A})$, 备择假设为 $H_1 : \mu_{i_2j}(\mathcal{A}) > (<)\ \mu_{i_1j}(\mathcal{A})$. 我们使用检验统计量 $\{\hat{\mu}_{i_2j}(\mathcal{A}) - \hat{\mu}_{i_1j}(\mathcal{A})\}/\hat{\tilde{\sigma}}_{i_2i_1,j}(\mathcal{A})$ 来对该问题进行检验. 类似地, 为了考虑区域间差异的显著性, 我们对原假设 $H_0 : \mu_{ij}(\mathcal{A}) = \mu_{ij}(\mathcal{B})$ 和备择假设 $H_1 : \mu_{ij}(\mathcal{A}) > (<)\ \mu_{ij}(\mathcal{B})$ 进行检验, 相应的统计量为 $\{\hat{\mu}_{ij}(\mathcal{A}) - \hat{\mu}_{ij}(\mathcal{B})\}/\hat{\tilde{\sigma}}_{ij}(\mathcal{A} - \mathcal{B})$. 根据定理 4.1 和 4.2, 这两个统计量都渐近服从标准正态分布, 从而可以获得统计显著性的 p 值.

6 北京空气污染数据的应用

6.1 模型诊断

首先对 (2.2) 中给出的回归函数进行了非参数核估计, 从而对非参数模型 (4.1) 的拟合效果进行分析. 首先, 表 1 提供了从 2013 年春季到 2016 年冬季的 6 种污染物的拟合度 R^2. 结果表明, 大多数 R^2 在 70% 以上, 这表明模型 (2.2) 对于数据的建模是合理的.

在对空间数据相关性的建模中, 一个重要的概念是半变差函数 (semi-variogram) [21], 我们将通过分析标准化残差的过程对其进行描述. 在时间平稳性的假设下, 第 i 年 j 季节 t 小时标准化残差过程 $\{e_{ijt}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{R}\}$ 的半变差函数是

$$\gamma_{ij}(\boldsymbol{s}, \boldsymbol{s}') = 2^{-1}\mathrm{E}[\{e_{ijt}(\boldsymbol{s}) - e_{ijt}(\boldsymbol{s}')\}^2], \quad \text{对于任意 } \boldsymbol{s}, \boldsymbol{s}' \in \mathcal{R}, \tag{6.1}$$

其中 $\|\cdot\|$ 表示 Euclid 距离. 如果 $\gamma_{ij}(\boldsymbol{s}, \boldsymbol{s}') = \gamma_{ij}(\boldsymbol{s} - \boldsymbol{s}')$, 称过程 $\{e_{ijt}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{R}\}$ 在是空间上平稳的; 如果 $\gamma_{ij}(\boldsymbol{s}, \boldsymbol{s}') = \gamma_{ij}(\|\boldsymbol{s} - \boldsymbol{s}'\|)$, 称其是各向同性的 (isotropic). 其他过程 (如 $\mathrm{PM}_{2.5}$ 和 SO_2 过程) 的半变差函数也可以类似定义.

表 1 模型拟合的 $\mathbf{R^2}$. 站点平均 $\mathbf{R^2}$ 通过将所有站点的模拟拟合的 $\mathbf{R^2}$ 取平均而得到. 站点合并 $\mathbf{R^2}$ 通过先将所有站点的观测值和模型拟合值放在一起构成一个合并的数据集, 再对这个数据集计算一个统一的 $\mathbf{R^2}$ 而得到

污染物	季节	站点平均				站点合并			
		2013	2014	2015	2016	2013	2014	2015	2016
$PM_{2.5}$	春季	0.82	0.79	0.88	0.86	0.82	0.79	0.88	0.86
	夏季	0.72	0.80	0.75	0.77	0.72	0.80	0.75	0.77
	秋季	0.85	0.78	0.87	0.87	0.85	0.79	0.88	0.87
	冬季	0.84	0.74	0.92	0.86	0.84	0.75	0.92	0.87
PM_{10}	春季	0.77	0.76	0.86	0.81	0.77	0.77	0.86	0.82
	夏季	0.64	0.67	0.69	0.68	0.65	0.68	0.69	0.68
	秋季	0.80	0.76	0.86	0.85	0.81	0.76	0.87	0.85
	冬季	0.82	0.71	0.91	0.84	0.82	0.72	0.91	0.85
SO_2	春季	0.74	0.81	0.87	0.84	0.76	0.82	0.87	0.84
	夏季	0.75	0.62	0.65	0.74	0.82	0.72	0.60	0.76
	秋季	0.81	0.70	0.91	0.87	0.83	0.72	0.91	0.89
	冬季	0.78	0.67	0.83	0.78	0.79	0.69	0.83	0.78
NO_2	春季	0.76	0.76	0.82	0.82	0.79	0.81	0.85	0.84
	夏季	0.65	0.55	0.65	0.71	0.76	0.73	0.79	0.80
	秋季	0.82	0.76	0.82	0.83	0.84	0.79	0.84	0.85
	冬季	0.82	0.72	0.89	0.85	0.83	0.74	0.90	0.86
CO	春季	0.78	0.82	0.88	0.86	0.77	0.83	0.88	0.86
	夏季	0.78	0.71	0.74	0.77	0.77	0.71	0.77	0.78
	秋季	0.79	0.75	0.92	0.88	0.79	0.76	0.92	0.88
	冬季	0.80	0.72	0.89	0.82	0.80	0.73	0.89	0.83
8 小时 O_3	春季	0.92	0.89	0.95	0.95	0.94	0.91	0.95	0.95
	夏季	0.87	0.73	0.88	0.86	0.89	0.77	0.89	0.88
	秋季	0.86	0.89	0.96	0.94	0.88	0.90	0.96	0.94
	冬季	0.90	0.79	0.96	0.90	0.90	0.82	0.96	0.91

特别地, $\gamma_{ij}(0)$ 被称作块金效应 (nugget effect). 它表示不能由空间相关性解释的变异性. 块金效应是由测量误差导致的, 并且需要密集分布的站点才能被精确估计. 在各向同性的假设下, 随着 h 增加, $\gamma_{ij}(h)$ 一开始会逐渐增加, 然后在超出一定距离后逐渐平稳, 该距离通常称为范围 (range). 若两个站点的距离大于范围, 则它们在空间上将没有相关性. 半变差函数在这个范围处的值称为门槛值 (sill).

在时间平稳的假设下, 在标准化残差过程中, 任何两个站点 $\boldsymbol{s}_{l_1}$ 和 $\boldsymbol{s}_{l_2}$ 的半变差函数可以通过下式估计:

$$\hat{\gamma}_{ij}(\|\boldsymbol{s}_{l_1}-\boldsymbol{s}_{l_2}\|)=\frac{1}{2n_{ij}}\sum_{t=1}^{n_{ij}}\{\hat{e}_{ijt}(\boldsymbol{s}_{l_1})-\hat{e}_{ijt}(\boldsymbol{s}_{l_2})\}^2,\quad l_1,l_2=1,\ldots,L, \tag{6.2}$$

我们在图 2 和 3 中用点来表示这些半变差函数.

受 Jun 和 Stein [22] 的启发, 为了获得有关大范围空间相关性的信息, 我们在图 3 中展示了 2015 和 2016 年夏季和冬季的 $PM_{2.5}$、SO_2、NO_2 和 8 小时 O_3 原始污染浓度的半变差函数、非参数回归的拟合值和估计的残差 $\hat{\boldsymbol{\epsilon}}_{ijt}=Y_{ijt}(\boldsymbol{s})-\hat{m}_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{s}\}$.

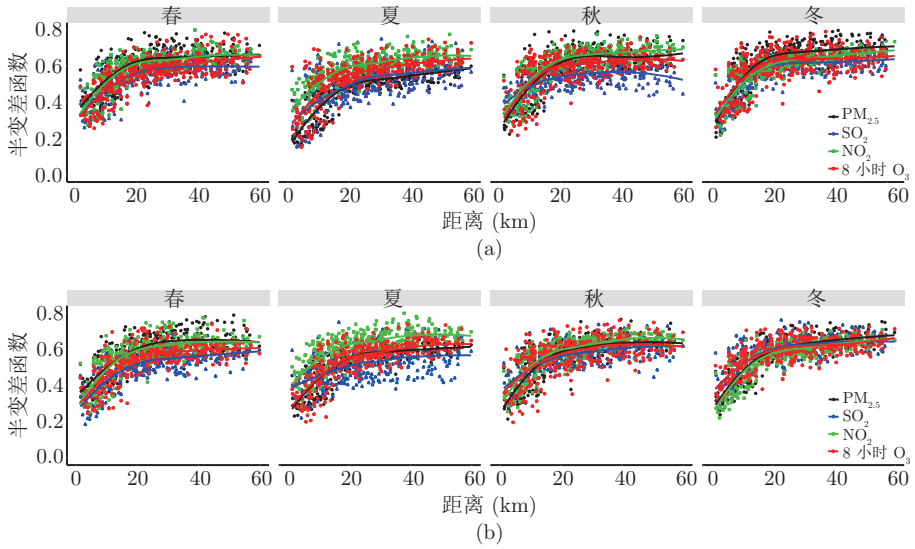


图 2 (网络版彩图) **2015 (a) 和 2016 年 (b) $PM_{2.5}$ (黑色)、SO_2 (蓝色)、NO_2 (绿色) 和 8 小时 O_3 (红色) 的非参数模型标准化残差的半变差函数. 散点表示半变差函数的估计, 曲线是通过非参数 LOESS 方法进行平滑的**

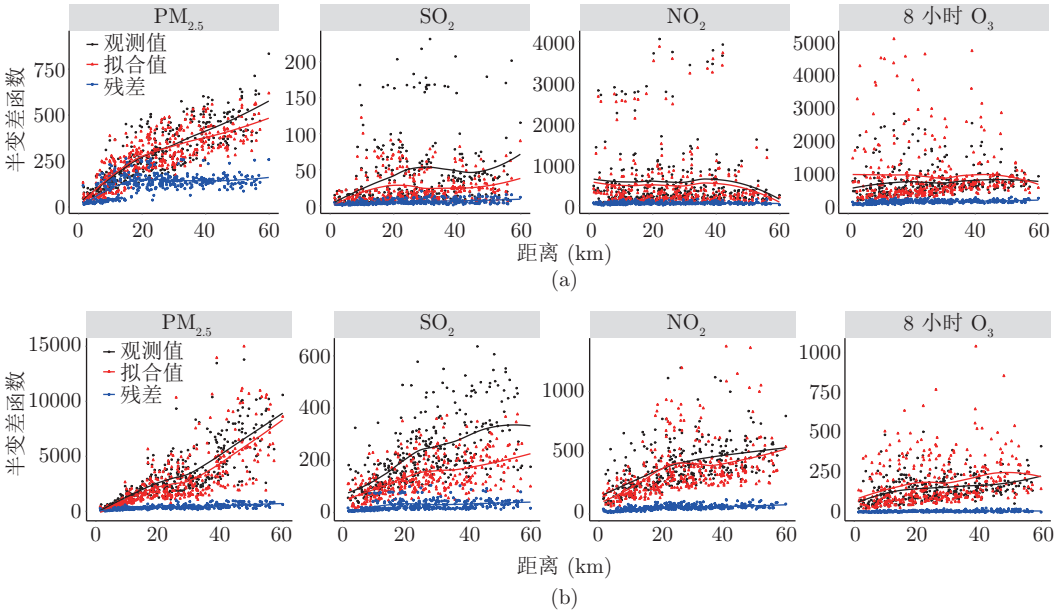


图 3 (网络版彩图) **2015 年夏季 (a) 和冬季 (b) $PM_{2.5}$、SO_2、NO_2 和 8 小时 O_3 的观测值 (黑色)、拟合值 (红色) 和残差 (蓝色) 的半变差函数. 点表示实际估计的半变差, 光滑曲线是通过非参数 LOESS 方法得到的**

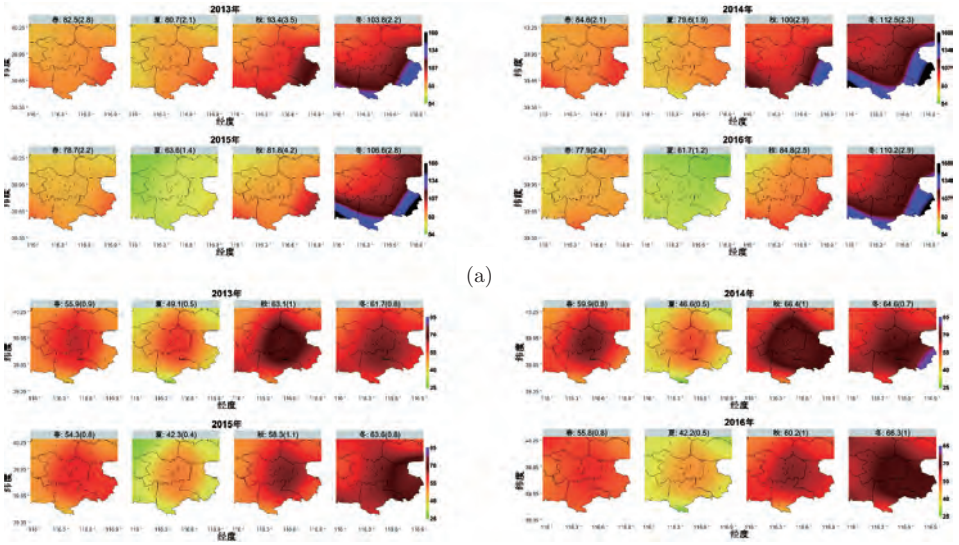
图 3 显示, 原始 $PM_{2.5}$ 和 SO_2 的半变差函数显示出更强的非平稳性和更远范围的相关性, 而原始 NO_2 和 O_3 的半变差函数即使在较大距离下也相对平坦. 后者显示出 NO_2 和 O_3 的空间相关性较弱, 这是由于它们的存活期较短, 而这两种气体都具有更高的化学反应活性, 因此无法远距离传播. 该图还显示, 原始污染物的半变差函数与拟合值的半变差函数相似, 这从空间相关性的角度也表明了核回归方法的合理性. 残差的半变差函数显示出弱得多的相关性, 这也表明了回归模型具有发现大规模趋势和变异的能力.

图 2 对 $PM_{2.5}$、SO_2、NO_2 和 8 小时 O_3 (中午 12 点至晚上 7 点) 的半变差函数 $\hat{\gamma}_{ij}(h)$ 进行了非参数局部加权回归 (locally estimated scatterplot smoothing, LOESS) [23] 平滑. 这些 LOESS 拟合曲线表明, 在大多数图上, 距离超过 20 km 都没有太大的空间相关性, 因为半变差函数在 20 km 之后不再显著增加. 这也表明模型 (2.2) 捕获了空间相关性的主要部分. 图 4 提供了 3 个监测点的 $PM_{2.5}$ 标准化残差的自回归函数及相应的长期协方差函数. 它们表明夏季的时间相关性比其他 3 个季节更强. 这可能是由于北京夏季的静态天气模式所致. 注意到, 图 2 中的这些半变差函数具有类似的形状、大小和块金效应, 表明 4 种污染物的误差过程 $\{e_{ijt}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{R}\}$ 的空间相关性有一些共同特征.

