Nonparametric estimation of the spatio-temporal covariance structure

Article *in* Statistics in Medicine · July 2019

DOI: 10.1002/sim.8315

2 authors:

Kai Yang
University of Florida
13 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Peihua Qiu
University of Florida
190 PUBLICATIONS **4,103** CITATIONS
SEE PROFILE

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Nonparametric Estimation of the Spatio-Temporal Covariance Structure

Kai Yang | Peihua Qiu*

Department of Biostatistics, University of Florida, Gainesville, Florida, USA

Correspondence
Peihua Qiu, Department of Biostatistics, University of Florida, Gainesville, Florida, USA.
Email: pqiu@ufl.edu

Abstract

Spatio-temporal modelling is an active research problem with broad applications. In this problem, proper description and estimation of the data covariance structure plays an important role. In the literature, most available methods assume that the data covariance is stationary and follows a specific parametric form. In practice, however, such assumptions are hardly valid or difficult to verify. In this paper, we propose a new and flexible method for estimating the underlying covariance structure. Our proposed method does not require the covariance to be stationary or follow a parametric form. It can accommodate nonparametric space-time-varying mean structure of the observed data. Under some mild regularity conditions, it is shown that our estimated covariance structure converges to the true covariance structure. The proposed method is also justified numerically by a simulation study and an application to a hand, foot and mouth disease data.

KEYWORDS:
consistency, covariance estimation, local smoothing, nonstationarity, spatio-temporal data

1 | INTRODUCTION

Spatio-temporal data modelling and analysis is an active interdisciplinary research problem with broad applications in different fields, including geography, environment, public health, and medicine. There have been many existing methods proposed for solving this problem. See, for instance, Cressie and Wikle,[1] Diggle,[2] Gneiting and Guttorp,[3] Schabenberger and Gotway,[4] and Yang and Qiu.[5] To analyze spatio-temporal data properly, one critical issue is to describe and estimate the underlying covariance structure of the observed data properly, which is the focus of this paper. [1]

In the literature, there has been some discussion about estimation of the covariance structure of spatio-temporal data.[6–9] However, most existing methods assume that the covariance structure is separable and/or stationary in space and time. The separability assumption can simplify computation greatly, but it may not be valid in many applications, because of the complicated space-time interaction. In addition, it can be observed in many real spatio-temporal datasets that the sample covariances change significantly over both space and time, indicating that the stationarity assumption may not be appropriate in such cases. To model nonstationary covariance structure, Hsu et al[10] suggested a method based on the assumptions that the related covariance function had a complex parametric form, the nonstationarity is in the spatial domain only, and the possible nonstationarity in the time domain can be ignored. Recently, Shand and Li[11] proposed a different method for modelling nonstationary covariance structure, by which a nonstationary process was assumed to be a projection of a stationary process in a higher-dimensional

[1]Data available on request from the authors.

space. This method can model the nonstationarity in both space and time; but, estimation of the extra dimensions depends on the choice of the parametric covariance model.

In this paper, we suggest a flexible method to effectively estimate the underlying covariance structure of a spatio-temporal dataset. By this method, the mean function of the spatio-temporal response is first estimated by a local smoothing approach. Then, the covariance function is estimated nonparametrically from the residuals. Our proposed method is described in detail in Sections 2 and 3. Its numerical performance is evaluated in Section 4, in comparison with two state-of-the-art existing methods. The method is demonstrated in a real-data example in Section 5. Several remarks conclude the article in Section 6. The proofs of three theorems are given in the supplementary file.

2 | SPATIO-TEMPORAL MODEL AND ITS ESTIMATION

2.1 | The model and its estimation

Suppose there are n observation times $\{t_i = i/n, i = 1, \dots, n\}$ and m_i observation locations $\{\boldsymbol{s}_{ij}, j = 1, \dots, m_i\}$ in the spatial region $\Omega \subseteq \mathbb{R}^2$ at the ith time, for $i = 1, \dots, n$. The observed spatio-temporal observations $\{y(t_i, \boldsymbol{s}_{ij}), j = 1, \dots, m_i, i = 1, \dots, n\}$ are assumed to follow the model

$$y(t_i, \boldsymbol{s}_{ij}) = \lambda(t_i, \boldsymbol{s}_{ij}) + \varepsilon(t_i, \boldsymbol{s}_{ij}), \text{ for } j = 1, \dots, m_i, i = 1, \dots, n, \tag{1}$$

where $\lambda(t_i, \boldsymbol{s}_{ij})$ is the mean of $y(t_i, \boldsymbol{s}_{ij})$, and $\varepsilon(t_i, \boldsymbol{s}_{ij})$ is the zero-mean random error. Under model (1), the variance-covariance structure of the spatio-temporal data can be described by

$$E\left[\varepsilon(t, \boldsymbol{s})\varepsilon(t', \boldsymbol{s}')\right] = \begin{cases} \sigma^2(t, \boldsymbol{s}), & \text{if } (t, \boldsymbol{s}) = (t', \boldsymbol{s}'), \\ V(t, t'; \boldsymbol{s}, \boldsymbol{s}'), & \text{otherwise,} \end{cases} \tag{2}$$

where $t, t' \in [0, 1]$, and $\boldsymbol{s}, \boldsymbol{s}' \in \Omega$. Note that observation times in practice are often integer numbers in the unit of year, month, day, or others. In the above expressions, the mean $\lambda(t, \boldsymbol{s})$ and the variance/covariance $\sigma^2(t, \boldsymbol{s})$ and $V(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ are all formulated in terms of the rescaled time $t_i = i/n \in [0, 1]$, rather than i. This formulation is commonly used in the literature.[12–14] The rescaled time is necessary to use for studying asymptotic properties of the estimated mean and variance/covariance functions in the time domain. Otherwise, the distance between two consecutive observation times is 1, which does not go to 0 when n increases, and thus the asymptotic properties in time cannot be discussed properly.

To estimate the mean function $\lambda(t, \boldsymbol{s})$, for $(t, \boldsymbol{s}) \in [0, 1] \times \Omega$, a natural and computationally convenient method is to use the following local linear kernel (LLK) smoothing procedure:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^4}{\arg\min} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left[y(t_i, \boldsymbol{s}_{ij}) - \boldsymbol{\beta}^T \mathbf{X}_{ij}\right]^2 w_0(i, j), \tag{3}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_t, \boldsymbol{\beta}_{\boldsymbol{s}}^T)^T = (\beta_0, \beta_t, \beta_{s,1}, \beta_{s,2})^T$ is the coefficient vector, $\mathbf{X}_{ij} = (1, t_i - t, (\boldsymbol{s}_{ij} - \boldsymbol{s})^T)^T$, $w_0(i, j) = K_1\left((t_i - t)/h_1\right) K_2\left(d_E(\boldsymbol{s}_{ij}, \boldsymbol{s})/h_2\right)$, $d_E(\cdot, \cdot)$ is the Euclidean distance, $K_1(\cdot)$ and $K_2(\cdot)$ are two univariate kernel functions, and $h_1 > 0$ and $h_2 > 0$ are two bandwidths. The solution of (3) to β_0 is then the LLK estimator of $\lambda(t, \boldsymbol{s})$, which has the expression

$$\widehat{\lambda}(t, \boldsymbol{s}) = \boldsymbol{e}_1^T \left(\mathbf{X}^T \mathbf{W}_0 \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}_0 \mathbf{Y}, \tag{4}$$

where $\boldsymbol{e}_1 = (1, 0, 0, 0)^T$, $\mathbf{X} = (\mathbf{X}_{11}, \dots, \mathbf{X}_{nm_n})^T$ is the design matrix, $\mathbf{W}_0 = \text{diag}\{w_0(1, 1), \dots, w_0(n, m_n)\}$, and $\mathbf{Y} = (y(t_1, \boldsymbol{s}_{11}), \dots, y(t_n, \boldsymbol{s}_{nm_n}))^T$ is the vector of all observations. The two kernel functions used in (3) are usually chosen to have finite supports. Therefore, the estimator in (4) is a weighted average of observations in a neighborhood of $(t, \boldsymbol{s})$, the neighborhood size is controlled by the bandwidths, and the weights are controlled by the kernel functions.