图 4　(网络版彩图) 奥体中心、丰台和顺义站点 2014 至 2016 年 4 个季节 $PM_{2.5}$ 标准化残差的自协方差函数. 每张图标题中的数字表示谱函数在零点的取值乘以 2π, 其等于自协方差函数对所有时间间隔的求和, 因此可以反映每个季节标准化残差序列的长期相关性的大小. (a) 奥体中心, 2014; (b) 奥体中心, 2015; (c) 奥体中心, 2016; (d) 丰台, 2014; (e) 丰台, 2015; (f) 丰台, 2016; (g) 顺义, 2014; (h) 顺义, 2015; (i) 顺义, 2016

6.2 污染地图

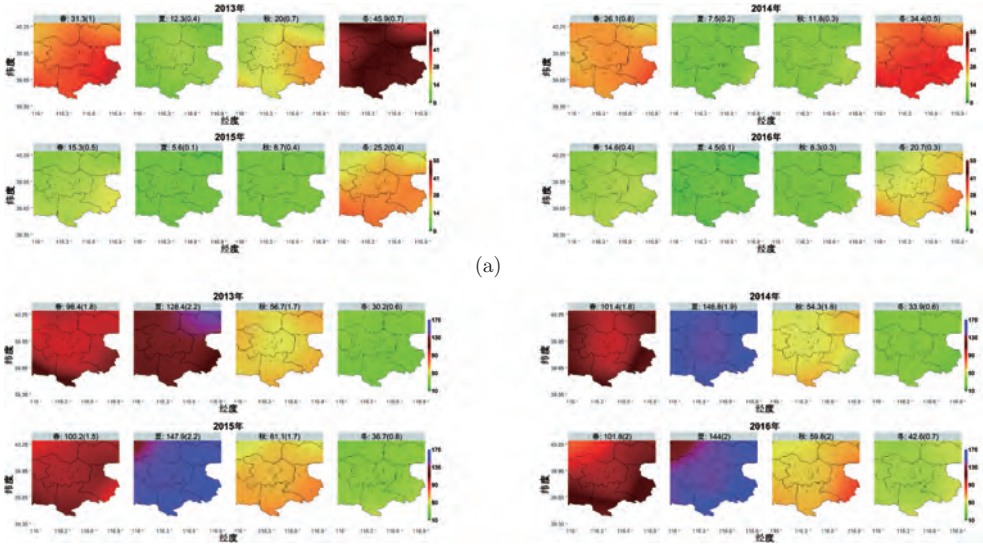
首先采用之前提出的调整方法来得到所有 28 个污染检测站点的每个季节不同年份的调整均值 $\hat{\mu}_{ij}(\boldsymbol{s})$. 然后, 使用二元 Gauss 乘积核, 在研究区域内对调整后均值 $\hat{\mu}_{ij}(\boldsymbol{s})$ 进行空间核平滑, 其中经度和纬度的平滑窗宽均为 $h = 0.15$ 度. 这样就得到了不同年份和季节的空气污染物的污染地图. 图 5 显示了 2013 至 2016 年 $PM_{2.5}$ 和 NO_2 的污染地图, 图 6 提供了 2003 至 2016 年 SO_2 和 8 小时 O_3 的污染地图.



(a)

(b)

图 5 (网络版彩图) **2013** 至 **2016** 年北京地区 $\mathbf{PM_{2.5}}$ **(a)** 和 $\mathbf{NO_2}$ **(b)** 时空调整后的各季节平均浓度 ($\mu g/m^3$) 的污染地图. 每张图上方的数字是北京地区调整后的平均值, 括号内的数字是标准误; 平滑窗宽为 **0.15**



(a)

(b)

图 6 (网络版彩图) **2013** 至 **2016** 年北京地区 $\mathbf{SO_2}$ **(a)** 和 **8** 小时 $\mathbf{O_3}$ **(b)** 时空调整后的各季节平均浓度 ($\mu g/m^3$) 的污染地图. 每张图上方的数字是北京地区调整后的平均值, 括号内的数字是标准误; 平滑窗宽为 **0.15**

图 5 和 6 显示, $PM_{2.5}$、SO_2 和 NO_2 的浓度具有相似的季节性模式, 即冬季高夏季低, 春季和秋季则介于两者之间. 而 8 小时 O_3 具有相反的季节性, 即春夏季浓度高而秋冬季浓度低. 这是因为导致臭氧产生的光化学过程需要来自太阳的紫外线, 因而中午 12 点到下午 7 点 O_3 浓度最高, 而且在春夏季臭氧污染严重.

除这些季节性模式外, $PM_{2.5}$ 还表现出较大的空间变异性, 即北京南部的浓度远高于其他地区, 尤其是在污染严重的冬季. 其他 3 种污染物的空间变异性远小于 $PM_{2.5}$. 图 5 显示北京市中心的 NO_2 呈现隆起的环形, 这在 2014 和 2016 年最为明显. 这主要是由于机动车的 NO 和 NO_2 排放, 特别是在交通拥堵的北京. 环形脊的峰位于城市的东部, 在三环路和四环路之间, 这里也是北京最拥挤的地区. 同时, NO_2 的环形分布也说明, NO_2 由于具有较强的化学活性, 因此其存在时间比较短而无法远距离传输.

对图 6 中臭氧浓度图的仔细观察, 我们发现其低浓度区域位于 NO_2 浓度高的区域, 该现象在 2014 年夏季和春季最为明显. 这是由于 NO_2 和 O_3 之间通过化学反应方程 $NO_2 + O_2 \overset{\text{u.v.}}{\rightleftarrows} NO + O_3$ 进行了相互转化. 由于机动车的排放主要是 NO (和 CO), 因此, 该方程式意味着 NO 的直接排放会消耗 O_3 来生成 NO_2. 当然, 在紫外线 (u.v.) 辐射的条件下, 逆反应也会发生, 这也解释了为什么 O_3 在下午和夏季最高.

图 5 显示了 2014 至 2015 年间 $PM_{2.5}$ 显著减少, 特别是在夏季和秋季. 但是, 我们很难发现 2015 至 2016 年有明显的改变. 同时, 我们发现 2014 至 2016 年所有季节的 SO_2 都有明显减少. 我们将在下一小节讨论这些下降在统计上是否是显著的. 相比之下, 图 5 和 6 中 NO_2 和 O_3 的改善量 (如果有的话) 还不清楚, 这也需要在下一小节进行讨论.

6.3 区域空气质量评估

本节利用第 5 节中描述的假设检验对空气污染水平的年度和区域差异进行统计推断. 这里关注的是区域 $\mathcal{A}$ 的时间差异 $\mu_{(i+1),j}(\mathcal{A}) - \mu_{ij}(\mathcal{A})$, 以及两个区域之间的空间差异 $\mu_{ij}(\mathcal{A}) - \mu_{ij}(\mathcal{B})$.

图 7 显示了 2013 至 2016 年中心和南部区域 $PM_{2.5}$、SO_2、NO_2 和 8 小时 O_3 的各季度平均浓度. 结果表明, 各季度 $PM_{2.5}$ 平均水平持续高于世界卫生组织 (WHO) 设定的过渡期目标限值 35 $\mu g/m^3$. 中心区域所有季节的 NO_2 平均浓度都始终超过 WHO 设定的 40 $\mu g/m^3$ 限值. 而 SO_2 和 O_3 相对较好, 其中 SO_2 仅在最近两年冬季超过 WHO 限值, O_3 仅在最近两年夏季超过 WHO 限值. 南部地区的 $PM_{2.5}$ 就要高得多, 但 NO_2 却比中心区域低. 南部地区的 $PM_{2.5}$ 较高, 反映了重工业发达的河北对北京细颗粒物的传输. 由于北京 70% 以上的人口都居住在中心区域, 交通拥堵导致的机动车排放量大大增加, 这也导致中心区域 NO_2 浓度较高.

为了检验这两个区域 (南部与中心) 之间是否存在统计上显著差异, 表 2 给出了南部与中心区域的平均值之差以及它们的标准误和 p 值. 这里假设检验的原假设是 "南部的 $PM_{2.5}$、SO_2 和 8 小时 O_3 高于中心区域, 而 NO_2 则相反". 从该表可以发现 SO_2 和 8 小时 O_3 的污染存在着 "南高中低" 的模式, 而这一点在图 7 中却不易察觉. 该表报告了 4 类 p 值: 大于 0.01 的值, 标记为 $*$ 的值 (介于 10^{-9} 和 0.01 之间), 标记为 $**$ 的值 (介于 10^{-16} 和 10^{-9} 之间), 标记为 $***$ 的值 (小于 10^{-16}). 如果用 16 个季节中 p 值标记为 2 个或 3 个 $*$ 的数量来衡量差异大小的话, 则南中差异最大的是 NO_2 (15/16), 其次是 $PM_{2.5}$ (11/16) 和 O_3 (9/16). SO_2 的南中差异最小, 同时 SO_2 在 4 种污染物中没有 $*$ 的季节也最多 (3/16).

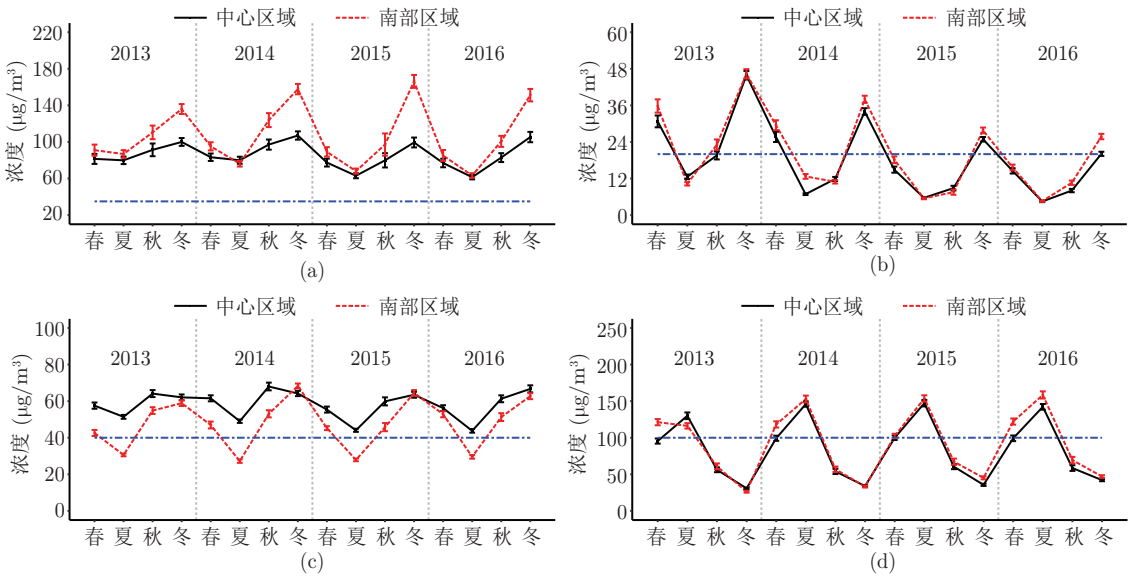


图 7 (网络版彩图) 季节性调整后的均值 ($\mu g/m^3$), 横条表示 95% 的置信区间. 区域均值是通过对该地区所有站点上调整后的均值进行平均而得到的. 蓝色虚线表示 WHO 设定的标准, (a) $PM_{2.5}$、(b) SO_2、(c) NO_2 和 (d) 8 小时 O_3 分别为 35、20、40 和 100 $\mu g/m^3$

表 2 $PM_{2.5}$、SO_2、NO_2 和 8 小时 O_3 在不同季节和年份的区域差异 (南部减中心), 其中 * 的数量表示年际间增加或减少的显著性水平 (*: $10^{-9} \leqslant p$ 值 $< 10^{-2}$; **: $10^{-16} \leqslant p$ 值 $< 10^{-9}$; ***: p 值 $< 10^{-16}$), 括号内的数表示差异的标准误

污染物	季节	2013	2014	2015	2016
$PM_{2.5}$	春季	9.5 (1.2)**	11.9 (1.1)***	11.7 (1.2)***	8.2 (1.6)*
	夏季	6.9 (1.3)*	−3.7 (1.2)*	5.1 (0.8)**	1.8 (1.0)
	秋季	19.1 (2.1)***	26.8 (2.4)***	17.2 (3.3)*	17.7 (1.8)***
	冬季	36.0 (1.7)***	50.8 (1.8)***	66.8 (2.2)***	45.8 (2.1)***
	平均	17.9 (0.8)***	21.4 (0.9)***	25.2 (1.1)***	18.4 (0.8)***
SO_2	春季	5.1 (0.5)***	3.8 (0.5)**	3.2 (0.3)***	1.1 (0.3)*
	夏季	−2.3 (0.3)**	5.8 (0.4)***	−0.2 (0.1)	0.1 (0.1)
	秋季	3.5 (0.6)**	−0.9 (0.2)*	−1.3 (0.3)*	2.6 (0.2)***
	冬季	0.5 (0.6)	4.0 (0.5)***	2.9 (0.4)***	5.7 (0.3)***
	平均	1.7 (0.3)**	3.2 (0.2)***	1.2 (0.2)**	2.4 (0.1)***
NO_2	春季	−15.0 (0.5)***	−14.6 (0.7)***	−10.2 (0.6)***	−3.4 (0.6)**
	夏季	−20.8 (0.4)***	−21.8 (0.4)***	−16.2 (0.3)***	−14.2 (0.4)***
	秋季	−9.3 (0.7)***	−14.9 (0.7)***	−14.1 (0.7)***	−10.0 (0.7)***
	冬季	−3.2 (0.5)**	4.1 (0.4)***	1.0 (0.5)	−4.0 (0.5)**
	平均	−12.1 (0.3)***	−11.8 (0.3)***	−9.9 (0.3)***	−7.9 (0.3)***
8 小时 O_3	春季	25.8 (1.5)***	18.9 (1.0)***	2.5 (0.9)*	22.9 (1.1)***
	夏季	−13.8 (1.6)***	6.9 (1.7)*	6.4 (1.2)*	16.2 (1.5)***
	秋季	4.6 (1.2)*	2.9 (1.0)*	6.9 (1.3)*	10.9 (1.1)***
	冬季	−4.5 (0.5)***	−0.7 (0.6)	9.5 (0.7)***	4.9 (0.6)***
	平均	3.0 (0.6)*	7.0 (0.6)***	6.3 (0.6)***	13.7 (0.6)***

为了得到空气质量的年度变化信息, 我们计算了相邻年份之间调整均值的差异. 同时, 我们采用之前提出的时空自助法来得到其标准误和 p 值. 图 8 中展示了 4 种污染物的相关结果. 注意到, 时间差异并不如表 2 所显示的地域差异显著. 实际上, 中心区域 $PM_{2.5}$ 的 12 个季节年际差异中只有 2 个是显著的, 而南部地区有 7 个是显著的. 这些显著的 p 值基本上处于 $*$ 级别, 这表明变化很小. 在这 4 种污染物中, SO_2 在两个区域中的年际差异是最显著的, 其次是 8 小时 O_3 和 NO_2. 在过去的 4 年中, $PM_{2.5}$ 在 4 种污染物中变化最小.

2014 和 2015 年 SO_2 的减少量非常可观, 这也是过去 4 年北京空气质量管理的一个亮点. 与 SO_2 相比, 过去 4 年其他 3 种污染物并没有持续显著下降. 注意到 2015 年是空气改善最显著的一年, 这体现在 $PM_{2.5}$、SO_2 和 NO_2 的显著减少. 这在很大程度上与上一个经济周期的经济放缓有关. 尽管 2015 年的 $PM_{2.5}$ 以 1% 的显著水平显著降低, 但 2016 年没有改善, 中心区域反而有不显著的增长. 上述现象令人担忧, 因为它表明 2016 年的 $PM_{2.5}$ 没有改善, 我们在 NO_2 和 O_3 中也观察到了类似的现象.

以上分析也揭示了北京空气质量管理面临的挑战. 尽管 SO_2 已大大降低, 但这没有转化成 $PM_{2.5}$ 的持续降低. 我们的分析表明, 当前迫切需要减少由机动车排放引起的 NO_2, 从而为 $PM_{2.5}$ 的下降提供新的驱动力. Chen 等 [3] 研究发现, 在京津冀地区, 减少 NO_2 还可以降低近年来持续上升的 O_3.

我们也将新提出的调整方法与两种已有的方法进行了比较. 一种是第 3 节简要叙述的 Thompson 等 [10] 提出的趋势分析方法, 另一种是美国环境保护署 (EPA) 提倡的 3 年滑动平均方法. 滑动平均方法的细节及存在的问题已经在文献 [3] 中进行了说明. 图 9 展示了采用 3 种方法得到的中心和南部地区 $PM_{2.5}$ 和 SO_2 的平均浓度. NO_2 和 8 小时 O_3 的情形展示在图 10 中. 这些图都显示出我们提出的方法与其他两种方法的明显差异, 尤其是秋冬两季的 $PM_{2.5}$、SO_2 和 NO_2. 同时, 表 3 也展示了新提出的方法与其他两种方法差异的数值平均. 从表中我们发现, 这两个地区冬季 $PM_{2.5}$ 的平均差异分别大于 6 和 10 $\mu g/m^3$, 这也体现了北京在前几个冬季 4 种污染物的显著改进.