After the estimator $\widehat{\lambda}(t, \boldsymbol{s})$ is obtained, estimation of the variance and covariance functions $\sigma^2(t, \boldsymbol{s})$ and $V(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ can be discussed as follows. First, the variance function $\sigma^2(t, \boldsymbol{s})$ can be estimated by

$$\widehat{\sigma}^2(t, \boldsymbol{s}) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \widehat{\varepsilon}^2(t_i, \boldsymbol{s}_{ij}) w_1(i, j)}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_1(i, j)}, \tag{5}$$

where $\widehat{\varepsilon}(t_i, \boldsymbol{s}_{ij}) = y(t_i, \boldsymbol{s}_{ij}) - \widehat{\lambda}(t_i, \boldsymbol{s}_{ij})$, for $j = 1, ..., m_i,\ i = 1, ..., n$, and $w_1(i, j) = K_1\left((t_i - t)/h_3\right) K_2\left(d_E(\boldsymbol{s}_{ij}, \boldsymbol{s})/h_4\right)$. Second, for two different points $(t, \boldsymbol{s}), (t', \boldsymbol{s}') \in [0, 1] \times \Omega$, $V(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ can be estimated by

$$\widehat{V}(t, t'; \boldsymbol{s}, \boldsymbol{s}') = \frac{\sum_{i,j} \sum_{(k,l)\neq(i,j)} \widehat{\varepsilon}(t_i, \boldsymbol{s}_{ij})\widehat{\varepsilon}(t_k, \boldsymbol{s}_{kl}) w_2(i, j, k, l)}{\sum_{i,j} \sum_{(k,l)\neq(i,j)} w_2(i, j, k, l)}, \tag{6}$$

where $w_2(i, j, k, l) = K_1((t_i - t)/h_3)K_1((t_k - t')/h_3)K_2(d_E(\boldsymbol{s}_{ij}, \boldsymbol{s})/h_4)K_2(d_E(\boldsymbol{s}_{kl}, \boldsymbol{s}')/h_4)$ if $|t_k - t_i| \in (|t - t'| - 1/n, |t - t'| + 1/n)$ and 0 otherwise. In (6), $\widehat{V}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ is actually defined as a weighted sample covariance computed from pairs of the residuals $\{\widehat{\varepsilon}(t_i, \boldsymbol{s}_{ij})\}$ in the related neighborhoods of $(t, \boldsymbol{s})$ and $(t', \boldsymbol{s}')$, and the weights are determined by the kernel functions. In (5) and (6), the two bandwidths (h_3, h_4) could be chosen differently from the bandwidths (h_1, h_2) used in estimating the mean function in (4).

2.2 | Modification of the estimated covariance matrix

Let $\boldsymbol{\varepsilon}$ be a long vector with the elements $\{\varepsilon(t_i, \boldsymbol{s}_{ij}), j = 1, 2, ..., m_i, i = 1, 2, ..., n\}$, $\mathbf{V}$ be its covariance matrix, and $\widehat{\mathbf{V}}$ be the estimate of $\mathbf{V}$ obtained from the estimated variance/convariance functions in (5) and (6). Then, $\widehat{\mathbf{V}}$ may not be a positive semidefinite matrix; thus, it may not be a legitimate covariance matrix. In the literature, there are existing modifications to make a symmetric matrix positive semidefinite (e.g., Hall, Fisher and Hoffmann,[15] Higham[16]). Some of them require assumptions that are not made in this paper. For instance, the method by Hall, Fisher and Hoffmann[15] is mainly for stationary processes, and its computation is extensive. Here, we suggest using the one by Higham[16], described briefly below. Let $\|\cdot\|_F$ be the Frobenius matrix norm, defined to be the square root of the sum of squares of a matrix's elements, and $\mathcal{P}$ be the set of all symmetric positive semidefinite matrices of the same dimensions as those of $\widehat{\mathbf{V}}$. Then, we consider the projection of $\widehat{\mathbf{V}}$ on $\mathcal{P}$, which is the solution of the problem

$$\widetilde{\mathbf{V}} = \underset{\mathbf{P}\in\mathcal{P}}{\arg\min} \|\mathbf{P} - \widehat{\mathbf{V}}\|_F.$$

It can be shown that the solution is $(\widehat{\mathbf{V}} + \widehat{\mathbf{V}}_p)/2$, where $\widehat{\mathbf{V}}_p$ is the symmetric polar factor of $\widehat{\mathbf{V}}$.[16] So, the modified covariance matrix is $\widetilde{\mathbf{V}} = (\widehat{\mathbf{V}} + \widehat{\mathbf{V}}_p)/2$.

2.3 | Selection of the kernel functions and bandwidths

In the mean and variance/covariance estimation procedures (4)-(6), the kernel functions $K_1(\cdot)$ and $K_2(\cdot)$ and the bandwidths h_1, h_2, h_3 and h_4 should be chosen properly in advance. In the kernel smoothing literature, it has been shown that the Epanechnikov kernel function has some good theoretical properties, see Epanechnikov.[17] So, both $K_1(\cdot)$ and $K_2(\cdot)$ are chosen to be that function. Namely, $K_1(x) = K_2(x) = 0.75(1 - x^2)\mathrm{I}(|x| \leq 1)$. Because $K_2(\cdot)$ is used in a 2-dimensional space, its normalization constant should be chosen differently from 0.75 to become a density kernel. However, this constant will be cancelled out in the estimators defined in (4)-(6). Thus, its normalization constant does not need to be chosen separately.

The bandwidths h_1 and h_2 are used in estimating the mean function $\lambda(t, \boldsymbol{s})$. In the literature on nonparametric estimation of regression functions from correlated data, it has been well discussed that the bandwidths selected by the conventional cross-validation (CV) procedure like the leave-one-out cross-validation (LOOCV) would not perform well, because it cannot properly distinguish the data correlation structure from the data mean function; see Altman[18] and Opsomer et al.[19] To overcome this difficulty, Brabanter et al.[20] suggested a modified CV (MCV) procedure in the univariate regression setup, by which h_1 and h_2 can be chosen by minimizing the following MCV score:

$$\mathrm{MCV}(h_1, h_2) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m_i}\sum_{j=1}^{m_i}\left\{\widehat{\lambda}_{-(ij)}(t_i, \boldsymbol{s}_{ij}) - y(t_i, \boldsymbol{s}_{ij})\right\}^2\right], \tag{7}$$

where $\widehat{\lambda}_{-(ij)}(t_i, \boldsymbol{s}_{ij})$ is the leave-one-out estimate of $\lambda(t, \boldsymbol{s})$, obtained by (4) when the observation $y(t_i, \boldsymbol{s}_{ij})$ is omitted in the computation and $K_1(\cdot)$ and $K_2(\cdot)$ are both chosen to be

$$K_\epsilon(u) = \frac{4}{4 - 3\epsilon - \epsilon^3}\begin{cases} \frac{3}{4}(1 - u^2)\mathrm{I}(|u| \leq 1), & \text{if } |u| \geq \epsilon, \\ \frac{3(1-\epsilon^2)}{4\epsilon}|u|, & \text{otherwise,} \end{cases} \tag{8}$$

where $\epsilon \in (0,1)$ is a constant. By (7), observations around $(t_i, \boldsymbol{s}_{ij})$ are down-weighted when computing $\widehat{\lambda}_{-(ij)}(t_i, \boldsymbol{s}_{ij})$ to reduce the impact of data correlation on bandwidth selection. Regarding ϵ, Brabanter et al.[20] suggested choosing it to be 0.1, based on a large simulation study. This suggestion is adopted here.

For selection of h_3 and h_4, we suggest a new method based on spatio-temporal prediction. More specifically, we first define the cross-validation mean squared prediction error (MSPE) by

$$\text{MSPE}(h_3, h_4) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{m_i}\sum_{j=1}^{m_i}\big\{y(t_i, \boldsymbol{s}_{ij}) - \widehat{y}_{-(ij)}(t_i, \boldsymbol{s}_{ij})\big\}^2\Big], \tag{9}$$

where the predicted values $\{\widehat{y}_{-(ij)}(t_i, \boldsymbol{s}_{ij}), j = 1, \ldots, m_i, i = 1, \ldots, n\}$ are obtained by the simple kriging method (Cressie[1]), described below. For each $1 \le j \le m_i$ and $1 \le i \le n$, let $\widehat{V}_{-(ij)}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ be the estimated covariance function by (6) when the (i,j)th residual $\widehat{\varepsilon}(t_i, \boldsymbol{s}_{ij})$ is omitted, $\boldsymbol{\varepsilon}_{-(ij)}$ be the vector with elements $\{\varepsilon(t_k, \boldsymbol{s}_{kl}), l = 1, \ldots, m_k, k = 1, \ldots, n, (k,l) \neq (i,j)\}$, $\mathbf{V}_{ij,-(ij)}$ be the covariance between $\varepsilon(t_i, \boldsymbol{s}_{ij})$ and $\boldsymbol{\varepsilon}_{-(ij)}$, and $\mathbf{V}_{-(ij),-(ij)}$ be the covariance matrix of $\boldsymbol{\varepsilon}_{-(ij)}$. Then, the predicted values $\{\widehat{y}_{-(ij)}(t_i, \boldsymbol{s}_{ij}), j = 1, \ldots, m_i, i = 1, \ldots, n\}$ are defined by

$$\widehat{y}_{-(ij)}(t_i, \boldsymbol{s}_{ij}) = \widehat{\lambda}(t_i, \boldsymbol{s}_{ij}) + \widehat{\mathbf{V}}^T_{ij,-(ij)}\widehat{\mathbf{V}}^{-1}_{-(ij),-(ij)}\widehat{\boldsymbol{\varepsilon}}_{-(ij)}, \tag{10}$$

where $\widehat{\mathbf{V}}_{ij,-(ij)}$ and $\widehat{\mathbf{V}}_{-(ij),-(ij)}$ are estimates of $\mathbf{V}_{ij,-(ij)}$ and $\mathbf{V}_{-(ij),-(ij)}$, respectively, computed from $\widehat{V}_{-(ij)}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$, and $\widehat{\boldsymbol{\varepsilon}}_{-(ij)}$ is a long vector of the residuals $\{\widehat{\varepsilon}(t_k, \boldsymbol{s}_{kl}), l = 1, \ldots, m_k, k = 1, \ldots, n, (k,l) \neq (i,j)\}$ arranged in the same order as those in $\boldsymbol{\varepsilon}_{-(ij)}$. The bandwidths h_3 and h_4 can then be selected by minimizing $\text{MSPE}(h_3, h_4)$.

Let $N = \sum_{i=1}^{n} m_i$ be the total number of observations, note that $\widehat{\mathbf{V}}_{ij,-(ij)}$ is a $(N-1)$-dimensional vector, and $\widehat{\mathbf{V}}_{-(ij),-(ij)}$ is a $(N-1)\times(N-1)$ matrix. So, when the total sample size N is large, the computation and storage for obtaining the inverse matrix $\widehat{\mathbf{V}}^{-1}_{-(ij),-(ij)}$ could be demanding. To overcome this difficulty, we suggest using the observations in a local neighborhood of $(t_i, \boldsymbol{s}_{ij})$ only when computing the predicted value $\widehat{y}_{-(ij)}(t_i, \boldsymbol{s}_{ij})$ in the kriging procedure (10). More specifically, let $\Delta_{ij}(\theta_t, \theta_{s,ij}) = \{(k,l) : |t_k - t_i| \le \theta_t, d_E(\boldsymbol{s}_{kl}, \boldsymbol{s}_{ij}) \le \theta_{s,ij}, (k,l) \neq (i,j)\}$ be a set of indices around (i,j). Then, when computing $\widehat{\mathbf{V}}_{ij,-(ij)}$ and $\widehat{\mathbf{V}}_{-(ij),-(ij)}$ used in (10), only those points $(t_k, \boldsymbol{s}_{kl})$ whose indices are included in $\Delta_{ij}(\theta_t, \theta_{s,ij})$ are used. Also, $\widehat{\boldsymbol{\varepsilon}}_{-(ij)}$ in (10) needs to be replaced by the one that includes the residuals with indices in $\Delta_{ij}(\theta_t, \theta_{s,ij})$ only. Then, the computation involved in calculating the predicted value $\widehat{y}_{-(ij)}(t_i, \boldsymbol{s}_{ij})$ is greatly reduced. The amount of computational reduction is controlled by the parameters θ_t and $\theta_{s,ij}$. Generally speaking, if their values are chosen smaller, then the computational reduction would be more substantial; but, the resulting predicted value $\widehat{y}_{-(ij)}(t_i, \boldsymbol{s}_{ij})$ could be less accurate. Based on extensive numerical studies, we suggest choosing θ_t and $\theta_{s,ij}$ as follows. For each $1 \le j \le m_i$ and $1 \le i \le n$, we first compute $\theta(i,j) = \min\{d_E(\boldsymbol{s}_{ij}, \boldsymbol{s}_{il}), l = 1, \ldots, m_i, l \neq j\}$. Then, θ_t and $\theta_{s,ij}$ are chosen such that $\theta_t \ge 5/n$ and $\theta_{s,ij} \ge 3\theta(i,j)$.