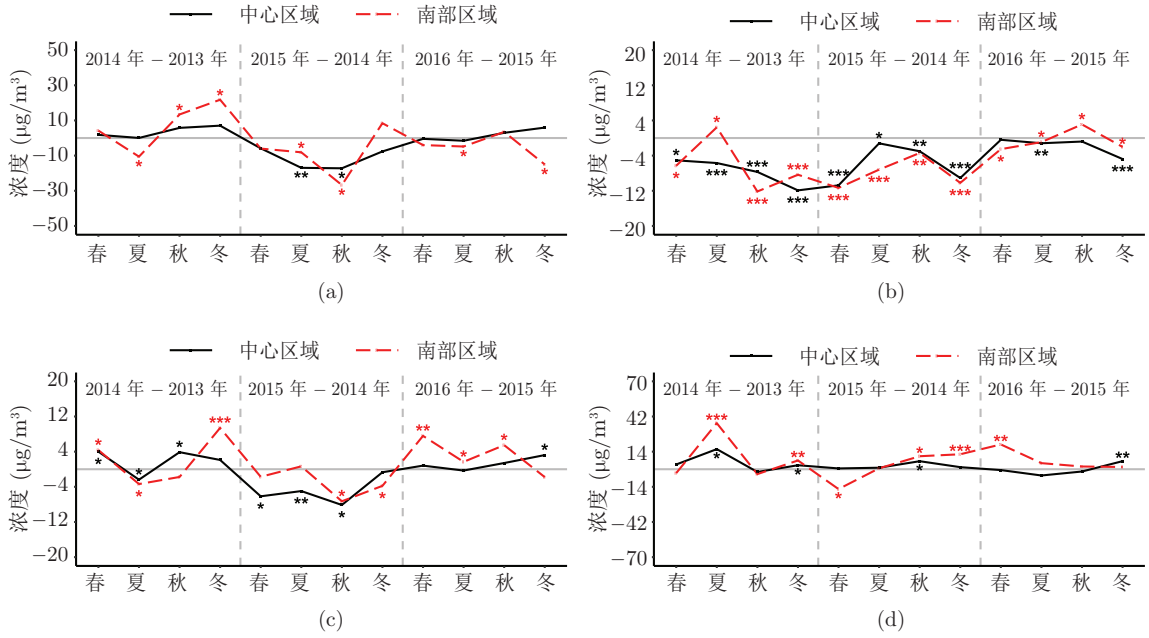


图 8 (网络版彩图) 调整均值的年度差异 ($\mu g/m^3$). (a) $PM_{2.5}$; (b) SO_2; (c) NO_2; (d) 8 小时 O_3, 其中 * 的数量表示年际间增加或减少的显著性水平 (*: $10^{-9} \leqslant p$ 值 $< 10^{-2}$; **: $10^{-16} \leqslant p$ 值 $< 10^{-9}$; ***: p 值 $< 10^{-16}$)

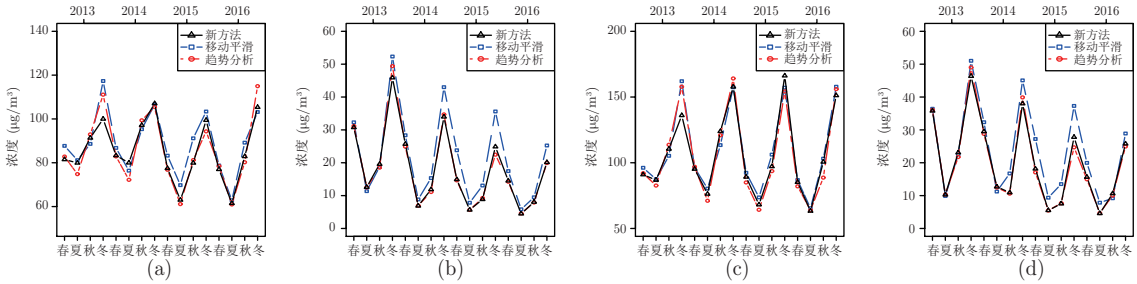


图 9 (网络版彩图) 新提出的方法和已有的移动平滑 (moving average)、趋势分析 (trend analysis) 在中心和南部区域的 $PM_{2.5}$ 和 SO_2 的季节调整均值. (a) 中心区域 $PM_{2.5}$; (b) 中心区域 SO_2; (c) 南部区域 $PM_{2.5}$; (d) 南部区域 SO_2

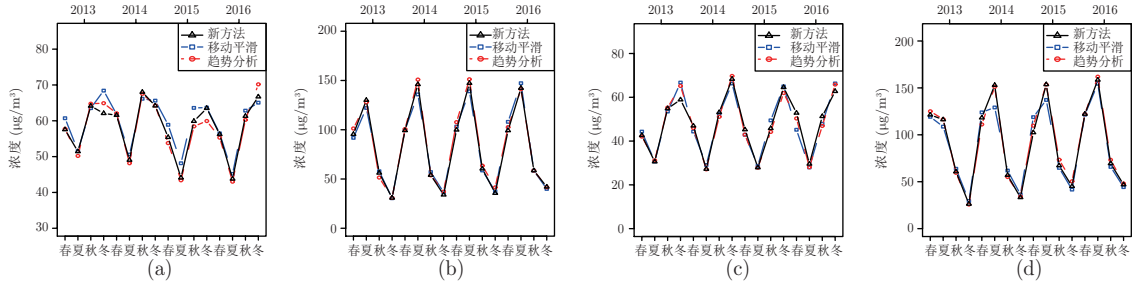


图 10 (网络版彩图) 新提出的方法和已有的移动平滑 (moving average)、趋势分析 (trend analysis) 在南部和中心区域的 NO_2 和 8 小时 O_3 的季节调整均值. (a) 中心区域 NO_2; (b) 中心区域 8 小时 O_3; (c) 南部区域 NO_2; (d) 南部区域 8 小时 O_3

表 3 新提出的方法和已有的移动平滑、趋势分析在中心和南部区域的差异 (标准误, $\mu g/m^3$)

季节	方法	中心区域				南部区域			
		$PM_{2.5}$	SO_2	NO_2	8 小时 O_3	$PM_{2.5}$	SO_2	NO_2	8 小时 O_3
春季	移动平滑	4.3 (1.0)	4.0 (0.4)	1.8 (0.4)	4.1 (0.8)	2.8 (1.3)	4.2 (0.4)	3.6 (0.5)	6.3 (1.2)
	趋势分析	1.0 (0.9)	0.4 (0.3)	0.8 (0.4)	4.6 (0.9)	2.5 (1.8)	0.7 (0.5)	1.6 (0.6)	4.6 (1.4)
夏季	移动平滑	3.3 (0.8)	1.7 (0.1)	1.8 (0.3)	7.8 (1.2)	3.1 (0.9)	2.2 (0.2)	1.1 (0.3)	12.9 (1.6)
	趋势分析	3.8 (0.8)	0.1 (0.2)	0.9 (0.3)	3.0 (1.3)	3.4 (1.2)	0.1 (0.2)	0.6 (0.3)	1.4 (3.2)
秋季	移动平滑	5.5 (1.5)	2.5 (0.2)	1.9 (0.5)	1.6 (0.8)	6.8 (2.3)	3.4 (0.3)	2.3 (0.6)	3.5 (1.2)
	趋势分析	2.0 (1.3)	0.6 (0.3)	0.9 (0.5)	2.0 (1.0)	5.4 (2.4)	0.6 (0.6)	2.1 (0.7)	3.3 (1.5)
冬季	移动平滑	6.0 (1.2)	7.8 (0.4)	2.4 (0.4)	1.7 (0.3)	10.5 (2.2)	6.1 (0.6)	3.5 (0.6)	3.1 (0.6)
	趋势分析	6.9 (1.3)	1.7 (0.3)	4.5 (0.5)	2.4 (0.4)	11.3 (1.9)	2.1 (0.4)	3.3 (0.6)	2.0 (0.6)
平均	移动平滑	4.8 (0.6)	4.0 (0.2)	2.0 (0.2)	3.8 (0.4)	5.8 (0.9)	4.0 (0.3)	2.6 (0.3)	6.5 (0.6)
	趋势分析	3.4 (0.6)	0.7 (0.1)	1.3 (0.2)	3.0 (0.4)	5.6 (1.0)	0.9 (0.2)	1.9 (0.3)	2.8 (1.4)

7 讨论

本文提出了一种新的时空调整方法, 该方法消除了气象混杂因素并产生了时空可比的空气质量统计量, 因此可以用于客观地评估某个地区的空气质量. 同时该方法能够量化排放的潜在变化, 而这如果利用排放源清单的话将需要很长的时间. 我们建立了该方法的理论性质, 并将其应用到了对北京地区

几种主要污染物模式和趋势的综合评估中. 理论结果和数值实验为调整方法的表现提供了必要保证.

本文报告的研究主要集中于北京地区, 这里空气质量监测站与气象站点彼此之间距离相对较近. 如果它们距离较远, 为了估计回归方程, 我们可以采用空间 Kriging 法来获得各个污染观测站点的气象变量. 然而, 在中国的大部分城市, 气象站比国控空气污染监测站更为密集, 所以这不是一个问题. 另一个值得考虑的问题是, 我们的方法在空气质量监测站相距较远的情形下是否有效. 一般来说, 只要这些站点的气象协变量有着相互重叠的取值范围, 使得可以定义时空的基准密度 $f_{\cdot j}(\boldsymbol{x})$, 那么新提出的调整方法的表现就可以得到保证. 我们的经验表明, 我们提出的方法可以用来评估相当大区域的空气质量, 如华北平原, 因为华北平原拥有相似的气象特征.

我们的分析结果显示, 到 2017 年初 (数据截止时间), SO_2 显著减少, 而 $PM_{2.5}$ 和 NO_2 的改善却微乎其微. 此外, 地面 O_3 有上升趋势, 值得关注. 尽管在研究中我们使用了非参数回归讨论了空气质量评估, 但是合适的参数或半参数回归模型在这里也可以应用.

致谢 作者衷心感谢匿名审稿人对本文提出的宝贵意见. 陈松蹊的研究得到北京大学数量经济与数理金融教育部重点实验室的支持, 林伟的研究得到北京智源人工智能研究院 (Beijing Academy of Artificial Intelligence, BAAI) 的支持, 郭斌的研究得到西南财经大学统计研究中心的支持.

参考文献

1 Zhang X Y, Wang Y Q, Niu T, et al. Atmospheric aerosol compositions in China: Spatial/temporal variability, chemical signature, regional haze distribution and comparisons with global aerosols. Atmos Chem Phys, 2012, 12: 779–799
2 Guo S, Hu M, Zamora M L, et al. Elucidating severe urban haze formation in China. Proc Natl Acad Sci USA, 2014, 111: 17373–17378
3 Chen L, Guo B, Huang J, et al. Assessing air-quality in Beijing-Tianjin-Hebei region: The method and mixed tales of PM2.5 and O3. Atmos Environ, 2018, 193: 290–301
4 Kuykendal W. Emissions Inventory Guidance for Implementation of Ozone and Particulate Matter National Ambient Air Quality Standards (NAAQS) and Regional Haze Regulations. Washington: Environmental Protection Agency, 2017
5 Liang X, Zou T, Guo B, et al. Assessing Beijing's PM2.5 pollution: Severity, weather impact, APEC and winter heating. Proc R Soc Lond Ser A Math Phys Eng Sci, 2015, 471: 20150257
6 Finazzi F, Scott E M, Fassó A. A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. J Roy Statist Soc Ser C, 2013, 62: 287–308
7 Rosenbaum P R. Observational Studies. New York: Springer-Verlag, 2002
8 Qin J. Biased Sampling, Over-Identified Parameter Problems and Beyond. New York: Springer, 2017
9 Huang C Y, Qin J, Follmann D A. Empirical likelihood-based estimation of the treatment effect in a pretest-posttest study. J Amer Statist Assoc, 2008, 103: 1270–1280
10 Thompson M L, Reynolds J, Cox L H, et al. A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmos Environ, 2001, 35: 617–630
11 Liang X, Li S, Zhang S, et al. $PM_{2.5}$ data reliability, consistency, and air quality assessment in five Chinese cities. J Geophys Res Atmos, 2016, 121: 10,220–10,236
12 Alduchov O A, Eskridge R E. Improved magnus form approximation of saturation vapor pressure. J Appl Meteor, 1996, 35: 601–609
13 Fan J, Yao Q. Nonlinear Time Series: Nonparametric and Parametric Methods. New York: Springer-Verlag, 2003
14 Rosenbaum P R, Rubin D B. The central role of the propensity score in observational studies for causal effects. Biometrika, 1983, 70: 41–55
15 Härdle W. Applied Nonparametric Regression. Cambridge: Cambridge University Press, 1990
16 Carlstein E. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. Ann Statist, 1986, 14: 1171–1179
17 Liu R Y. Bootstrap procedures under some Non-I.I.D. models. Ann Statist, 1988, 16: 1696–1708
18 Härdle W, Mammen E. Comparing nonparametric versus parametric regression fits. Ann Statist, 1993, 21: 1926–1947
19 Kreiss J P, Neumann M H, Yao Q. Bootstrap tests for simple structures in nonparametric time series regression. Stat Its Interface, 2008, 1: 367–380

20 Davison A C, Hinkley D V. Bootstrap Methods and Their Application. Cambridge: Cambridge University Press, 1997
21 Cressie N. Statistics for Spatial Data. New York: John Wiley & Sons, 1993
22 Jun M, Stein M L. Statistical comparison of observed and CMAQ modeled daily sulfate levels. Atmos Environ, 2004, 38: 4427–4436
23 Cleveland W S, Devlin S J. Locally weighted regression: An approach to regression analysis by local fitting. J Amer Statist Assoc, 1988, 83: 596–610
24 Bosq D. Nonparametric Statistics for Stochastic Processes: Estimation and Prediction. New York: Springer-Verlag, 1998

附录 A 假设

这里列出我们需要的假设条件. 一个严平稳过程 $\boldsymbol{\xi}_t$ 是 α 混合的, 如果它的 α 混合系数 $\alpha_\xi(k)$ 满足 $\lim_{k\to\infty}\alpha_\xi(k)=0$, 其中 $\alpha_\xi(k)$ 的定义参见文献 [24]. 为了得到正文第 4 节的渐近性质, 我们需要下面一些条件.

假设 A.1 对于任意给定的 $i=1,\ldots,A_j$, $j=1,\ldots,4$ 和 $s\in\mathcal{R}$, 潜变量 $\{\boldsymbol{U}_{ijt}(s)\}_{t=1}^{n_{ij}}$ 具有相同的分布函数. 进一步, 在给定气象向量 $\boldsymbol{X}_{ijt}(s)$ 的条件下, $\boldsymbol{U}_{ijt}(s)$ 的条件分布函数对于 $t=1,2,\ldots,n_{ij}$ 是相同的. 但在不同的年份 i、季节 j 和空间点 s, $\{\boldsymbol{U}_{ijt}(s)\}_{t=1}^{n_{ij}}$ 的分布和条件分布可以不同.

假设 A.2 关于气象向量 $\boldsymbol{X}_{ijt}(s)$, 我们给出如下假设.

(i) 对于任意给定的 $i=1,\ldots,A_j$ 和 $j=1,\ldots,4$, 多维时间序列 $\{\boldsymbol{X}_{ijt}=\{\boldsymbol{X}_{ijt}(s_1)^{\mathrm{T}},\boldsymbol{X}_{ijt}(s_2)^{\mathrm{T}},\ldots,\boldsymbol{X}_{ijt}(s_L)^{\mathrm{T}}\}^{\mathrm{T}}\}_{t=1}^{n_{ij}}$ 关于时间指标 t 是严平稳的, 并且存在正实数 $a_1>0$ 和 $a_2>1$, 使得 $\boldsymbol{X}_{ijt}$ 的 α- 混合系数 $\alpha_{\boldsymbol{X}}(k)$ 满足对于任意 $k\geqslant 0$ 有 $\alpha_{\boldsymbol{X}}(k)\leqslant a_1k^{-a_2}$.

(ii) 对于任意给定的 $i=1,\ldots,A_j$, $j=1,\ldots,4$ 和 $s\in\mathcal{W}$, 气象向量 $\boldsymbol{X}_{ijt}(s)$ 的概率密度函数 $f_{ij}(\boldsymbol{x},s)$ 关于 $\boldsymbol{x}$ 处处 v 阶连续可微. 进一步, 存在正实数 $c_1,c_2>0$ 使得

$$c_1<\inf_{x\in\mathrm{supp}\{f_{ij}(\boldsymbol{x},s)\}}f_{ij}(\boldsymbol{x},s)\leqslant\sup_{x\in\mathrm{supp}\{f_{ij}(\boldsymbol{x},s)\}}f_{ij}(\boldsymbol{x},s)<c_2,$$

其中 $\mathrm{supp}\{f_{ij}(\boldsymbol{x},s)\}$ 是概率密度函数 $f_{ij}(\boldsymbol{x},s)$ 的支撑集, 定义为

$$\mathrm{supp}\{f_{ij}(\boldsymbol{x},s)\}=\overline{\{\boldsymbol{x}\in\mathbb{R}^d:f_{ij}(\boldsymbol{x},s)>0\}},$$

即集合 $\{x\in\mathbb{R}^d:f_{ij}(\boldsymbol{x},s)>0\}$ 的闭包.