3 | STATISTICAL PROPERTIES

In this section, we present some statistical properties of the estimators $\widehat{\lambda}(t, \boldsymbol{s})$, $\widehat{\sigma}^2(t, \boldsymbol{s})$ and $\widehat{V}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ defined in (4)-(6). In the discussion, it is assumed that $\{m_i, i = 1, \ldots, n\}$ are in the same order of an integer number m. It is further assumed that $\{\boldsymbol{s}_{ij}, i = 1, \ldots, n, j = 1, \ldots, m_i\}$ follow a distribution with the density $f(\boldsymbol{s})$, for $\boldsymbol{s} \in \Omega$, and at each time point t_i, the corresponding spatial locations $\{\boldsymbol{s}_{ij}, j = 1, \ldots, m_i\}$ are independent. Furthermore, $\{\boldsymbol{s}_{ij}, j = 1, \ldots, m_i, i = 1, \ldots, n\}$ are assumed to be independent of the random errors $\{\varepsilon(t_i, \boldsymbol{s}_{ij}), j = 1, \ldots, m_i, i = 1, \ldots, n\}$ in model (1). All these assumptions on the spatial locations are denoted as *(Assumption-SL)*. For the random error $\{\varepsilon(t_i, \boldsymbol{s}_{ij}), j = 1, \ldots, m_i, i = 1, \ldots, n\}$, the strong mixing coefficient in the time domain is defined as

$$\alpha(k) = \sup_{n \ge 1, 1 \le i \le n-k} \sup_{A,B} \left\{|P(AB) - P(A)P(B)| : A \in \mathcal{F}_1^i, B \in \mathcal{F}_{i+k}^n\right\},$$

where $\mathcal{F}_{k_0}^{k_1}$ is the σ-algebra generated by $\{\varepsilon(t_k, \boldsymbol{s}_{kl}), k_0 \le k \le k_1, l = 1, \ldots, m_k\}$. Then, we have the following results.

Theorem 1. In model (1), it is assumed that Ω is a bounded and closed set in $\mathbb{R}^2$, the observation locations $\{\boldsymbol{s}_{ij}, j = 1, \ldots, m_i, i = 1, \ldots, n\}$ satisfy the assumptions (Assumption-SL), $f(\boldsymbol{s})$ is twice continuously differentiable and it has a non-zero lower bound in Ω, the mean function $\lambda(t, \boldsymbol{s})$ is twice continuously differentiable in $[0,1]\times\Omega$, the strong mixing coefficient $\alpha(k)$ satisfies the condition that $\alpha(k) \le C_0 \exp(-C_1 k)$, where $C_0 > 0$ and $C_1 > 0$ are two constants, and there are some constants $\delta > 5$ and $0 \le C_\varepsilon < \infty$ such that $E|\varepsilon(t_i, \boldsymbol{s}_{ij})|^\delta \le C_\varepsilon$, for all $1 \le j \le m_i$ and $1 \le i \le n$, the kernel functions $K_1(\cdot)$ and $K_2(\cdot)$ are bounded, symmetric, Lipschitz-1 continuous density functions with finite supports, $1/(nh_1) = o(1)$, $\log(n)/(mh_2^2) = O(1)$, $h_1 = o(1)$, and

$h_2 = o(1)$. Then, for any $(t, \boldsymbol{s}) \in [0,1] \times \Omega$, we have

$$\left|\widehat{\lambda}(t,\boldsymbol{s}) - \lambda(t,\boldsymbol{s})\right| = O_p\left(h_1^2 + h_2^2 + \left\{1/(nh_1)\right\}^{1/2}\right). \tag{11}$$

Theorem 2. Besides the conditions in Theorem 1, if we further assume that $\log(n)^2/(nh_1^2) = o(1)$, and $h_1/h_2 = O(1)$, then we have

$$\sup_{t\in[0,1],\boldsymbol{s}\in\Omega}\left|\widehat{\lambda}(t,\boldsymbol{s}) - \lambda(t,\boldsymbol{s})\right| = O_p\left(h_1^2 + h_2^2 + \{\log^2(n)/(nh_1^2)\}^{1/2}\right). \tag{12}$$

Theorem 3. Besides the conditions in Theorem 2, if we further assume that both the variance function $\sigma^2(t,\boldsymbol{s})$ and the covariance function $V(t,t';\boldsymbol{s},\boldsymbol{s}')$ are twice continuously differentiable, $h_3 = o(1)$, $h_4 = o(1)$, $1/(nh_3) = o(1)$, and $\log(n)/(mh_4^2) = O(1)$, then

(i) for any $(t,\boldsymbol{s}) \in [0,1] \times \Omega$, it holds that

$$\begin{aligned}\left|\widehat{\sigma}^2(t,\boldsymbol{s}) - \sigma^2(t,\boldsymbol{s})\right| = O_p\big(&h_1^2 + h_2^2 + \{\log^2(n)/(nh_1^2)\}^{1/2} \\ &+h_3^2 + h_4^2 + \{1/(nh_3)\}^{1/2}\big),\end{aligned} \tag{13}$$

(ii) for any $\boldsymbol{s}, \boldsymbol{s}' \in \Omega$, and $t, t' \in [0,1]$ such that $n(t'-t) = \varpi + o(1)$, where $\varpi \geq 0$ is a constant, we have

$$\begin{aligned}\left|\widehat{V}(t,t';\boldsymbol{s},\boldsymbol{s}') - V(t,t';\boldsymbol{s},\boldsymbol{s}')\right| = O_p\big(&h_1^2 + h_2^2 + \{\log^2(n)/(nh_1^2)\}^{1/2} \\ &+h_3^2 + h_4^2 + \{1/(nh_3)\}^{1/2}\big).\end{aligned} \tag{14}$$

Remark: In some applications, the spatial locations do not change over time. In such cases, the number of spatial locations $\{m_i, i = 1,\ldots,n\}$ at different time points are the same to be m, and the spatial locations can be simply denoted as $\{\boldsymbol{s}_j, j = 1\ldots,m\}$. In such cases, let (Assumption-SL*) denote the assumptions that $\{\boldsymbol{s}_j, j = 1,\ldots,m\}$ are independent and identically distributed with a density function $f(\boldsymbol{s})$, for $\boldsymbol{s} \in \Omega$, and they are independent of the random errors $\{\varepsilon(t_i,\boldsymbol{s}_{ij}), j = 1,\ldots,m, i = 1,\ldots,n\}$ in model (1). Then, it is clear that all assumptions in (Assumption-SL) are satisfied if (Assumption-SL*) is true in the special cases discussed here. Consequently, the results in Theorems 1-3 are still true if (Assumption-SL) is replaced by (Assumption-SL*) in such cases.

The proofs of Theorems 1-3 are given in a supplementary file.