(iii) 气象向量 $\boldsymbol{X}_{ijt}(s)$ 和 $\boldsymbol{X}_{ijt'}(s')$ 的联合概率密度函数 $q_{ij,t-t'}(\boldsymbol{x},\boldsymbol{x}';s,s')$ 关于 $(\boldsymbol{x},\boldsymbol{x}')$ 处处 v 阶连续可微. 进一步, 对于任意给定的 $r=0,1,\ldots,v-1$, $q=1,2,\ldots,d$ 和 $\boldsymbol{x},\boldsymbol{x}'\in\mathbb{R}^d$, 当 $n_{ij}\to+\infty$ 时,

$$n_{ij}^{-1}\sum_{k=1}^{n_{ij}}\frac{\partial^r q_{ij,k}(\boldsymbol{x},\boldsymbol{x}';s,s')}{\partial x_q^r}=O(1). \tag{A.1}$$

(iv) 对于任意给定的 $s'\in\mathcal{W}$ 和 (t,t'), 在给定 $\boldsymbol{X}_{ijt'}(s')$ 的条件下, $\boldsymbol{X}_{ijt}(s)$ 的条件概率密度函数存在且有限, 并且如果假定 $\boldsymbol{X}_{ijt}(s)$ 是严平稳的时间序列, 则可以记为 $p_{ij,t-t'}(\boldsymbol{x}|\boldsymbol{x}';s|s')$. 假设 $p_{ij,t-t'}(\boldsymbol{x}|\boldsymbol{x}';s|s')$ 关于 $(\boldsymbol{x},\boldsymbol{x}')$ 处处 v 阶连续可微, 并且对于任意给定的 $r=0,1,\ldots,v-1$, $q=1,2,\ldots,d$ 和 $\boldsymbol{x},\boldsymbol{x}'\in\mathbb{R}^d$, 当 $n_{ij}\to+\infty$ 时,

$$n_{ij}^{-1}\sum_{k=1}^{n_{ij}}\frac{\partial^r p_{ij,k}(\boldsymbol{x}|\boldsymbol{x}';s|s')}{\partial x_q^r}=O(1). \tag{A.2}$$

(v) 定义 $\|\boldsymbol{X}_{ijt}\|_r := \sup_{s\in\mathcal{W},k=1,\ldots,d}\|X_{ijt,k}(s)\|_r$, 其中 $\|X_{ijt,k}(s)\|_r = \{\mathrm{E}|X_{ijt,k}(s)|^r\}^{1/r}$, $k = 1, \ldots, d$. 假设存在正整数 $r > 2$ 使得 $\|\boldsymbol{X}_{ijt}\|_r < +\infty$.

(vi) 对于任意给定的 $r = 0, 1, \ldots, v-1$, $q = 1, 2, \ldots, d$ 和 $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$, 当 $k \to +\infty$ 时,

$$\frac{\partial^r p_{ij,k}(\boldsymbol{x}\,|\,\boldsymbol{x}';s\,|\,s')}{\partial x_q^r} \to \frac{\partial^r f_{ij}(\boldsymbol{x},s)}{\partial x_q^r}.$$

假设 A.3 对于任意给定的 $i = 1, \ldots, A_j$, $j = 1, \ldots, 4$ 和 $s \in \mathcal{R}$, 假设 (i) 回归函数 $m_{ij}(\boldsymbol{x}, s)$ 关于 $\boldsymbol{x}$ 处处 v 阶连续可微; (ii) 对于任意给定的 $a = 1, \ldots, A_j$, $s' \in \mathcal{W}$, $r = 0, 1, \ldots, v$ 和 $q = 1, \ldots, d$, 积分 $\int f_{aj}(\boldsymbol{x}, s')\frac{\partial^r m_{ij}(\boldsymbol{x},s)}{\partial x_q^r}d\boldsymbol{x}$ 存在且有限.

对于任意给定的 $i = 1, \ldots, A_j$ 和 $j = 1, \ldots, 4$, 定义

$$\mathcal{F}_{ij} = \sigma(\boldsymbol{X}_{ijt}, t \geqslant 0) = \sigma\{\boldsymbol{X}_{ijt}(s_1), \boldsymbol{X}_{ijt}(s_2), \ldots, \boldsymbol{X}_{ijt}(s_L), t \geqslant 0\}$$

为由 $\{\boldsymbol{X}_{ijt} : t \geqslant 0\}$ 或者等价地由 $\{\boldsymbol{X}_{ijt}(s_1), \boldsymbol{X}_{ijt}(s_2), \ldots, \boldsymbol{X}_{ijt}(s_L) : t \geqslant 0\}$ 生成的 σ- 代数.

假设 A.4 对于任意 $i_1 \neq i_2$, σ- 代数 $\mathcal{F}_{i_1 j}$ 和 $\mathcal{F}_{i_2 j}$ 是相互独立的.

假设 A.5 对于任意给定的 $a, i = 1, \ldots, A_j$, $j = 1, \ldots, 4$, $s_1, s_2 \in \mathcal{W}$ 和 $s_3, s_4 \in \mathcal{R}$, 级数

$$\sum_{k=-\infty}^{\infty} |C_{j,ii,k}^a(s_1, s_2; s_3, s_4)| < \infty.$$

假设 A.6 关于标准化残差序列 $e_{ijt}(s)$, 我们给出如下假设.

(i) 对于任意给定的 $i = 1, \ldots, A_j$ 和 $j = 1, \ldots, 4$, 标准化残差向量

$$\{\boldsymbol{e}_{ijt} = \{e_{ijt}(s_1), e_{ijt}(s_2), \ldots, e_{ijt}(s_L)\}^{\mathrm{T}}\}_{t=1}^{n_{ij}}$$

关于时间指标 t 是严平稳的, 且存在正实数 $b_1 > 0$ 和 $b_2 > 1$, 使得标准化残差向量序列 $\boldsymbol{e}_{ijt}$ 的 α- 混合系数满足对任意 $k \geqslant 0$ 都有 $\alpha_e(k) \leqslant b_1 k^{-b_2}$.

(ii) $\mathrm{E}(\boldsymbol{e}_{ijt} \mid \mathcal{F}_{ijt}) = 0$.

(iii) 对任意的 $s, s' \in \mathcal{R}$, 级数 $\sum_{k=-\infty}^{\infty}|\rho_{ij}(k; s, s')| < \infty$ 且 $0 < \sum_{k=-\infty}^{\infty}|\rho_{ij}(k; s, s')|q_{ij,k}(\boldsymbol{x}, \boldsymbol{x}';, s, s') < +\infty$.

假设 A.7 假设在回归函数 $m_{ij}(\boldsymbol{x}, s)$ 的估计量 (4.1) 中, 核函数 $K(\cdot)$ 满足以下条件.

(i) $K(\cdot)$ 是 d 元函数, 其中 d 为连续型随机向量 $\boldsymbol{X}_{ijt}(s)$ 的维数. $K(\cdot)$ 满足

$$\iint\cdots\int K(u_1, u_2, \ldots, u_d)du_1du_2\cdots du_d = 1;$$

(ii) $K(\cdot)$ 是球面对称的, 即对于任意 $\boldsymbol{u}_1 = (u_{1,1}, \ldots, u_{1,d})^{\mathrm{T}} \in \mathbb{R}^d$ 和 $\boldsymbol{u}_2 = (u_{2,1}, \ldots, u_{2,d})^{\mathrm{T}} \in \mathbb{R}^d$, 只要 $\sum_{i=1}^d u_{1,i}^2 = \sum_{i=1}^d u_{2,i}^2$, 就有 $K(\boldsymbol{u}_1) = K(\boldsymbol{u}_2)$;

(iii) $K(\cdot)$ 是 v ($v \geqslant 2$) 阶核函数, 即对于任意的正整数 l 及 $\boldsymbol{r} = (r_1, r_2, \ldots, r_d)^{\mathrm{T}} \in \mathbb{N}$ 满足 $\sum_{i=1}^d r_i = l$, 其中 $\mathbb{N} = \{0, 1, 2, \ldots\}$, 当 $1 \leqslant l < v$ 时,

$$\iint\cdots\int u_1^{r_1}u_2^{r_2}\cdots u_d^{r_d}K(u_1, u_2, \ldots, u_d)du_1du_2\cdots du_d = 0;$$

当 $l = v$ 时,

$$\iint\cdots\int u_1^{r_1}u_2^{r_2}\cdots u_d^{r_d}K(u_1, u_2, \ldots, u_d)du_1du_2\cdots du_d \neq 0.$$

此时, 称 v 为核函数 $K(\cdot)$ 的阶数.

假设 A.8 在 (4.1) 中, 窗宽 $\boldsymbol{H}=(h_1,h_2,\ldots,h_d)^{\mathrm{T}}$ 满足, 当 $n_{ij}\to\infty$ 时,

$$\sum_{q=1}^{d}|h_q|\to 0,\quad n_{ij}\prod_{q=1}^{d}h_q\to\infty\quad 且\quad n_{ij}\prod_{q=1}^{d}h_q^{2v}\to 0.$$

假设 A.9 当 $n_{ij}\to\infty$ 时, $\sup_{i,j}|n_{ij}(\sum_{a=1}^{A_j}n_{aj})^{-1}-A_j^{-1}|=o(\sum_{q=1}^{d}h_q^v)$, 其中 A_j 是对于季节 j 而言所有年份的个数.

附录 B 理论证明

考虑 $\boldsymbol{X}_{ijt}(\boldsymbol{s})$ 中所有协变量都是连续的, 因此,

$$\hat{m}_{ij}(\boldsymbol{x},\boldsymbol{s})=\frac{\sum_{t=1}^{n_{ij}}K_{\boldsymbol{H}}\{\boldsymbol{x}-\boldsymbol{X}_{ijt}(\boldsymbol{s})\}Y_{ijt}(\boldsymbol{s})}{\sum_{t=1}^{n_{ij}}K_{\boldsymbol{H}}\{\boldsymbol{x}-\boldsymbol{X}_{ijt}(\boldsymbol{s})\}},$$
$$\hat{\mu}_{ij}(\boldsymbol{s})=\int\hat{m}_{ij}(\boldsymbol{x},\boldsymbol{s})d\hat{F}_{\cdot j}(\boldsymbol{x})=S^{-1}\Big(\sum_{a=1}^{A_j}n_{aj}\Big)^{-1}\sum_{a=1}^{A_j}\sum_{\boldsymbol{s}'\in\mathcal{W}}\sum_{t=1}^{n_{aj}}\hat{m}_{ij}\{\boldsymbol{X}_{ajt}(\boldsymbol{s}'),\boldsymbol{s}\}.$$

附录 B.1 $\hat{\mu}_{ij}(\boldsymbol{s})$ 的偏差和方差

定义以下与 $\hat{\mu}_{ij}(\boldsymbol{s})$ 的偏差相关的量:

$$b_{ij,a}^{(1)}(\boldsymbol{s};n_{ij})=\frac{\mu_v(K)}{v!S}\sum_{q=1}^{d}\Bigg\{\sum_{\boldsymbol{s}'\in\mathcal{W}}\sum_{r=1}^{v}\binom{v}{r}\int\frac{f_{aj}(\boldsymbol{x},\boldsymbol{s}')}{f_{ij}(\boldsymbol{x},\boldsymbol{s})}\frac{\partial^r m_{ij}(\boldsymbol{x},\boldsymbol{s})}{\partial x_q^r}\frac{\partial^{v-r}f_{ij}(\boldsymbol{x},\boldsymbol{s})}{\partial x_q^{v-r}}d\boldsymbol{x}\Bigg\}h_q^v,$$
$$b_{ij,a}^{(2)}(\boldsymbol{s};n_{ij})=\frac{\mu_v(K)}{v!Sn_{ij}}\sum_{q=1}^{d}\Bigg\{\sum_{\boldsymbol{s}'\in\mathcal{W}}\sum_{r=1}^{v}\sum_{k=-(n_{ij}-1)}^{n_{ij}-1}\binom{v}{r}\int\frac{f_{aj}(\boldsymbol{x},\boldsymbol{s}')}{f_{ij}(\boldsymbol{x},\boldsymbol{s})}\frac{\partial^r m_{ij}(\boldsymbol{x},\boldsymbol{s})}{\partial x_q^r}$$
$$\times\frac{\partial^{v-r}p_{ij,k}(\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{s}|\boldsymbol{s}')}{\partial(x_q')^{v-r}}\bigg|_{\boldsymbol{x}'=\boldsymbol{x}}d\boldsymbol{x}\Bigg\}h_q^v.$$

下面这些量与 $\hat{\mu}_{ij}(\boldsymbol{s})$ 的方差相关:

$$\gamma_{ij}(\boldsymbol{s}_1,\boldsymbol{s}_2;n_{ij})=n_{ij}^{-1}\sum_{k=-n_{ij}+1}^{n_{ij}-1}\rho(|k|;\boldsymbol{s}_1,\boldsymbol{s}_2)\Big(1-\frac{|k|}{n_{ij}}\Big)\iint\sigma_{ij}(\boldsymbol{x}_1,\boldsymbol{s}_1)\sigma_{ij}(\boldsymbol{x}_2,\boldsymbol{s}_2)$$
$$\times q_{ij,k}(\boldsymbol{x}_1,\boldsymbol{x}_2;\boldsymbol{s}_1,\boldsymbol{s}_2)\frac{f_{\cdot j}(\boldsymbol{x}_1)f_{\cdot j}(\boldsymbol{x}_2)}{f_{ij}(\boldsymbol{x}_1,\boldsymbol{s}_1)f_{ij}(\boldsymbol{x}_2,\boldsymbol{s}_2)}d\boldsymbol{x}_1d\boldsymbol{x}_2,$$
$$\lambda_{ii,j}(\boldsymbol{s}_1,\boldsymbol{s}_2;n_{ij})=n_{ij}^{-1}S^{-2}A_j^{-2}\sum_{\boldsymbol{s}_1',\boldsymbol{s}_2'\in\mathcal{W}}\sum_{a=1}^{A_j}\sum_{k=-n_{aj}+1}^{n_{aj}-1}\Big(1-\frac{|k|}{n_{aj}}\Big)C_{ii,j,k}^{a}(\boldsymbol{s}_1',\boldsymbol{s}_2';\boldsymbol{s}_1,\boldsymbol{s}_2).$$

定理 B.1 若假设 A.1、A.2(ii)–A.2(v)、A.3、A.4、A.6(ii)、A.7(i)–A.7(iii) 和 A.9 成立, 那么对于 $j=1,\ldots,4$, $i=1,\ldots,A_j$ 和 $\boldsymbol{s}\in\mathcal{R}$, 当 $n_{ij}\to\infty$ 时, $\hat{\mu}_{ij}(\boldsymbol{s})$ 的偏差和方差分别为

$$\mathrm{Bias}\{\hat{\mu}_{ij}(\boldsymbol{s})\}=A_j^{-1}\Big\{\sum_{a\neq i}b_{ij,a}^{(1)}(\boldsymbol{s};n_{ij})+b_{ij,i}^{(2)}(\boldsymbol{s};n_{ij})\Big\}\{1+o(1)\},$$
$$\mathrm{Var}\{\hat{\mu}_{ij}(\boldsymbol{s})\}=\{\gamma_{ij}(\boldsymbol{s},\boldsymbol{s};n_{ij})+\lambda_{ii,j}(\boldsymbol{s},\boldsymbol{s};n_{ij})\}\{1+o(1)\},$$

此外, 如果条件 A.2(vi) 成立, 当 $n_{ij}\to\infty$ 时,

$$\mathrm{Bias}\{\hat{\mu}_{ij}(\boldsymbol{s})\}=\Big\{A_j^{-1}\sum_{a=1}^{A_j}\delta_a b_{ij,a}^{(1)}(\boldsymbol{s};n_{ij})\Big\}\{1+o(1)\},$$

其中 $\delta_a=I(a\neq i)+2I(a=i)$. 进一步, 在上述两种情形下, 我们有 $\mathrm{Bias}\{\hat{\mu}_{ij}(\boldsymbol{s})\}=O(\sum_{q=1}^d h_q^v)$ 和 $\mathrm{Var}\{\hat{\mu}_{ij}(\boldsymbol{s})\}=O(n_{ij}^{-1})$.