4 | NUMERICAL STUDY

In this section, we present some simulation results about the numerical performance of the proposed method described in the previous sections. For simplicity, let the observation times be $\{t_i = i/n, i = 1,\ldots,n\}$, the observation locations at each time point be equally spaced in $\Omega = [0,1]\times[0,1]$ with the same number m, and they do not change over time. In such a setup, for simplicity, we will use $\{\boldsymbol{s}_j, j = 1,\ldots,m\}$ to denote the observation locations. In the simulation study, (m,n) are chosen to be $(64,200)$ or $(100,500)$, and the mean function is chosen to be

$$\lambda(t,\boldsymbol{s}) = 1.5 + e^{-(s_x^2+s_y^2)} + \cos(2\pi t), \text{ for } t \in [0,1],\ \boldsymbol{s} \in \Omega,$$

where $\boldsymbol{s} = (s_x, s_y)^T$. Two different scenarios when $\lambda(t,\boldsymbol{s})$ is assumed known or unknown are considered. In the scenario when $\lambda(t,\boldsymbol{s})$ is assumed unknown, it needs to be estimated by (4). For the spatio-temporal data correlation structure, the following three cases are considered:

Case 1: The random errors $\{\varepsilon(t_i,\boldsymbol{s}_j)\}$ are generated from a non-stationary spatio-temporal model, as discussed in Hsu et al.[10] More specifically, it is assumed that

$$\varepsilon(t_i,\boldsymbol{s}_j) = W(t_i,\boldsymbol{s}_j) + \sum_{\gamma=1}^{\Upsilon}\xi_\gamma(t_i)\Psi_\gamma(\boldsymbol{s}_j), \text{ for } i = 1,\ldots,n,\ j = 1,\ldots,m,$$

where $W(t_i,\boldsymbol{s}_j)$ is a zero-mean stationary spatio-temporal process, $\{\Psi_\gamma(\boldsymbol{s}), \gamma = 1,\ldots,\Upsilon\}$ are pre-specified basis functions, and $\{\xi_\gamma(t_i), i = 1,\ldots,n\}$, for $\gamma = 1,\ldots,\Upsilon$, are mutually independent zero-mean stationary time series that are independent of $W(t_i,\boldsymbol{s}_j)$. The stationary process $W(t_i,\boldsymbol{s}_j)$ is generated from an AR(1) process $W(t_i,\boldsymbol{s}_j) = \phi_w W(t_{i-1},\boldsymbol{s}_j) + (1-\phi_w^2)^{1/2}\eta_w(t_i,\boldsymbol{s}_j)$, where $|\phi_w| < 1$ is a constant, and $\eta_w(t_i,\cdot)$ are temporally independent spatial processes whose spatial correlation are described by $\mathrm{Cov}(\eta_w(t_i,\boldsymbol{s}_j),\eta_w(t_i,\boldsymbol{s}_l)) = \rho_w(d_E(\boldsymbol{s}_j,\boldsymbol{s}_l))$, for any i, j, and l. At each $(t_i,\boldsymbol{s}_j)$, $\eta_w(t_i,\boldsymbol{s}_j)$ has the

$N(0,1)$ distribution. Let $\boldsymbol{\xi}(t_i) = (\xi_1(t_i), \ldots, \xi_\Upsilon(t_i))^T$. Then, $\boldsymbol{\xi}(t_i)$ is generated from the following stationary AR(1) process: $\boldsymbol{\xi}(t_i) = \Phi\boldsymbol{\xi}(t_{i-1}) + \boldsymbol{\eta}(t_i)$, where $\Phi = \text{diag}\{\phi_1, \ldots, \phi_\Upsilon\}$ is a $\Upsilon \times \Upsilon$ diagonal matrix, $|\phi_\gamma| < 1$, for $\gamma = 1, ..., \Upsilon$, $\boldsymbol{\eta}(t_i)$ are temporally independent random vectors with the common distribution $N_\Upsilon(\mathbf{0}, \Sigma)$, and $\Sigma = \text{diag}\{1-\phi_1^2, \ldots, 1-\phi_\Upsilon^2\}$. Note that the random process $\{\varepsilon(t_i, \boldsymbol{s}_j)\}$ is stationary in the time domain, and it can be checked that the covariance between $\varepsilon(t_i, \boldsymbol{s}_j)$ and $\varepsilon(t_k, \boldsymbol{s}_l)$ has the expression

$$V(t_i, t_k; \boldsymbol{s}_j, \boldsymbol{s}_l) = \phi_w^{|k-i|}\rho_w(d_E(\boldsymbol{s}_j, \boldsymbol{s}_l)) + \sum_{\gamma=1}^{\Upsilon} \phi_\gamma^{|k-i|}\Psi_\gamma(\boldsymbol{s}_j)\Psi_\gamma(\boldsymbol{s}_l).$$

In the case when $t_i = t_k$ and $\boldsymbol{s}_j = \boldsymbol{s}_l$, $V(t_i, t_k; \boldsymbol{s}_j, \boldsymbol{s}_l)$ is actually the variance of $\varepsilon(t_i, \boldsymbol{s}_j)$. In the simulation, we set $\rho_w(d) = \exp(-d)$, and $\Upsilon = 30$. Unless stated otherwise, ϕ_w is chosen to be 0.6. The quantity ϕ_γ is chosen randomly from the set $\{0.2, 0.4, 0.6\}$, for $\gamma = 1, \ldots, \Upsilon$. The basis functions are generated as follows. Let $d_j = \min\{d_E(\boldsymbol{s}_j, \boldsymbol{s}_l), l \neq j\}$, for $j = 1, 2, ..., m$, and $\Gamma = \{(\boldsymbol{\zeta}, \varsigma) : \boldsymbol{\zeta} = \boldsymbol{s}_j, \varsigma = 0.75d_j, d_j \text{ or } 1.25d_j, \text{ for } j = 1, \ldots, m\}$. Then, there are $3m$ elements in Γ. Next, we choose Υ different elements randomly without replacement from Γ, and the set of chosen elements is denoted as $\Gamma' = \{(\boldsymbol{\zeta}_\gamma, \varsigma_\gamma), \gamma = 1, \ldots, \Upsilon\}$. Then, the basis functions are defined as $\Psi_\gamma(\boldsymbol{s}) = \exp\{-d_E(\boldsymbol{s}, \boldsymbol{\zeta}_\gamma)^2/(2\varsigma_\gamma^2)\}$, for $\gamma = 1, \ldots, \Upsilon$.

Case 2: Let $\boldsymbol{\varepsilon}(t_i) = (\varepsilon(t_i, \boldsymbol{s}_1), ..., \varepsilon(t_i, \boldsymbol{s}_m))^T$. Then, $\boldsymbol{\varepsilon}(t_i)$ is generated from the AR(1) process

$$\boldsymbol{\varepsilon}(t_i) = \phi\boldsymbol{\varepsilon}(t_{i-1}) + (1-\phi^2)^{1/2}\boldsymbol{\eta}(t_i),$$

where ϕ is a coefficient, and $\{\boldsymbol{\eta}(t_i) = (\eta(t_i, \boldsymbol{s}_1), \ldots, \eta(t_i, \boldsymbol{s}_m))^T, i = 1, \ldots, n\}$ are temporally independent spatial processes. At each $(t_i, \boldsymbol{s}_j)$, $\eta(t_i, \boldsymbol{s}_j)$ is assumed to have the distribution $N(0,1)$. At each observation time t_i, the spatial correlation among the elements of $\boldsymbol{\eta}(t_i)$ is described by $\text{Cov}(\eta(t_i, \boldsymbol{s}_j), \eta(t_i, \boldsymbol{s}_l)) = \rho(d_E(\boldsymbol{s}_j, \boldsymbol{s}_l))$, for any j and l. In such cases, it can be checked that the covariance between $\varepsilon(t_i, \boldsymbol{s}_j)$ and $\varepsilon(t_k, \boldsymbol{s}_l)$ is $V(t_i, t_k; \boldsymbol{s}_j, \boldsymbol{s}_l) = \phi^{|k-i|}\rho(d_E(\boldsymbol{s}_j, \boldsymbol{s}_l))$. In the simulation, we choose $\rho(d) = \exp(-d^2)$, and ϕ is always chosen to be 0.8 except in cases when we study the effect of correlation level on performance of the bandwidth selection for (h_1, h_2).

Case 3: The random error vector $\boldsymbol{\varepsilon}(t_i)$ is generated from the AR(2) process

$$\boldsymbol{\varepsilon}(t_i) = \phi_1\boldsymbol{\varepsilon}(t_{i-1}) + \phi_2\boldsymbol{\varepsilon}(t_{i-2}) + \boldsymbol{\eta}(t_i),$$

where ϕ_1 and ϕ_2 are two coefficients, and $\boldsymbol{\eta}(t_i)$ is the same as the one in the AR(1) model in Case 2, except that its spatial correlation is determined by a correlation function $\rho(\boldsymbol{s}, \boldsymbol{s}')$, for any $\boldsymbol{s}, \boldsymbol{s}' \in \Omega$. In such cases, the covariance between $\varepsilon(t_i, \boldsymbol{s}_j)$ and $\varepsilon(t_k, \boldsymbol{s}_l)$ is $V(t_i, t_k; \boldsymbol{s}_j, \boldsymbol{s}_l) = \varrho_{|k-i|}\rho(\boldsymbol{s}_j, \boldsymbol{s}_l)$, where $\varrho_0 = (1-\phi_2)/\{(1+\phi_2)((1-\phi_2)^2 - \phi_1^2)\}$, $\varrho_1 = \frac{\phi_1}{1-\phi_2}\varrho_0$, and $\varrho_d = \phi_1\varrho_{d-1} + \phi_2\varrho_{d-2}$ for $d > 1$. In the simulation, we set $\rho(\boldsymbol{s}, \boldsymbol{s}') = \exp\left[-\left\{(s_x - s_x') + (s_y - s_y')\right\}\right]$, where $\boldsymbol{s} = (s_x, s_y)^T$ and $\boldsymbol{s}' = (s_x', s_y')^T$. For the two parameters (ϕ_1, ϕ_2), unless stated otherwise, we choose $(\phi_1, \phi_2) = (0.5, 0.3)$.

4.1 | Performance evaluation of MCV

When the mean function $\lambda(t, \boldsymbol{s})$ is unknown, we need to estimate it by the LLK smoothing procedure (3). To use the local kernel estimating procedure, the two bandwidths (h_1, h_2) should be chosen properly. In this part, we study the performance of the MCV bandwidth selection procedure (7) and compare it with the conventional LOOCV procedure. To calculate the LOOCV score, we can also use the equation (7) with the bimodal kernel function $K_\epsilon(u)$ in (8) being replaced by the Epanechnikov kernel $K(u) = 0.75(1-u^2)\text{I}(|u| \leq 1)$. Then, the bandwidths can be chosen by minimizing the LOOCV score. To evaluate the performance of the mean estimate $\widehat{\lambda}(t, \boldsymbol{s})$, a natural metric is the mean average squared error (MASE) defined as

$$\text{MASE} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=i}^{m}\left[\lambda(t_i, \boldsymbol{s}_{ij}) - \widehat{\lambda}(t_i, \boldsymbol{s}_{ij})\right]^2.$$

So, in simulations, we can also choose the bandwidths (h_1, h_2) by minimizing the above MASE metric. This method of bandwidth selection procedure is denoted as OPTIMAL, because the resulting bandwidths would be optimal for estimating $\lambda(t, \boldsymbol{s})$. Then, to consider different cases of spatio-temporal data correlation, we let ϕ_w change among 0.3, 0.6 and 0.9 in Case 1, ϕ change among 0.2, 0.4 and 0.8 in Case 2, and (ϕ_1, ϕ_2) change among (0.1,0.1), (0.3,0.2) and (0.5,0.3) in Case 3. Based on 100 replicated simulations, the results of the selected bandwidths by different procedures together with the corresponding MASE values are presented in Table 1. From the table, it can be seen that MCV outperforms LOOCV quite dramatically in all cases considered,

in the sense that the selected bandwidths by MCV are much closer to the optimal bandwidths and the corresponding MASE values are much smaller than those of LOOCV.

[Table 1 about here.]

4.2 | Performance evaluation of the proposed covariance estimation method

Next, we evaluate the performance of the proposed covariance estimation method, denoted as NP, in comparison with two representative existing methods: the group lasso (GL) method by Hsu et al[10] and the B-spline (BS) method by Choi et al.[6] In the proposed method NP, the bandwidths are chosen by the MCV procedure (7)-(8) and the MSPE procedure (9). In the method GL, its parameters are chosen by the CV procedure, as discussed in Hsu et al.[10] In the method BS, its parameters are chosen to be the recommended values given in Choi et al.[6] The temporal correlation is assumed to be exponentially decayed in both GL and BS. As a comparison, we did not make this assumption in NP. To make the comparison fair among different methods, we use the following truncated mean average squared error (TMASE) to measure the performance of the estimated covariance structure:

$$\mathrm{TMASE}(\vartheta) = \frac{\sum_{i=1}^{n}\sum_{k=i}^{n}\sum_{j=1}^{m}\sum_{l=1}^{m}\left[V(t_i,t_k;\boldsymbol{s}_j,\boldsymbol{s}_l)-\widehat{V}(t_i,t_k;\boldsymbol{s}_j,\boldsymbol{s}_l)\right]^2 \mathrm{I}(k-i\le\vartheta)}{m^2\sum_{i=1}^{n}\sum_{k=i}^{n}\mathrm{I}(k-i\le\vartheta)},$$

where ϑ is an positive integer for truncation. The results based on 100 replicated simulations when ϑ changes among 5, 10, or 20 are presented in Tables 2 and 3, respectively, for the two scenarios when the mean function $\lambda(t,\boldsymbol{s})$ is assumed known or unknown. From Table 2, it can be seen that i) the TMASE(ϑ) values of the proposed method NP decrease when the sample size (m,n) increases from $(64,200)$ to $(100,500)$, and ii) in most cases, TMASE(ϑ) values of NP are much smaller than those of GL and BS, especially when the sample size is large (i.e., $(m,n)=(100,500)$). Similar conclusions can be made from Table 3, although the TMASE(ϑ) values in Table 3 are generally larger than those in Table 2, because the randomness in the estimated mean function $\widehat{\lambda}(t,\boldsymbol{s})$ would be added to the randomness in the estimated covariance structure in the case of Table 3.

[Table 2 about here.]

[Table 3 about here.]