证明 **情形 1** $\hat{\mu}_{ij}(\boldsymbol{s})$ 的偏差推导. 求 $\hat{\mu}_{ij}(\boldsymbol{s})$ 的一阶矩可以归结为计算 $\hat{m}_{ij}(\boldsymbol{X}_{ajt'}(\boldsymbol{s}'),\boldsymbol{s})$ 的一阶矩. 这里应该考虑 $\boldsymbol{X}_{ajt'}(\boldsymbol{s}')$ 的条件分布. 首先, 注意到, 在定理 B.1 的条件下, $b_{ij,a}^{(1)}(\boldsymbol{s};n_{ij})=O(\sum_{q=1}^d h_q^v)$, $b_{ij,a}^{(2)}(\boldsymbol{s};n_{ij})=O(\sum_{q=1}^d h_q^v)$. 根据 $\boldsymbol{X}_{ijt}(\boldsymbol{s})$ 和 $\boldsymbol{X}_{ajt'}(\boldsymbol{s}')$ 的相关性, 考虑以下两种情形.

情形 1.1 如果 $a\neq i$, 因为 $\mathcal{F}_{ij}$ 和 $\mathcal{F}_{aj}$ 是独立的, 当 $n_{ij}\to\infty$ 时, 我们有

$$\begin{aligned}\mathrm{E}[\hat{m}_{ij}\{\boldsymbol{X}_{ajt'}(\boldsymbol{s}'),\boldsymbol{s}\}]&=\mathrm{E}(\mathrm{E}[\hat{m}_{ij}\{\boldsymbol{X}_{ajt'}(\boldsymbol{s}'),\boldsymbol{s}\}|\boldsymbol{X}_{ajt'}(\boldsymbol{s}')])\\&=\mathrm{E}[m_{ij}\{\boldsymbol{X}_{ajt'}(\boldsymbol{s}'),\boldsymbol{s}\}]+\frac{\mu_v(K)}{v!}\sum_{q=1}^d\Big\{\sum_{r=1}^v\binom{v}{r}\int\frac{f_{aj}(\boldsymbol{x},\boldsymbol{s}')}{f_{ij}(\boldsymbol{x},\boldsymbol{s})}\frac{\partial^r m_{ij}(\boldsymbol{x},\boldsymbol{s})}{\partial\boldsymbol{x}_q^r}\\&\quad\times\frac{\partial^{v-r}f_{ij}(\boldsymbol{x},\boldsymbol{s})}{\partial\boldsymbol{x}_q^{v-r}}d\boldsymbol{x}\Big\}h_q^v+o\Big(\sum_{q=1}^d h_q\Big).\end{aligned}$$

情形 1.2 如果 $a=i$, $\boldsymbol{X}_{ijt}(\boldsymbol{s})$ 和 $\boldsymbol{X}_{ajt'}(\boldsymbol{s}')$ 是时空相关的. 在假设 A.2 下, 可得当 $n_{ij}\to\infty$ 时,

$$\begin{aligned}\mathrm{E}[\hat{m}_{ij}\{\boldsymbol{X}_{ijt'}(\boldsymbol{s}'),\boldsymbol{s}\}]&=\mathrm{E}(\mathrm{E}[\hat{m}_{ij}\{\boldsymbol{X}_{ijt'}(\boldsymbol{s}'),\boldsymbol{s}\}|\boldsymbol{X}_{ijt'}(\boldsymbol{s}')])\\&=\mathrm{E}[m_{ij}\{\boldsymbol{X}_{ijt'}(\boldsymbol{s}'),\boldsymbol{s}\}]+\frac{\mu_v(K)}{v!n_{ij}}\sum_{q=1}^d\Big\{\sum_{t=1}^{n_{ij}}\sum_{r=1}^v\binom{v}{r}\int\frac{f_{ij}(\boldsymbol{x},\boldsymbol{s}')}{f_{ij}(\boldsymbol{x},\boldsymbol{s})}\frac{\partial^r m_{ij}(\boldsymbol{x},\boldsymbol{s})}{\partial x_q^r}\\&\quad\times\frac{\partial^{v-r}p_{ij,t-t'}(\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{s}|\boldsymbol{s}')}{\partial(x_q')^{v-r}}\Big|_{\boldsymbol{x}'=\boldsymbol{x}}d\boldsymbol{x}\Big\}h_q^v+o\Big(\sum_{q=1}^d h_q\Big).\end{aligned}$$

此外, 在假设 A.9 下, 当 $n_{ij}\to\infty$ 时,

$$S^{-1}\Big(\sum_{a=1}^{A_j}n_{aj}\Big)^{-1}\sum_{a=1}^{A_j}\sum_{\boldsymbol{s}'\in\mathcal{W}}\sum_{t'=1}^{n_{aj}}\mathrm{E}[m_{ij}\{\boldsymbol{X}_{ajt'}(\boldsymbol{s}'),\boldsymbol{s}\}]-\mu_{ij}(\boldsymbol{s})=o(h_1^v+\cdots+h_d^v).\tag{B.1}$$

由情形 1.1 和 1.2 中的结果及 (B.1), 我们可以得到当 $n_{ij}\to\infty$ 时,

$$\mathrm{Bias}\{\hat{\mu}_{ij}(\boldsymbol{s})\}=A_j^{-1}\Big\{\sum_{a\neq i}b_{ij,a}^{(1)}(\boldsymbol{s};n_{ij})+b_{ij,i}^{(2)}(\boldsymbol{s};n_{ij})\Big\}+o\Big(\sum_{q=1}^d h_q^v\Big).$$

进一步, 在假设 A.2(vi) 下, 根据 Stolz-Cesàro 定理, 可得

$$\mathrm{Bias}\{\hat{\mu}_{ij}(\boldsymbol{s})\}=A_j^{-1}\sum_{a=1}^{A_j}\delta_a b_{ij,a}^{(1)}(\boldsymbol{s};n_{ij})+o\Big(\sum_{q=1}^d h_q^v\Big).$$

情形 2 $\hat{\mu}_{ij}(\boldsymbol{s})$ 的方差推导. 注意到在定理 B.1 的假设下, $\gamma_{ij}(\boldsymbol{s},\boldsymbol{s};n_{ij})=O(n_{ij}^{-1})$ 和 $\lambda_{ii,j}(\boldsymbol{s},\boldsymbol{s};n_{ij})=O(n_{ij}^{-1})$. 为了得到 $\hat{\mu}_{ij}(\boldsymbol{s})$ 的方差, 首先注意到有下面分解式:

$$\hat{\mu}_{ij}(\boldsymbol{s})=\int\hat{m}_{ij}(\boldsymbol{x},\boldsymbol{s})d\hat{F}_{\cdot j}(\boldsymbol{x})=\mu_{ij}(\boldsymbol{s})+T_{ij,1}(\boldsymbol{s})+T_{ij,2}(\boldsymbol{s})+T_{ij,3}(\boldsymbol{s}),\tag{B.2}$$

其中

$$T_{ij,1}(\boldsymbol{s})=\int\{\hat{m}_{ij}(\boldsymbol{x},\boldsymbol{s})-m_{ij}(\boldsymbol{x},\boldsymbol{s})\}dF_{\cdot j}(\boldsymbol{x}),\quad T_{ij,2}(\boldsymbol{s})=\int m_{ij}(\boldsymbol{x},\boldsymbol{s})d\{\hat{F}_{\cdot j}(\boldsymbol{x})-F_{\cdot j}(\boldsymbol{x})\},$$
$$T_{ij,3}(\boldsymbol{s})=\int\{\hat{m}_{ij}(\boldsymbol{x},\boldsymbol{s})-m_{ij}(\boldsymbol{x},\boldsymbol{s})\}d\{\hat{F}_{\cdot j}(\boldsymbol{x})-F_{\cdot j}(\boldsymbol{x})\}.$$

可以证明 $\mathrm{Var}\{\hat{\mu}_{ij}(\boldsymbol{s})\}=[\mathrm{Var}\{T_{ij,1}(\boldsymbol{s})\}+\mathrm{Var}\{T_{ij,2}(\boldsymbol{s})\}+2\mathrm{Cov}\{T_{ij,1}(\boldsymbol{s}),T_{ij,2}(\boldsymbol{s})\}]\{1+o(1)\}$. 首先处理 $\mathrm{Var}\{T_{ij,1}(\boldsymbol{s})\}$. 由于 $\hat{f}_{ij}(\boldsymbol{x},\boldsymbol{s})$ 是 $f_{ij}(\boldsymbol{x},\boldsymbol{s})$ 的相合估计, 可以得到, 当 $n_{ij}\to\infty$ 时,

$$T_{ij,1}(\boldsymbol{s})=\{T_{ij,1}^{(1)}(\boldsymbol{s})+T_{ij,1}^{(2)}(\boldsymbol{s})\}\{1+o_P(1)\},\tag{B.3}$$

其中

$$T_{ij,1}^{(1)}(\boldsymbol{s})=\frac{1}{n_{ij}}\sum_{t=1}^{n_{ij}}\int\frac{1}{f_{ij}(\boldsymbol{x},\boldsymbol{s})}K_{\boldsymbol{H}}\{\boldsymbol{x}-\boldsymbol{X}_{ijt}(\boldsymbol{s})\}[m_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{s}\}-m_{ij}(\boldsymbol{x},\boldsymbol{s})]f_{\cdot j}(\boldsymbol{x})dx,$$
$$T_{ij,1}^{(2)}(\boldsymbol{s})=\frac{1}{n_{ij}}\sum_{t=1}^{n_{ij}}\sigma_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{s}\}e_{ijt}(\boldsymbol{s})\int\frac{1}{f_{ij}(\boldsymbol{x},\boldsymbol{s})}K_{\boldsymbol{H}}\{\boldsymbol{x}-\boldsymbol{X}_{ijt}(\boldsymbol{s})\}f_{\cdot j}(\boldsymbol{x})dx.$$

通过计算, 得到 $\mathrm{Var}\{T_{ij,1}^{(1)}(\boldsymbol{s})\}=O(\sum_{q=1}^{d}h_q^{2v})$ 和 $\mathrm{Var}\{T_{ij,1}^{(2)}(\boldsymbol{s})\}=\gamma_{ij}(\boldsymbol{s},\boldsymbol{s};n_{ij})\{1+o(1)\}$. 因此, 当 $n_{ij}\to\infty$ 时,

$$\mathrm{Var}\{T_{ij,1}(\boldsymbol{s})\}=\gamma_{ij}(\boldsymbol{s},\boldsymbol{s};n_{ij})+o(n_{ij}^{-1}).$$

对于第二项 $T_{ij,2}(\boldsymbol{s})=\int m_{ij}(\boldsymbol{x},\boldsymbol{s})d\{\hat{F}_{\cdot j}(\boldsymbol{x})-F_{\cdot j}(\boldsymbol{x})\}$, 可以证明

$$T_{ij,2}(\boldsymbol{s})=S^{-1}\sum_{a=1}^{A_j}\sum_{\boldsymbol{s}'\in\mathcal{W}}\left[\frac{\hat{\omega}_{aj}}{n_{aj}}\sum_{t=1}^{n_{aj}}m_{ij}\{\boldsymbol{X}_{ajt}(\boldsymbol{s}'),\boldsymbol{s}\}-A_j^{-1}\int m_{ij}(\boldsymbol{x},\boldsymbol{s})dF_{aj}(\boldsymbol{x},\boldsymbol{s}')\right].\tag{B.4}$$

通过分别推导 $\mathrm{E}\{T_{ij,2}(\boldsymbol{s})\}$ 和 $\mathrm{E}\{T_{ij,2}^2(\boldsymbol{s})\}$, 我们有 $n_{ij}\to\infty$ 时,

$$\mathrm{Var}\{T_{ij,2}(\boldsymbol{s})\}=\lambda_{ii,j}(\boldsymbol{s},\boldsymbol{s};n_{ij})+o(n_{ij}^{-1}).$$

$T_{ij,1}(\boldsymbol{s})$ 和 $T_{ij,2}(\boldsymbol{s})$ 的协方差满足 $\mathrm{Cov}\{T_{ij,1}(\boldsymbol{s}),T_{ij,2}(\boldsymbol{s})\}=O(\sum_{q=1}^{d}h_q^{v})$. 因此, 当 $n_{ij}\to\infty$ 时, $\hat{\mu}_{ij}(\boldsymbol{s})$ 的方差是 $\mathrm{Var}\{\hat{\mu}_{ij}(\boldsymbol{s})\}=\gamma_{ij}(\boldsymbol{s},\boldsymbol{s};n_{ij})+\lambda_{ii,j}(\boldsymbol{s},\boldsymbol{s};n_{ij})+o(n_{ij}^{-1})$. □

附录 B.2 定理 4.1 的证明

除了前面定义的一些量, 这里定义

$$\gamma_{ij,k}(\boldsymbol{s}_1,\boldsymbol{s}_2)=\rho(|k|;\boldsymbol{s}_1,\boldsymbol{s}_2)\iint\sigma_{ij}(\boldsymbol{x}_1,\boldsymbol{s}_1)\sigma_{ij}(\boldsymbol{x}_2,\boldsymbol{s}_2)\frac{q_{ij,k}(\boldsymbol{x}_1,\boldsymbol{x}_2;\boldsymbol{s}_1,\boldsymbol{s}_2)}{f_{ij}(\boldsymbol{x}_1,\boldsymbol{s}_1)f_{ij}(\boldsymbol{x}_2,\boldsymbol{s}_2)}dF_{\cdot j}(\boldsymbol{x}_1)dF_{\cdot j}(\boldsymbol{x}_2).$$

通过对 $\gamma_{ij,k}(\boldsymbol{s}_1,\boldsymbol{s}_2)$ 的 k 进行求和, 我们得到 $\gamma_{ij}(\boldsymbol{s}_1,\boldsymbol{s}_2)=\sum_{k=-\infty}^{+\infty}\gamma_{ij,k}(\boldsymbol{s}_1,\boldsymbol{s}_2)$.

附录 B.2.1 $\hat{\mu}_{ij}(\boldsymbol{s})$ 的渐近正态性

为了得到 $\hat{\mu}_{ij}(\boldsymbol{s})$ 的渐近正态性, 我们仍然考虑分解式 (B.2)–(B.4). 注意到 $T_{ij,3}(\boldsymbol{s})=o_P\{T_{ij,1}(\boldsymbol{s})+T_{ij,2}(\boldsymbol{s})\}$. 当 $n_{ij}\to\infty$ 时, 我们重新对 $\hat{\mu}_{ij}(\boldsymbol{s})$ 进行分解:

$$\hat{\mu}_{ij}(\boldsymbol{s})=\{T_{ij}^{(1)}(\boldsymbol{s})+T_{ij}^{(2)}(\boldsymbol{s})+T_{ij}^{(3)}(\boldsymbol{s})\}\{1+o_P(1)\},\tag{B.5}$$

其中

$$T_{ij}^{(1)}(\boldsymbol{s})=\frac{1}{n_{ij}}\sum_{t=1}^{n_{ij}}\int K_{\boldsymbol{H}}\{\boldsymbol{x}-\boldsymbol{X}_{ijt}(\boldsymbol{s})\}[m_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{s}\}-m_{ij}(\boldsymbol{x},\boldsymbol{s})]\frac{f_{\cdot j}(\boldsymbol{x})}{f_{ij}(\boldsymbol{x},\boldsymbol{s})}d\boldsymbol{x},$$

$$T_{ij}^{(2)}(\boldsymbol{s})=\frac{1}{n_{ij}}\sum_{t=1}^{n_{ij}}\left[\sigma_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{s}\}e_{ijt}(\boldsymbol{s})\int\frac{K_{\boldsymbol{H}}\{\boldsymbol{x}-\boldsymbol{X}_{ijt}(\boldsymbol{s})\}}{f_{ij}(\boldsymbol{x},\boldsymbol{s})}dF_{\cdot j}(\boldsymbol{x})+\sum_{\boldsymbol{s}'\in\mathcal{W}}\frac{m_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}'),\boldsymbol{s}\}}{SA_j}\right],$$

$$T_{ij}^{(3)}(\boldsymbol{s})=S^{-1}A_j^{-1}\sum_{a\neq i}\frac{1}{n_{aj}}\sum_{t=1}^{n_{aj}}\sum_{\boldsymbol{s}'\in\mathcal{W}}m_{ij}\{\boldsymbol{X}_{ajt}(\boldsymbol{s}'),\boldsymbol{s}\}.$$

注意到 $\mathrm{Var}\{T_{ij}^{(1)}(\boldsymbol{s})\}=o(n_{ij}^{-1}\sum_{q=1}^{d}h_q^v)$. 因此, 当 $n_{ij}\to\infty$ 时,

$$\sqrt{n_{ij}}[T_{ij}^{(1)}(\boldsymbol{s})-\mathrm{E}\{T_{ij}^{(1)}(\boldsymbol{s})\}]\xrightarrow{P}0.$$

为了得到 $T_{ij}^{(2)}(\boldsymbol{s})$ 的渐近性质, 定义 $\varsigma_{ijt}(\boldsymbol{s})=\varsigma_{ijt,1}(\boldsymbol{s})+\varsigma_{ijt,2}(\boldsymbol{s})$, 其中

$$\varsigma_{ijt,1}(\boldsymbol{s})=\sigma_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}),\boldsymbol{s}\}e_{ijt}(\boldsymbol{s})\int\frac{K_{\boldsymbol{H}}\{\boldsymbol{x}-\boldsymbol{X}_{ijt}(\boldsymbol{s})\}}{f_{ij}(\boldsymbol{x},\boldsymbol{s})}dF_{\cdot j}(\boldsymbol{x}),$$

$$\varsigma_{ijt,2}(\boldsymbol{s})=S^{-1}A_j^{-1}\sum_{\boldsymbol{s}'\in\mathcal{W}}m_{ij}\{\boldsymbol{X}_{ijt}(\boldsymbol{s}'),\boldsymbol{s}\}.$$

由此, $T_{ij}^{(2)}(\boldsymbol{s})=n_{ij}^{-1}\sum_{t=1}^{n_{ij}}\varsigma_{ijt}(\boldsymbol{s})$. 由于 $\mathrm{E}(\boldsymbol{e}_{ijt}\mid\mathcal{F}_{ijt})=0$, 因此, 对于 t_1 和 t_2, 有 $\mathrm{Cov}\{\varsigma_{ijt_1}(\boldsymbol{s}),\varsigma_{ijt_2}(\boldsymbol{s})\}=\mathrm{Cov}\{\varsigma_{ijt_1,1}(\boldsymbol{s}),\varsigma_{ijt_2,1}(\boldsymbol{s})\}+\mathrm{Cov}\{\varsigma_{ijt_1,2}(\boldsymbol{s}),\varsigma_{ijt_2,2}(\boldsymbol{s})\}$. 可以证明 $\mathrm{Cov}\{\varsigma_{ijt_1,1}(\boldsymbol{s}),\varsigma_{ij,t_2,1}(\boldsymbol{s})\}=\gamma_{ij,t_1-t_2}(\boldsymbol{s}_1,\boldsymbol{s}_2)$. 进一步, 有 $\mathrm{Cov}\{\varsigma_{ijt_1,2}(\boldsymbol{s}),\varsigma_{ijt_2,2}(\boldsymbol{s})\}=S^{-2}A_j^{-2}\sum_{\boldsymbol{s}_1',\boldsymbol{s}_2'\in\mathcal{W}}C_{ii,j,t_1-t_2}^{i}(\boldsymbol{s}_1',\boldsymbol{s}_2';\boldsymbol{s},\boldsymbol{s})$. 因此, $\{\boldsymbol{\varsigma}_{ijt}(\boldsymbol{s})\}_{t=1}^{n_{ij}}$ 的长期方差是

$$\sigma_{ij}^{(2)}(\boldsymbol{s})=\sum_{k=-\infty}^{\infty}\mathrm{Cov}\{\varsigma_{ij0}(\boldsymbol{s}),\varsigma_{ijk}(\boldsymbol{s})\}=\gamma_{ij}(\boldsymbol{s},\boldsymbol{s})+S^{-2}A_j^{-2}\sum_{k=-\infty}^{+\infty}\sum_{\boldsymbol{s}_1',\boldsymbol{s}_2'\in\mathcal{W}}C_{ii,j,k}^{i}(\boldsymbol{s}_1',\boldsymbol{s}_2';\boldsymbol{s},\boldsymbol{s}).$$

于是, 由弱相依数据的中心极限定理 [24], 当 $n_{ij}\to\infty$ 时,

$$\sqrt{n_{ij}}[T_{ij}^{(2)}(\boldsymbol{s})-\mathrm{E}\{T_{ij}^{(2)}(\boldsymbol{s})\}]\xrightarrow{d}N(0,\sigma_{ij}^{(2)}(\boldsymbol{s})).$$

类似地, 可以证明, 当 $n_{ij}\to\infty$ 时,

$$\sqrt{n_{ij}}[T_{ij}^{(3)}(\boldsymbol{s})-\mathrm{E}\{T_{ij}^{(3)}(\boldsymbol{s})\}]\xrightarrow{d}N(0,\sigma_{ij}^{(3)}(\boldsymbol{s})),$$

其中 $\sigma_{ij}^{(3)}(\boldsymbol{s})=S^{-2}A_j^{-2}\sum_{a\neq i}\sum_{k=-\infty}^{\infty}\sum_{\boldsymbol{s}_1',\boldsymbol{s}_2'\in\mathcal{W}}C_{ii,j,k}^{a}(\boldsymbol{s}_1',\boldsymbol{s}_2';\boldsymbol{s},\boldsymbol{s})$. 由于 $T_{ij}^{(3)}(\boldsymbol{s})$ 和 $T_{ij}^{(1)}(\boldsymbol{s})+T_{ij}^{(2)}(\boldsymbol{s})$ 是独立的, 而且 $\mathrm{Bias}\{\hat{\mu}_{ij}(\boldsymbol{s})\}=O(\sum_{q=1}^{d}h_q^v)$, 由 Slutsky 定理可知, 当 $n_{ij}\to\infty$ 时, $\sqrt{n_{ij}}\{\hat{\mu}_{ij}(\boldsymbol{s})-\mu_{ij}(\boldsymbol{s})\}\xrightarrow{d}N(0,\tilde{\sigma}_{ij}^2(\boldsymbol{s},\boldsymbol{s}))$, 其中 $\tilde{\sigma}_{ij}^2(\boldsymbol{s},\boldsymbol{s})=\gamma_{ij}(\boldsymbol{s},\boldsymbol{s})+\lambda_{ii,j}(\boldsymbol{s},\boldsymbol{s})$.

附录 B.2.2 $\hat{\boldsymbol{\mu}}_{\boldsymbol{ij}}(\boldsymbol{\mathcal{A}})$ 的渐近正态性

由证明 $\hat{\mu}_{ij}(\boldsymbol{s})$ 的渐近正态性中 (B.5), 当 $n_{ij}\to\infty$ 时,

$$\sum_{\boldsymbol{s}\in\mathcal{A}}\hat{\mu}_{ij}(\boldsymbol{s})=\sum_{\boldsymbol{s}\in\mathcal{A}}\{T_{ij}^{(1)}(\boldsymbol{s})+T_{ij}^{(2)}(\boldsymbol{s})+T_{ij}^{(3)}(\boldsymbol{s})\}\{1+o_P(1)\},$$

其中 $T_{ij}^{(1)}(\boldsymbol{s})$、$T_{ij}^{(2)}(\boldsymbol{s})$ 和 $T_{ij}^{(3)}(\boldsymbol{s})$ 定义在 (B.5). 由于 $\mathrm{Var}\{T_{ij}^{(1)}(\boldsymbol{s})\} = o(n_{ij}^{-1}\sum_{q=1}^{d} h_q^v)$, 当 $n_{ij}\to\infty$ 时,

$$\sqrt{n_{ij}}\Big[\sum_{\boldsymbol{s}\in\mathcal{A}} T_{ij}^{(1)}(\boldsymbol{s}) - \mathrm{E}\Big\{\sum_{\boldsymbol{s}\in\mathcal{A}} T_{ij}^{(1)}(\boldsymbol{s})\Big\}\Big] \xrightarrow{P} 0. \tag{B.6}$$

类似于之前的证明, 当 $n_{ij}\to\infty$ 时,

$$\sqrt{n_{ij}}\Big[\sum_{\boldsymbol{s}\in\mathcal{A}} T_{ij}^{(2)}(\boldsymbol{s}) - \mathrm{E}\Big\{\sum_{\boldsymbol{s}\in\mathcal{A}} T_{ij}^{(2)}(\boldsymbol{s})\Big\}\Big] \xrightarrow{d} N\Big(0, \sum_{k=-\infty}^{\infty}\Xi_{ij,k}^{(2)}\Big), \tag{B.7}$$

$$\sqrt{n_{ij}}\Big[\sum_{\boldsymbol{s}\in\mathcal{A}} T_{ij}^{(3)}(\boldsymbol{s}) - \mathrm{E}\Big\{\sum_{\boldsymbol{s}\in\mathcal{A}} T_{ij}^{(3)}(\boldsymbol{s})\Big\}\Big] \xrightarrow{d} N\Big(0, \sum_{k=-\infty}^{\infty}\Xi_{ij,k}^{(3)}\Big), \tag{B.8}$$

其中

$$\Xi_{ij,k}^{(2)} = \sum_{\boldsymbol{s}_1,\boldsymbol{s}_2\in\mathcal{A}}\gamma_{ij,k}(\boldsymbol{s}_1,\boldsymbol{s}_2) + S^{-2}A_j^{-2}\sum_{\boldsymbol{s}_1,\boldsymbol{s}_2\in\mathcal{A}}\sum_{\boldsymbol{s}_1',\boldsymbol{s}_2'\in\mathcal{W}} C_{ii,j,k}^{i}(\boldsymbol{s}_1',\boldsymbol{s}_2';\boldsymbol{s}_1,\boldsymbol{s}_2),$$
$$\Xi_{ij,k}^{(3)} = S^{-2}A_j^{-2}\sum_{a\neq i}\sum_{\boldsymbol{s}_1,\boldsymbol{s}_2\in\mathcal{A}}\sum_{\boldsymbol{s}_1',\boldsymbol{s}_2'\in\mathcal{W}} C_{ii,j,k}^{a}(\boldsymbol{s}_1',\boldsymbol{s}_2';\boldsymbol{s}_1,\boldsymbol{s}_2).$$

由 Slutsky 定理, 当 $n_{ij}\to\infty$ 时,

$$\sqrt{n_{ij}}\Big[\sum_{\boldsymbol{s}\in\mathcal{A}}\hat{\mu}_{ij}(\boldsymbol{s}) - \mathrm{E}\Big\{\sum_{\boldsymbol{s}\in\mathcal{A}}\hat{\mu}_{ij}(\boldsymbol{s})\Big\}\Big] \xrightarrow{d} N(0,\tilde{\sigma}_{ij}^0(\mathcal{A})),$$

其中 $\tilde{\sigma}_{ij}^0(\mathcal{A}) = \sum_{\boldsymbol{s}_1,\boldsymbol{s}_2\in\mathcal{A}}\{\gamma_{ij}(\boldsymbol{s}_1,\boldsymbol{s}_2)+\lambda_{ii,j}(\boldsymbol{s}_1,\boldsymbol{s}_2)\}$. 进一步, 当 $n_{ij}\to\infty$ 时, $\mathrm{Bias}\{\hat{\mu}_{ij}(\mathcal{A})\} = O(\sum_{q=1}^{d} h_q^v)$. 由此, 根据连续映射定理, 我们可以直接得到定理 4.1 中 $\hat{\mu}_{ij}(\mathcal{A})$ 的渐近正态性.

附录 B.3　定理 4.2 的证明

为了扩展 $\lambda_{i_1i_2,j}(\boldsymbol{s}_1,\boldsymbol{s}_2)$ 的定义, 定义

$$\begin{aligned}\phi_{i_1i_2,j,k}^{a}(\boldsymbol{s}_1,\boldsymbol{s}_2) = S^{-2}\sum_{\boldsymbol{s}_1',\boldsymbol{s}_2'\in\mathcal{W}}\{&C_{i_1i_1,j,k}^{a}(\boldsymbol{s}_1',\boldsymbol{s}_2';\boldsymbol{s}_1,\boldsymbol{s}_2) + C_{i_2i_2,j,k}^{a}(\boldsymbol{s}_1',\boldsymbol{s}_2';\boldsymbol{s}_1,\boldsymbol{s}_2)\\ &- C_{i_1i_2,j,k}^{a}(\boldsymbol{s}_1',\boldsymbol{s}_2';\boldsymbol{s}_1,\boldsymbol{s}_2) - C_{i_2i_1,j,k}^{a}(\boldsymbol{s}_1',\boldsymbol{s}_2';\boldsymbol{s}_1,\boldsymbol{s}_2)\}.\end{aligned}$$

附录 B.3.1　$\hat{\mu}_{i_2j}(\mathcal{A}) - \hat{\mu}_{i_1j}(\mathcal{A})$ 的渐近正态性

由定理 4.1 证明中 (B.5), 对于任意的 $i_1\neq i_2$, 当 $n_{ij}\to\infty$ 时,

$$\sum_{\boldsymbol{s}\in\mathcal{A}}\{\hat{\mu}_{i_2j}(\boldsymbol{s}) - \hat{\mu}_{i_1j}(\boldsymbol{s})\} = \sum_{\boldsymbol{s}\in\mathcal{A}}\{T_{i_2i_1,j}^{(1)}(\boldsymbol{s}) + T_{i_2i_1,j}^{(2)}(\boldsymbol{s}) - T_{i_1i_2,j}^{(2)}(\boldsymbol{s}) + T_{i_2i_1,j}^{(3)}(\boldsymbol{s})\}\{1+o_P(1)\},$$

其中

$$T_{i_2i_1,j}^{(1)}(\boldsymbol{s}) = T_{i_2j}^{(1)}(\boldsymbol{s}) - T_{i_1j}^{(1)}(\boldsymbol{s}),\quad T_{i_2i_1,j}^{(2)}(\boldsymbol{s}) = T_{i_2j}^{(2)}(\boldsymbol{s}) - S^{-1}A_j^{-1}n_{i_2j}^{-1}\sum_{t=1}^{n_{i_2j}}\sum_{\boldsymbol{s}'\in\mathcal{W}} m_{i_1j}\{\boldsymbol{X}_{i_2jt}(\boldsymbol{s}'),\boldsymbol{s}\},$$

$$T_{i_2i_1,j}^{(3)}(\boldsymbol{s}) = S^{-1}A_j^{-1}\sum_{a\neq i_1,i_2}\frac{1}{n_{aj}}\sum_{t=1}^{n_{aj}}\sum_{\boldsymbol{s}'\in\mathcal{W}}[m_{i_2j}\{\boldsymbol{X}_{ajt}(\boldsymbol{s}'),\boldsymbol{s}\} - m_{i_1j}\{\boldsymbol{X}_{ajt}(\boldsymbol{s}'),\boldsymbol{s}\}].$$

由 (B.6) 和 (B.8), 当 $n_{i_1j}, n_{i_2j} \to +\infty$ 时, 有

$$\sqrt{n_{i_1j}}\bigg[\sum_{\boldsymbol{s}\in\mathcal{A}} T^{(1)}_{i_2i_1,j}(\boldsymbol{s}) - \mathrm{E}\bigg\{\sum_{\boldsymbol{s}\in\mathcal{A}} T^{(1)}_{i_2i_1,j}(\boldsymbol{s})\bigg\}\bigg] \xrightarrow{P} 0,$$
$$\sqrt{n_{i_1j}}\bigg[\sum_{\boldsymbol{s}\in\mathcal{A}} T^{(3)}_{i_2i_1,j}(\boldsymbol{s}) - \mathrm{E}\bigg\{\sum_{\boldsymbol{s}\in\mathcal{A}} T^{(3)}_{i_2i_1,j}(\boldsymbol{s})\bigg\}\bigg] \xrightarrow{d} N\bigg(0, \sum_{k=-\infty}^{\infty} \Omega^{(3)}_{i_2i_1,j,k}\bigg),$$

其中 $\Omega^{(3)}_{i_2i_1,j,k} = A_j^{-2}\sum_{a\neq i_1,i_2}\sum_{\boldsymbol{s}_1,\boldsymbol{s}_2\in\mathcal{A}} \phi^{a}_{i_2i_1,j,k}(\boldsymbol{s}_1,\boldsymbol{s}_2)$. 类似于 (B.7), 当 $n_{i_2j}\to+\infty$ 时,

$$\sqrt{n_{i_1j}}\bigg[\sum_{\boldsymbol{s}\in\mathcal{A}} T^{(2)}_{i_2i_1,j}(\boldsymbol{s}) - \mathrm{E}\bigg\{\sum_{\boldsymbol{s}\in\mathcal{A}} T^{(2)}_{i_2i_1,j}(\boldsymbol{s})\bigg\}\bigg] \xrightarrow{d} N\bigg(0, \sum_{k=-\infty}^{\infty} \Omega^{(2)}_{i_2i_1,j,k}\bigg),$$

其中 $\Omega^{(2)}_{i_2i_1,j,k} = \sum_{\boldsymbol{s}_1,\boldsymbol{s}_2\in\mathcal{A}}\gamma_{i_2j,k}(\boldsymbol{s}_1,\boldsymbol{s}_2) + A_j^{-2}\sum_{\boldsymbol{s}_1,\boldsymbol{s}_2\in\mathcal{A}}\phi^{i_2}_{i_2i_1,j,k}(\boldsymbol{s}_1,\boldsymbol{s}_2)$. 由 Slutsky 定理知, 当 $n_{i_1j}, n_{i_2j}\to+\infty$ 时,

$$\sqrt{n_{i_1j}}\bigg(\sum_{\boldsymbol{s}\in\mathcal{A}}\{\hat{\mu}_{i_2j}(\boldsymbol{s}) - \hat{\mu}_{i_1j}(\boldsymbol{s})\} - \mathrm{E}\bigg[\sum_{\boldsymbol{s}\in\mathcal{A}}\{\hat{\mu}_{i_2j}(\boldsymbol{s}) - \hat{\mu}_{i_1j}(\boldsymbol{s})\}\bigg]\bigg) \xrightarrow{d} N(0, \tilde{\sigma}^0_{i_2i_1,j}(\mathcal{A})),$$

其中 $\tilde{\sigma}^0_{i_2i_1,j}(\mathcal{A}) = \sum_{\boldsymbol{s}_1,\boldsymbol{s}_2\in\mathcal{A}}\{\gamma_{i_2j}(\boldsymbol{s}_1,\boldsymbol{s}_2) + \phi_{i_2i_1,j}(\boldsymbol{s}_1,\boldsymbol{s}_2)\}$. 进一步, 当 $n_{i_1j}, n_{i_2j}\to+\infty$ 时,

$$\mathrm{Bias}\{\hat{\mu}_{i_2j}(\mathcal{A}) - \hat{\mu}_{i_1j}(\mathcal{A})\} = O\bigg(\sum_{q=1}^{d} h_q^v\bigg).$$

于是, 根据连续映射定理, 可以得到 $\hat{\mu}_{i_2j}(\mathcal{A}) - \hat{\mu}_{i_1j}(\mathcal{A})$ 的渐近正态性.