In Tables 2 and 3, the truncation parameter ϑ takes the value of 5, 10 or 20, which is quite small. Next, we compare the three methods in cases when ϑ could be chosen large. To this end, in the example of Table 3 when $(m,n)=(64,200)$, let ϑ change its value from 5 to 150, and other setups remain unchanged. The related results about the calculated TMASE(ϑ) values of the three methods are shown in Figure 1. From the plots, it can be seen that i) NP is much better than the other two methods when ϑ is relatively small in all cases, except in Case 1 where GL is slightly better because all its assumptions are satisfied in that case, and ii) when ϑ gets larger, performance of the three methods tends to be similar. It should be pointed out that both GL and BS are based on the assumptions that the temporal data correlation is stationary and it decays monotonically, which are satisfied in this example. But, the proposed method NP does not impose these assumptions, which puts it at a disadvantage in the comparison.

[Figure 1 about here.]

5 | APPLICATION TO A HAND, FOOT AND MOUTH DISEASE DATASET

We now present an application of our method to a hand, foot and mouth disease (HFMD) dataset. The dataset contains weekly incidence rates of HFMD in 21 cities of Sichuan province of China during 2009-2010 (52 weeks). Besides the proposed method NP, we also consider the methods GL and BS that are discussed in some simulation examples in Section 4. The setup of the three methods also remains the same as that in Section 4. Because the true spatio-temporal covariance structure is unknown in this case, the performance metric TMASE is not well defined. So, the MSPE metric defined in (9) is used here as the performance measure. The calculated MSPE values for NP, GL and BS are 3.62×10^{-12}, 6.04×10^{-12} and 7.14×10^{-12}, respectively. Therefore, the proposed method NP outperforms the other two methods GL and BS in quite large margins in this example.

The estimated mean function for 4 cities (Aba Tibetan and Qiang Autonomous Prefecture, Deyang, Mianyang, and Panzhihua) are presented in Figure 2, along with the observed data. From the plots in the figure, it can be seen that the estimated mean

function $\widehat{\lambda}(t, \boldsymbol{s})$ describes the observed data well. Because the data correlation is assumed temporally stationary in GL and BS, we calculate the estimated temporal correlation with a time lag τ as the average of the elements in the set

$$\left\{ \left[\widehat{\sigma}^2(t_i, \boldsymbol{s}_j)\widehat{\sigma}^2(t_{i+\tau}, \boldsymbol{s}_j)\right]^{-1/2} \widehat{V}(t_i, t_{i+\tau}; \boldsymbol{s}_j, \boldsymbol{s}_j) \,:\, i = 1, \ldots, 52-\tau, j = 1, \ldots, 21 \right\},$$

for $0 \leq \tau \leq 51$. The estimated temporal correlations of the three methods are presented in Figure 3. From the plots in the figure, it can be seen that the estimated temporal correlations by GL and BS are both smooth in τ, as required in the two methods, the one by BS decreases exponentially fast, and the one by GL decreases much slower. The estimated temporal correlation by NP is close to that of GL, but it is less smooth to reflect the variability in the observed data. Figure 4 presents the observed incidence rates and the prediction errors, defined as the observed values minus the predicted values, of the three methods at the 16th, 32nd and 48th week, respectively. From the plots in the figure, it can be seen that the proposed method has the best prediction performance, compared with the two alternative methods.

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

6 | CONCLUDING REMARKS

We have described a nonparametric approach for estimating the covariance structure of a spatio-temporal dataset in the previous several sections. The proposed method is based on local kernel estimation of the variance and covariance functions, after the mean function is estimated properly. The bandwidths used in local kernel estimation are chosen by a new metric defined in (9) based on spatio-temporal kriging. Both theoretical results and numerical studies show that the proposed method works well in practice.

Acknowledgments: The authors thank the editor, the associate editor and two referees for some constructive comments and suggestions that improved the quality of the paper greatly.

References

1. Cressie N, Wikle C. *Statistics for Spatio-Temporal Data*. New York: Wiley; 2011.

2. Diggle PJ. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. 3rd ed. London: Chapman & Hall/CRC; 2014.

3. Gneiting T, Guttorp P. Continuous parameter spatio-temporal processes. In: Gelfand AE, Diggle PJ, Fuentes M, Guttorp P, eds. *Handbook of Spatial Statistics*. Boca Raton, FL: CRC Press; 2010:427–436.

4. Schabenberger O, Gotway CA. *Statistical Methods for Spatial Data Analysis*. New York: Chapman and Hall; 2005.

5. Yang K, Qiu P. Spatiotemporal incidence rate data analysis by nonparametric regression. *Statist Med*. 2018;37:2094–2107.

6. Choi I, Li B, Wang X. Nonparametric estimation of spatial and space-time covariance functions. *J Agric Biol Environ Stat*. 2013;18:611–630.

7. Cressie N, Huang HC. Classes of nonseparable, spatio-temporal stationary covariance functions. *J Am Stat Assoc*. 1999;94:1330–1340.

8. Genton MG. Separable approximations of space-time covariance matrices. *Environmetrics*. 2007;18:681–695.

9. Gneiting T. Nonseparable stationary covariance functions for space-time data. *J Am Stat Assoc*. 2002;97:590–600.

10. Hsu NJ, Chang YM, Huang HC. A group lasso approach for non-stationary spatial-temporal covariance estimation. *Environmetrics*. 2012;23:12–23.

11. Shand L, Li B. Modeling nonstationarity in space and time. *Biometrics*. 2017;73:759–768.

12. Robinson PM. Nonparametric estimation of time-varying parameters. In: Hackl P, ed. *Statistical Analysis and Forecasting of Economic Structural Change*. Berlin: Springer; 1989:253-264.

13. Dahlaus R. Fitting time series models to nonstationary processes. *Ann Stat*. 1997;25:1-37.

14. Vogt M, Linton O. Nonparametric estimation of a periodic sequence in the presence of a smooth trend. *Biometrika*. 2014;101:121–140.

15. Hall P, Fisher NI, Hoffmann B. On the nonparametric estimation of covariance functions. *Ann Stat*. 1994;22:2115–2134.

16. Higham NJ. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Its Appl*. 1988;103:103–118.

17. Epanechnikov VA. Non-parametric estimation of a multivariate probability density. *Theory Probab Its Appl*. 1969;14:153–158.

18. Altman NS. Kernel smoothing of data with correlated errors. *J Am Stat Assoc*. 1990;85:749–758.

19. Opsomer J, Wang Y, Yang Y. Nonparametric regression with correlated errors. *Stat Sci*. 2001;16:134–153.

20. Brabanter KD, Brabanter JD, Suykens JAK, Moor BD. Kernel regression in the presence of correlated errors. *J Mach Learn Res*. 2011;12:1955–1976.

21. Liebscher E. Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stoch Process Their Appl*. 1996;65:69–80.

22. Diggle PJ, Verbyla AP. Nonparametric estimation of covariance structure in longitudinal data. *Biometrics*. 1998;54:401–415.

23. Hall P, Müller HG, Yao F. Properties of principal component methods for functional and longitudinal data analysis. *Ann Stat*. 2006;34:1493–1517.