附录 B.3.2 $\hat{\boldsymbol{\mu}}_{ij}(\mathcal{A}) - \hat{\boldsymbol{\mu}}_{ij}(\mathcal{B})$ 的渐近正态性

对于区域 $\mathcal{A}$ 和 $\mathcal{B}$ 使得 $\mathcal{A}\cap\mathcal{B} = \emptyset$, 令 $M_1 = |\mathcal{A}|$, $M_2 = |\mathcal{B}|$, $M = M_1 + M_2$. 不失一般性, 令 $\mathcal{A} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{M_1}\}$ 和 $\mathcal{B} = \{\boldsymbol{s}_{M_1+1}, \ldots, \boldsymbol{s}_M\}$. 可以证明

$$\begin{pmatrix}\hat{\mu}_{ij}(\mathcal{A})\\ \hat{\mu}_{ij}(\mathcal{B})\end{pmatrix} = \begin{pmatrix}|\mathcal{A}|^{-1}\sum_{\boldsymbol{s}\in\mathcal{A}}\hat{\mu}_{ij}(\boldsymbol{s})\\ |\mathcal{B}|^{-1}\sum_{s\in\mathcal{B}}\hat{\mu}_{ij}(\boldsymbol{s})\end{pmatrix} = \begin{pmatrix}|\mathcal{A}|^{-1}\mathbf{1}_{M_1} & \mathbf{0}\\ \mathbf{0} & |\mathcal{B}|^{-1}\mathbf{1}_{M_2}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\hat{\mu}_{ij}(\boldsymbol{s}_1)\\ \vdots\\ \hat{\mu}_{ij}(\boldsymbol{s}_M)\end{pmatrix} =: \boldsymbol{\Gamma}\hat{\boldsymbol{\mu}}_{ij}.$$

类似于定理 4.1 的证明, 当 $n_{ij}\to\infty$ 时, $\sqrt{n_{ij}}(\hat{\boldsymbol{\mu}}_{ij} - \boldsymbol{\mu}_{ij}) \xrightarrow{d} N(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_{ij}(\mathcal{A}\cup\mathcal{B}))$. 由连续映射定理, 当 $n_{ij}\to\infty$ 时,

$$\sqrt{n_{ij}}[\{\hat{\mu}_{ij}(\mathcal{A}), \hat{\mu}_{ij}(\mathcal{B})\}^{\mathrm{T}} - \{\mu_{ij}(\mathcal{A}), \mu_{ij}(\mathcal{B})\}^{\mathrm{T}}] \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma}\tilde{\boldsymbol{\Sigma}}_{ij}(\mathcal{A}\cup\mathcal{B})\boldsymbol{\Gamma}^{\mathrm{T}}).$$

令 $\boldsymbol{\Gamma}\tilde{\boldsymbol{\Sigma}}_{ij}(\mathcal{A}\cup\mathcal{B})\boldsymbol{\Gamma}^{\mathrm{T}} = \left(\begin{smallmatrix}\theta_{11} & \theta_{12}\\ \theta_{21} & \theta_{22}\end{smallmatrix}\right)$. 于是有

$$\theta_{11} = (|\mathcal{A}|^{-1}\mathbf{1}^{\mathrm{T}}_{|\mathcal{A}|}, \mathbf{0}^{\mathrm{T}})\tilde{\boldsymbol{\Sigma}}_{ij}(\mathcal{A}\cup\mathcal{B})(|\mathcal{A}|^{-1}\mathbf{1}^{\mathrm{T}}_{|\mathcal{A}|}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}} = \tilde{\sigma}^2_{ij}(\mathcal{A},\mathcal{A}),$$
$$\theta_{12} = (|\mathcal{A}|^{-1}\mathbf{1}^{\mathrm{T}}_{|\mathcal{A}|}, \mathbf{0}^{\mathrm{T}})\tilde{\boldsymbol{\Sigma}}_{ij}(\mathcal{A}\cup\mathcal{B})(|\mathcal{A}|^{-1}\mathbf{1}^{\mathrm{T}}_{|\mathcal{A}|}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}} = \tilde{\sigma}^2_{ij}(\mathcal{A},\mathcal{B}),$$
$$\theta_{21} = (\mathbf{0}^{\mathrm{T}}, |\mathcal{B}|^{-1}\mathbf{1}^{\mathrm{T}}_{|\mathcal{B}|})\tilde{\boldsymbol{\Sigma}}_{ij}(\mathcal{A}\cup\mathcal{B})(\mathbf{0}^{\mathrm{T}}, |\mathcal{B}|^{-1}\mathbf{1}^{\mathrm{T}}_{|\mathcal{B}|})^{\mathrm{T}} = \tilde{\sigma}^2_{ij}(\mathcal{B},\mathcal{A}),$$
$$\theta_{22} = (\mathbf{0}^{\mathrm{T}}, |\mathcal{B}|^{-1}\mathbf{1}^{\mathrm{T}}_{|\mathcal{B}|})\tilde{\boldsymbol{\Sigma}}_{ij}(\mathcal{A}\cup\mathcal{B})(\mathbf{0}^{\mathrm{T}}, |\mathcal{B}|^{-1}\mathbf{1}^{\mathrm{T}}_{|\mathcal{B}|})^{\mathrm{T}} = \tilde{\sigma}^2_{ij}(\mathcal{B},\mathcal{B}),$$

其中 $\mathbf{1}_{|\mathcal{A}|}^{\mathrm{T}}=(1,1,\ldots,1)_{|\mathcal{A}|\times 1}$. 由于 $\tilde{\sigma}_{ij}^2(\mathcal{B},\mathcal{A})=\tilde{\sigma}_{ij}^2(\mathcal{A},\mathcal{B})$, 利用 Cramér-Wold 方法, 当 $n_{ij}\to\infty$ 时,

$$\sqrt{n_{ij}}[\{\hat{\mu}_{ij}(\mathcal{A})-\hat{\mu}_{ij}(\mathcal{B})\}-\{\mu_{ij}(\mathcal{A})-\mu_{ij}(\mathcal{B})\}]\xrightarrow{d}N(0,\tilde{\sigma}_{ij}^2(\mathcal{A}-\mathcal{B})),$$

其中 $\tilde{\sigma}_{ij}^2(\mathcal{A}-\mathcal{B})=\tilde{\sigma}_{ij}^2(\mathcal{A},\mathcal{A})-2\tilde{\sigma}_{ij}^2(\mathcal{A},\mathcal{B})+\tilde{\sigma}_{ij}^2(\mathcal{B},\mathcal{B})$.

附录 C　模拟结果

本节将通过仿真实验分析所提出的调整均值浓度的效果. 我们将使用北京市中心城区和南部地区共 28 个污染监测站点的污染物浓度数据和 11 气象站点的气象数据作为基础来进行仿真模拟. 使用实际数据可以使仿真模拟和实际数据情形更加接近, 从而得到的结论更加适用于实际问题.

冬季是北京 $\mathrm{PM}_{2.5}$ 污染最为严重的季节. 为了更好地分析在 $\mathrm{PM}_{2.5}$ 重污染季节调整均值浓度的效果, 本文选择了冬季 (12 月至次年 2 月) 的观测数据作为研究对象. 其中原始气象数据是 2010 年 3 月至 2017 年 2 月的小时数据. 给定每个冬季, 首先把所有站点同一时刻的原始气象数据合并成一个向量, 再对向量在时间上进行分块抽样, 最后把抽样得到的时间分块重新拼凑成一个新的时间序列. 具体来讲, 模拟数据的样本量 $\tilde{n}$ 分别取为 1,080 和 2,160. 因为仿真模拟中考虑的是冬季的数据, 所以在本节中取定季节指标 $j=4$. 回顾记号 $\boldsymbol{X}_{ijt}=\{\boldsymbol{X}_{ijt}(s_1)^{\mathrm{T}},\ldots,\boldsymbol{X}_{ijt}(s_S)^{\mathrm{T}}\}^{\mathrm{T}}$. 对于给定的模拟数据的样本量 $\tilde{n}$, 在年份 i 和站点 s, 我们将原始气象向量 $\{X_{i4t}\}_{t=1}^{n_{i4}}$ 划分成长度 $l=12$ 的时间块, 然后从 n_{i4}/l 个时间块中有放回地独立等概率抽取 $\tilde{n}/l$ 个分块, 再将抽样得到的 $\tilde{n}/l$ 个分块重新拼凑成长度为 $\tilde{n}$ 的样本 $\{\boldsymbol{X}_{i4t}^*\}_{t=1}^{\tilde{n}}$, 最后将所有年份的抽样数据合并起来得到最终的气象模拟数据. 通过对观测到的时间序列进行分块抽样, 我们可以保持模拟数据与原始数据的时间相关性是一致的. 我们用 2015 和 2016 年的气象模拟数据建立回归模型, 用 2010 至 2016 年的气象模拟数据来构造均衡的气象条件.

在生成了气象向量的模拟数据之后, 我们根据模型 (2.2) 来生成 28 个污染物监测站点的污染物浓度的模拟数据. 我们将 2015 和 2016 年的冬季分别称为第一个冬季和第二个冬季, 并将第一个冬季和第二个冬季分别表示为 $k=I$ 和 $k=II$. 对于本节中的仿真模拟, 模型 (2.2) 可以写成如下形式:

$$Y_{i4t}^*(s)=m_{k4}\{\boldsymbol{X}_{i4t}^*(s),s\}+\sigma_{k4}\{\boldsymbol{X}_{i4t}^*(s),s\}\boldsymbol{e}_{k4t}(s),\quad t=1,\ldots,\tilde{n},$$

其中 $\boldsymbol{X}_{i4t}^*(s)$ 和 $Y_{i4t}^*(s)$ 分别表示气象向量和污染物浓度的模拟数据. $i=2015$ 对应于第一个冬季 $k=I$, $i=2016$ 对应于第二个冬季参数 $k=II$. 对于每个冬季, 设定回归函数具有参数形式 $m_{k4}(\boldsymbol{x},s)=m\{\boldsymbol{x};\boldsymbol{\beta}_k(w,s)\}$. 参数 $\boldsymbol{\beta}_k(w,s)$ 在两个冬季的不同站点 s 和风向 w 下不同. 进一步, 回归函数 $m\{\boldsymbol{x};\boldsymbol{\beta}_k(w,s)\}$ 关于参数 $\boldsymbol{\beta}_k(w,s)$ 是线性的, 满足 $\boldsymbol{\beta}_k(w,s)=\{\beta_{k,0}(w,s),\ldots,\beta_{k,6}(w,s)\}^{\mathrm{T}}$ 且

$$m\{\boldsymbol{x};\boldsymbol{\beta}_k(w,s)\}=\beta_{k,0}(w,s)+\sum_{r=1}^{4}m^{(r)}\{x_r;\boldsymbol{\beta}_k(w,s)\}.\tag{C.1}$$

上式中的各个求和项等于

$$m^{(1)}\{x_1;\boldsymbol{\beta}_k(w,s)\}=\beta_{k,1}(w,s)x_1+\beta_{k,2}(w,s)x_1^2,\quad m^{(2)}\{x_2;\boldsymbol{\beta}_k(w,s)\}=\beta_{k,3}(w,s)x_2+\beta_{k,4}(w,s)x_2^2,$$
$$m^{(3)}\{x_3;\boldsymbol{\beta}_k(w,s)\}=\beta_{k,5}(w,s)x_3,\quad m^{(4)}\{x_4;\boldsymbol{\beta}_k(w,s)\}=\beta_{k,6}(w,s)\log x_4,$$

其中 $\boldsymbol{x}=(x_1,x_2,x_3,x_4,w)^{\mathrm{T}}$ 表示由温度、露点温度、气压、累积风速和风向组成的向量. 在模拟仿真中, 我们没有考虑降水量. 这是因为在北京市冬季的降水量非常稀少, 大部分的降水量等于 0.

我们根据实际观测数据 $\{Y_{ijt}(s)\}$ 和 $\{\boldsymbol{X}_{ijt}(s)\}$ 对模型 (C.1) 中的参数分别进行了估计. 2015 年实际观测数据用于估计第一个冬季的参数 $\boldsymbol{\beta}_I(w,s)$, 2016 年即第二个冬季的参数 $\boldsymbol{\beta}_{II}(w,s)$ 基于第一个冬季的参数产生. 具体过程如下. 第一步, 对于每个站点, 污染物浓度和气象变量的观测数据分别被标准化. 第二步, 我们选取了北京奥体中心、房山、官园、美国大使馆、顺义和亦庄这 6 个污染物监测站点在 2015 年的标准化观测数据, 用于拟合模型 (C.1). 这 6 个基准站点覆盖了北京市的大部分区域, 因此具有很好的代表性. 这 6 个站点对应的空间位置组成的集合记为 $\mathcal{S}=\{s_1^0,s_2^0,\ldots,s_6^0\}$. 记用这 6 个站点的标准化观测数据拟合模型 (C.1) 得到的参数估计值为 $\boldsymbol{\beta}^*(w,s)$, $s\in\mathcal{S}$. 第三步, 为了减小观测数据的噪声对估计的影响并抓住主要的趋势特征, 令 $\boldsymbol{\beta}^*(w)=6^{-1}\sum_{s\in\mathcal{S}}\boldsymbol{\beta}^*(w,s)$. 则 $\boldsymbol{\beta}^*(w)$ 提供了给定风向 w 下各个站点标准化数据对应参数的基准, 我们称之为基准参数. 表 C1 给出了基准参数 $\beta^*(w)$ 的取值.

下面首先给出在第一个冬季 ($i=2015$, $k=I$) 的模型 (C.1) 参数设定方法. 对于给定的线性模型, 用原始数据和标准化数据得到的参数估计值之间存在确定的关系. 由此出发, 对于每一个污染物监测站点, 假定用标准化的污染物浓度和气象的观测数据得到的参数估计值为基准参数 $\boldsymbol{\beta}^*(w)$, 然后利用标准化数据与原始数据的参数估计值之间的关系, 得到其对应的模型参数. 具体而言, 令 $D_{i4t}(s)$、$T_{i4t}(s)$、$P_{i4t}(s)$、$C_{i4t}(s)$ 和 $W_{i4t}(s)$ 分别表示在年份 i 的冬季、小时 t 和空间点 s 实际观测到的露点温度、气温、气压、累积风速和风向. 令

$$\boldsymbol{V}_{i4t}(s)=\{D_{i4t}(s),D_{i4t}^2(s),T_{i4t}(s),T_{i4t}^2(s),P_{i4t}(s),\log C_{i4t}(s)\}^{\mathrm{T}},$$

$\boldsymbol{\beta}_k(w,s)=[\beta_{k,0}(w,s),\{\tilde{\boldsymbol{\beta}}_k(w,s)\}^{\mathrm{T}}]^{\mathrm{T}}$ 和 $\boldsymbol{\beta}^*(w)=[\beta_0^*(w),\{\tilde{\boldsymbol{\beta}}^*(w)\}^{\mathrm{T}}]^{\mathrm{T}}$, 则模型 (C.1) 可写为

$$m_{k4}\{\boldsymbol{X}_{i4t}(s),s\}=\beta_{k,0}\{W_{i4t}(s),s\}+V_{i4t}^{\mathrm{T}}(s)\tilde{\boldsymbol{\beta}}_k\{W_{i4t}(s),s\}. \tag{C.2}$$

令 $\mu_{i4}^V(w,s)$ 和 $\mu_{i4}^Y(w,s)$ 分别为 $\boldsymbol{V}_{i4t}(s)$ 和 $Y_{i4t}(s)$ 在给定风向 $W_{ijt}(s)=w$ 条件下的条件样本均值, 即 $\boldsymbol{\mu}_{i4}^V(s)=n_{i4}^{-1}(w)\sum_{t=1}^{n_{ij}}\boldsymbol{V}_{i4t}(s)I\{W_{ijt}(s)=w\}$, $\mu_{i4}^Y(s)=n_{i4}^{-1}(w)\sum_{t=1}^{n_{ij}}Y_{i4t}(s)I\{W_{ijt}(s)=w\}$, 其中 $n_{i4}(w)=\sum_{t=1}^{n_{ij}}I\{W_{ijt}(s)=w\}$. 污染物浓度 $Y_{i4t}(s)$ 在给定风向 $W_{ijt}(s)=w$ 的条件下的条件标准差记为 $\xi_{i4}^Y(w,s)$, 向量 $\boldsymbol{V}_{i4t}(s)$ 的第 l 个分量在给定风向 $W_{ijt}(s)=w$ 的条件下的条件标准差记为 $\xi_{i4,l}^V(w,s)$. 令 $\boldsymbol{\Xi}_{i4}^V(w,s)=\mathrm{diag}\{\xi_{i4,1}^V(w,s),\xi_{i4,2}^V(w,s),\ldots,\xi_{i4,6}^V(w,s)\}$ 表示主对角线元素等于 $\xi_{i4,1}^V(w,s),\xi_{i4,2}^V(w,s),\ldots,\xi_{i4,6}^V(w,s)$ 的对角矩阵. 对于 ($i=2015$, $k=I$) 和任意给定的风向 w, 假设每个站点的标准化观测数据对应的参数估计值为 $\boldsymbol{\beta}^*(w)$, 则其观测数据对应的参数估计值为

$$\begin{aligned}&\beta_{k,0}(w,s)=\beta_0^*(w)+\frac{\mu_{i4}^Y(w,s)}{\xi_{i4}^Y(w,s)}-\{\mu_{i4}^V(w,s)\}^{\mathrm{T}}\{\boldsymbol{\Xi}_{i4}^V(w,s)\}^{-1}\tilde{\boldsymbol{\beta}}^*(w),\\&\tilde{\boldsymbol{\beta}}_k(w,s)=\xi_{i4}^Y(w,s)\{\boldsymbol{\Xi}_{i4}^V(w,s)\}^{-1}\tilde{\boldsymbol{\beta}}^*(w).\end{aligned} \tag{C.3}$$

表 C1 仿真模拟中基准参数 $\boldsymbol{\beta}^*(w)$ 在不同风向下的取值

w	$\beta_0^*(w)$	$\beta_1^*(w)$	$\beta_2^*(w)$	$\beta_3^*(w)$	$\beta_4^*(w)$	$\beta_5^*(w)$	$\beta_6^*(w)$
CV	0.00	1.17	0.44	−0.59	−0.28	−0.24	0.01
NE	0.01	1.20	0.60	−0.41	−0.15	−0.26	−0.03
NW	0.00	1.49	1.13	−0.23	−0.14	−0.18	−0.23
SE	0.00	1.17	0.60	−0.36	−0.09	−0.35	0.02
SW	0.00	1.30	0.86	−0.36	−0.15	−0.36	0.04

因为在实际问题中, 观测到的污染物浓度是非负实数, 但是线性模型得到的污染物拟合值却有可能是负实数. 为了避免出现这一有悖常识的情形, 对于 $(i=2015, k=I)$, 我们在截距项 $\beta_{k,0}(w,s)$ 上加上一个正实数 $c(w,s)$, 其中 $c(w,s)$ 是由观测数据得到的满足 $\min_{t=1,\ldots,n_{ij}}[m_{k4}\{\boldsymbol{X}_{i4t}(s),s\}+c(w,s)]\geqslant 0$ 的最小正实数. 但是这样得到的污染物浓度的模拟数据会出现很大的数值, 为了与实际观测数据的范围保持一致, 我们又将所有的回归参数值除以 3, 即对于 $(i=2015, k=I)$, 将 (C.3) 中的回归系数进一步作如下变换: $\beta_{k,0}(w,s)$ 变换为 $\beta_{k,0}(w,s)+c(w,s)\}/3$, $\tilde{\boldsymbol{\beta}}_k(w,s)$ 变换为 $\tilde{\boldsymbol{\beta}}_k(w,s)/3$.

至此, 我们得到了第一个冬季 $(i=2015, k=I)$ 时模型 (C.1) 对应的回归系数的设定. 对于第二个冬季 $(i=2016, k=II)$, 对任意给定的站点 s 和风向 w, 我们在 $\beta_I(w,s)$ 上加上服从 7 维 Gauss 分布的白噪声向量, Gauss 分布的均值为 $\mathbf{0}$、协方差矩阵为主对角线等于 5、2、0.06、1、0.02、0.5 和 0.1 的对角矩阵, 得到了第二个冬季 $(i=2016, k=II)$ 对应的回归系数设定 $\beta_{II}(w,s)$. 对于这个系数设定, 我们计算了在 $(i=2016, k=II)$ 时 $[m_{k4}\{\boldsymbol{X}_{i4t}(s),s\}]_{t=1}^{n_{ij}}$ 的取值, 发现对于大部分时间点 $t=1,2,\ldots,n_{ij}$ 都能得到非负的浓度值且浓度值与 2016 年冬季实际污染物浓度数据的范围较一致. 如果在模拟数据中被抽到少数取值为负的时间点 t, 我们将令 $m_{k4}\{\boldsymbol{X}_{i4t}(s),s\}=0$. 换言之, 对于第二个冬季 $(i=2016, k=II)$, 在生成污染物浓度的模拟数据时, 我们将用 $m_{k4}\{\boldsymbol{X}_{i4t}(s),s\}I[m_{k4}\{\boldsymbol{X}_{i4t}(s),s\}\geqslant 0]$ 来替代 $m_{k4}\{\boldsymbol{X}_{i4t}(s),s\}$, 而由此得到的取值为 0 的情形是很少的.

为了得到污染物浓度的模拟数据, 由模型 (C.1), 对所关注的两个冬季, 我们还需要设定条件方差函数 $\sigma_{k4}^2(\boldsymbol{x},s)$ 和标准化残差序列 $\{\boldsymbol{e}_{k4t}=\{e_{k4t}(s_1),\ldots,e_{k4t}(s_{28})\}^{\mathrm{T}}\}_{t=1}^{\tilde{n}}$, $k=I,II$. 在正文中, 我们给出了基于观测数据的条件方差 $\sigma_{ij}(\boldsymbol{x},s)$ 的非参数核函数估计量 $\hat{\sigma}_{ij}(\boldsymbol{x},s)$. 在仿真模拟中, 我们设定 $\sigma_{k4}(\boldsymbol{x},s)=\hat{\sigma}_{i4}(\boldsymbol{x},s)$, 其中 i 等于 2015 和 2016 分别对应 k 等于 I 和 II. 标准化残差序列的模拟数据生成自向量值 1 阶时间自回归模型 $\boldsymbol{e}_{k4t}=\boldsymbol{A}_k\boldsymbol{e}_{k4,t-1}+\boldsymbol{u}_{k4t}$, 其中 $\boldsymbol{A}_k$ 为 28×28 维的矩阵, 噪声序列 $\boldsymbol{u}_{k4t}$ 是独立同分布的时间序列, 服从均值为 $\mathbf{0}$、协方差矩阵为 $\boldsymbol{\Omega}_k$ 的 28 维 Gauss 分布. 这里用标准化残差序列的估计值 $\{\hat{\boldsymbol{e}}_{i4t}=\{\hat{e}_{i4t}(s_1),\ldots,\hat{e}_{i4t}(s_{28})\}\}_{t=1}^{n_{i4}}$ 对自回归模型进行进行估计, 并将 $\boldsymbol{A}_k$ 和 $\boldsymbol{\Omega}_k$ 分别取为估计值, 其中 i 等于 2015 和 2016 分别对应 k 等于 I 和 II. 记标准化残差的模拟数据为 $e_{k4t}^*(s)$.

最终, $\mathrm{PM}_{2.5}$ 浓度的模拟数据 $\boldsymbol{Y}_{k4t}^*(s)$ 通过将气象模拟数据 $\boldsymbol{X}_{i4t}^*(s)$ 代入模型 (C.1) 得到, 即

$$Y_{k4t}^*(s)=m_{k4}\{\boldsymbol{X}_{i4t}^*(s),s\}+\sigma_{k4}\{\boldsymbol{X}_{i4t}^*(s),s\}\boldsymbol{e}_{k4t}^*(s),\quad t=1,\ldots,\tilde{n},\tag{C.4}$$

其中 $i=2015$ 对应于 $k=I$, $i=2016$ 对应于 $k=II$, 且回归函数 $m_{k4}(\boldsymbol{x},s)=m\{\boldsymbol{x};\boldsymbol{\beta}_k(w,s)\}$ 由模型 (C.1) 给出.

基于气象向量的模拟数据 $\boldsymbol{X}_{i4t}^*(s)$, $i=2010,2011,\ldots,2016$, 我们可以构造均衡的气象分布. 进一步, 可以得到在站点 s 和区域 $\mathcal{A}$ 的调整均值浓度的真实值 $\mu_{k4}(s)$ 和 $\mu_{k4}(\mathcal{A})$ 分别为

$$\mu_{k4}(s)=S^{-1}\Big(\sum_{a=1}^{A_4}n_{a4}\Big)^{-1}\sum_{s'\in\mathcal{W}}\sum_{a=1}^{A_4}\sum_{t=1}^{n_{a4}}m_{k4}\{\boldsymbol{X}_{a4t}(s'),s\}\quad\text{和}\quad\mu_{k4}(\mathcal{A})=|\mathcal{A}|^{-1}\sum_{s\in\mathcal{A}}\mu_{k4}(s).$$

在生成了 $\mathrm{PM}_{2.5}$ 浓度的模拟数据 $Y_{k4t}^*(s)$、气象向量的模拟数据 $\boldsymbol{X}_{i4t}^*(s)$、标准化残差序列的模拟数据 $e_{k4t}(s)$ 之后, 可以得到基于这些模拟数据的调整均值浓度的估计值. 记在站点 s 和区域 $\mathcal{A}$ 的调整均值浓度的估计值分别为 $\hat{\mu}_{k4}(s)$ 和 $\hat{\mu}_{k4}(\mathcal{A})$. 对于给定的两个样本量, 我们分别做了 1,000 次仿真模拟, 并计算了调整均值浓度的估计量 $\hat{\mu}_{k4}(s)$ 和 $\hat{\mu}_{k4}(\mathcal{A})$ 的标准误差及均方根误差.

表 C2 给出了在选定的时间段内 28 个污染物监测站点、5 个子区域和整个研究区域的调整均值浓度估计量的均方根误差和标准误差. 从表中可知, 对于选定的两个冬季, 当样本量增加时, 单站点和

表 C2 在仿真模拟中, 对于选定的两个冬季, 在 28 个污染物监测站点、5 个子区域内和整个区域上调整均值估计量的均方根误差 (root mean squared error) 和标准误差 (standard error). 黑色数据表示样本量

	冬季 I				冬季 II			
	均方根误差		标准误差		均方根误差		标准误差	
区域/站点	**1080**	**2160**	**1080**	**2160**	**1080**	**2160**	**1080**	**2160**
整个区域	1.04	0.67	0.94	0.63	0.94	0.73	0.93	0.62
东北区域	1.09	0.70	0.88	0.57	0.96	0.77	0.93	0.61
东四	1.35	0.95	0.95	0.60	0.93	0.71	0.91	0.59
东四环	1.07	0.71	0.87	0.56	1.18	0.93	1.10	0.73
美国大使馆	1.26	0.80	0.93	0.58	0.87	0.66	0.87	0.59
农展馆	1.02	0.66	0.86	0.57	1.08	0.94	0.97	0.62
顺义	1.10	0.69	1.08	0.69	1.12	0.82	1.09	0.71
西北区域	0.92	0.61	0.92	0.61	1.11	0.96	0.89	0.59
奥体中心	1.17	0.78	0.91	0.59	0.91	0.73	0.88	0.57
北部新区	1.06	0.70	1.01	0.69	1.47	1.33	1.07	0.71
昌平	0.85	0.59	0.84	0.56	1.15	0.99	0.73	0.45
官园	0.97	0.61	0.89	0.58	1.09	0.89	1.09	0.69
古城	1.04	0.68	1.04	0.67	1.50	1.34	1.10	0.77
门头沟	2.15	1.30	2.01	1.08	1.04	0.77	0.97	0.64
万柳	1.23	0.85	1.22	0.85	1.67	1.33	1.43	0.98
西直门北	1.35	0.98	1.32	0.94	1.93	1.50	1.56	1.12
植物园	1.02	0.73	1.02	0.72	1.27	0.90	1.23	0.83
东南区域	1.19	0.76	0.96	0.64	1.02	0.63	0.96	0.63
前门	1.10	0.69	1.01	0.68	1.16	0.73	1.10	0.71
天坛	0.98	0.60	0.93	0.60	0.95	0.69	0.94	0.62
通州	1.76	1.31	1.17	0.83	1.30	0.77	1.11	0.74
亦庄	1.41	0.85	1.07	0.69	1.49	1.03	0.93	0.61
永定门	1.18	0.78	1.05	0.71	1.31	1.06	1.21	0.79
西南区域	1.12	0.71	1.02	0.69	0.98	0.76	0.98	0.66
大兴	1.55	0.97	1.19	0.78	1.49	0.90	1.02	0.67
房山	1.48	0.96	1.21	0.84	1.11	0.82	1.10	0.72
丰台	1.15	0.80	1.15	0.78	1.08	0.77	1.08	0.71
南三环	1.34	0.84	1.18	0.79	1.35	1.05	1.31	0.92
万寿西宫	1.08	0.69	1.04	0.69	1.17	0.93	1.16	0.75
云冈	0.94	0.67	0.93	0.64	1.44	1.30	1.05	0.71
南部区域	1.57	1.03	1.27	0.86	1.40	0.87	1.26	0.86
琉璃河	1.86	1.30	1.48	1.00	1.62	1.23	1.61	1.10
永定门	1.73	1.07	1.60	1.05	1.60	0.97	1.42	0.96
榆垡	1.49	1.04	1.15	0.77	1.73	1.14	1.11	0.74

区域内的调整均值浓度估计量的均方根误差和标准误差均有明显下降. 这印证了调整均值浓度的估计

量的相合性. 在第一个冬季和第二个冬季, 28 个污染物监测站点调整均值浓度估计量的平均值分别等于 155 和 187 $\mu g/m^3$. 与其相比, 调整均值浓度估计量的均方根误差和标准误差则要小很多. 位于北京市南部的三个子区域 (东南部、西南部、南部靠近河北的地区) 的调整均值浓度估计量比其他北部子区域具有更大的均方根误差. 这是由于南部的 $PM_{2.5}$ 浓度本身高于北部, 因此, 南部的污染物的模拟数据也具有更大的变异性.

Regional air-quality assessment that adjusts for meteorological confounding

Shuyi Zhang, Songxi Chen, Bin Guo, Hengfang Wang & Wei Lin

Abstract Although air pollution is caused by emission of pollutants to the atmosphere, the observed pollution levels are confounded by meteorological conditions, which largely determine the dispersion of the pollutants. Hence, effective air-quality management requires the evaluation index and statistical measures that are immune to meteorological confounding and reflect changes in pollutant concentrations accurately and objectively. Motivated by the task of assessing changes in the underlying emission in a region near Beijing, we propose a spatial and temporal adjustment approach to remove meteorological confounding. The adjusted average pollutant concentration over space and time can capture changes in the underlying emission by controlling the meteorological variation. Estimation of the adjusted average is proposed together with theoretical and numerical analysis. We apply the approach to conducting air-quality assessments in the Beijing region, which reveals some intriguing patterns and trends that are useful for air-quality management.

Keywords **air-quality assessment, meteorological confounding, nonparametric regression, spatio-temporal adjustment**

MSC(2010) **62-07, 62P12, 62G08**

doi: 10.1360/SCM-2019-0368