24. Hansen BE. Uniform convergence rates for kernel estimation with dependent data. *Econom Theory*. 2008;24:726–748.

25. Hart JD, Wehrly TE. Kernel regression estimation using repeated measurements data. *J Am Stat Assoc*. 1986;81:1080–1088.

26. Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. *J Am Stat Assoc*. 2005;100:577–590.

TABLE 1 In each entry, numbers in line 1 are the MASE value and its standard error (in parenthesis) and numbers in line 2 are the selected values of the bandwidths.

	$(m,n)=(64,200)$			$(m,n)=(100,500)$		
	LOOCV	MCV	OPTIMAL	LOOCV	MCV	OPTIMAL
Case 1, $\phi_w=0.3$	0.11(0.003)	0.06(0.002)	0.05(0.002)	0.05(0.001)	0.02(0.001)	0.02(0.001)
	(0.08,0.39)	(0.18,0.42)	(0.21,0.54)	(0.05,0.37)	(0.15,0.40)	(0.17,0.46)
Case 1, $\phi_w=0.6$	0.22(0.005)	0.11(0.004)	0.08(0.004)	0.12(0.002)	0.04(0.001)	0.03(0.001)
	(0.06,0.35)	(0.16,0.37)	(0.24,0.58)	(0.04,0.34)	(0.13,0.38)	(0.22,0.49)
Case 1, $\phi_w=0.9$	0.64(0.018)	0.33(0.015)	0.24(0.014)	0.44(0.008)	0.18(0.005)	0.10(0.004)
	(0.04,0.33)	(0.15,0.35)	(0.26,0.61)	(0.03,0.31)	(0.11,0.38)	(0.25,0.53)
Case 2, $\phi=0.2$	0.05(0.002)	0.04(0.001)	0.03(0.001)	0.04(0.001)	0.02(0.001)	0.02(0.001)
	(0.12,0.52)	(0.20,0.52)	(0.20,0.54)	(0.04,0.38)	(0.16,0.42)	(0.17,0.51)
Case 2, $\phi=0.4$	0.08(0.003)	0.05(0.002)	0.04(0.002)	0.08(0.002)	0.03(0.001)	0.02(0.001)
	(0.09,0.47)	(0.18,0.48)	(0.24,0.59)	(0.03,0.36)	(0.15,0.40)	(0.20,0.55)
Case 2, $\phi=0.8$	0.32(0.010)	0.18(0.007)	0.14(0.006)	0.31(0.006)	0.09(0.004)	0.07(0.004)
	(0.07,0.45)	(0.17,0.46)	(0.28,0.68)	(0.03,0.32)	(0.14,0.39)	(0.25,0.59)
Case 3, $(\phi_1,\phi_2)=(0.1,0.1)$	0.05(0.001)	0.03(0.001)	0.03(0.001)	0.04(0.001)	0.02(0.001)	0.02(0.001)
	(0.09,0.52)	(0.17,0.62)	(0.25,0.54)	(0.04,0.39)	(0.17,0.45)	(0.18,0.48)
Case 3, $(\phi_1,\phi_2)=(0.3,0.2)$	0.13(0.004)	0.07(0.003)	0.06(0.002)	0.12(0.002)	0.03(0.002)	0.03(0.002)
	(0.08,0.46)	(0.16,0.57)	(0.27,0.58)	(0.03,0.35)	(0.16,0.43)	(0.22,0.53)
Case 3, $(\phi_1,\phi_2)=(0.5,0.3)$	0.71(0.019)	0.38(0.015)	0.27(0.012)	0.66(0.012)	0.20(0.008)	0.13(0.008)
	(0.06,0.41)	(0.15,0.54)	(0.32,0.61)	(0.03,0.30)	(0.14,0.41)	(0.27,0.57)

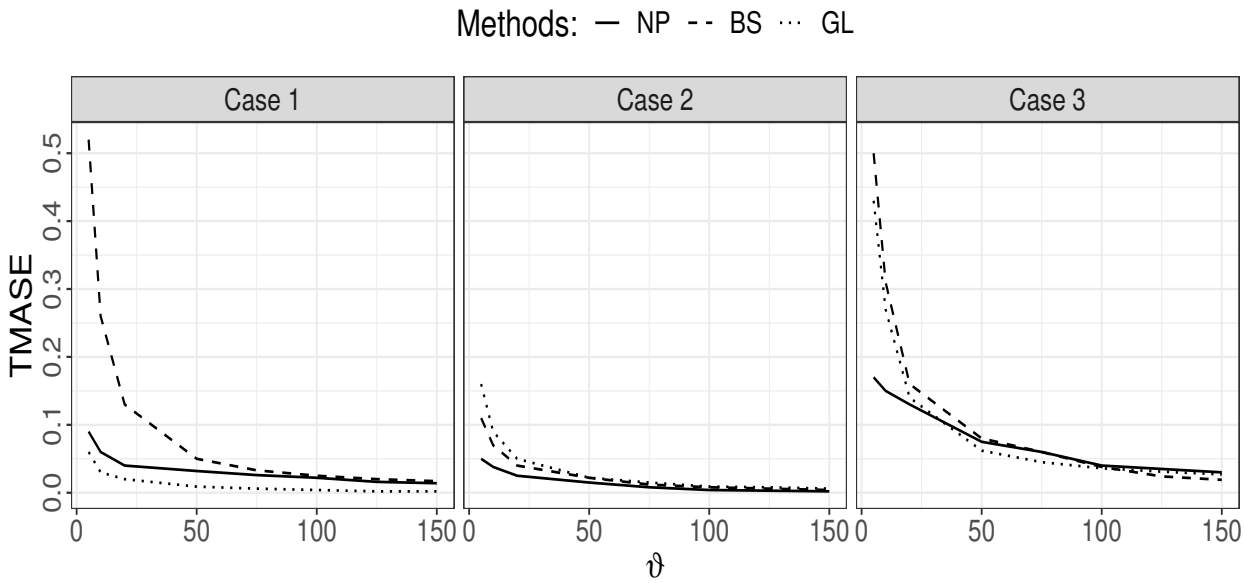

FIGURE 1 TAMSE values of the proposed method and two competing methods in cases when $(m,n)=(64,200)$, ϑ changes from 5 to 150, and the true mean function $\lambda(t,\boldsymbol{s})$ is assumed unknown.

TABLE 2 TMASE values and their standard errors (in parentheses) for the proposed method and two competing methods when the true mean function $\lambda(t, \boldsymbol{s})$ is assumed known. All numbers in the parentheses are in 1×10^{-2}, and 0.00 means the value is smaller than 5×10^{-5}.

		TMASE(5)			TMASE(10)			TMASE(20)		
	(m,n)	NP	GL	BS	NP	GL	BS	NP	GL	BS
Case 1	(64,200)	0.05	0.05	0.52	0.05	0.03	0.26	0.05	0.01	0.13
		(0.24)	(0.04)	(5.50)	(0.17)	(0.02)	(2.70)	(0.12)	(0.01)	(1.41)
	(100,500)	0.01	0.05	0.57	0.01	0.02	0.28	0.01	0.01	0.14
		(0.04)	(0.01)	(7.33)	(0.03)	(0.01)	(3.62)	(0.03)	(0.00)	(1.84)
Case 2	(64,200)	0.04	0.15	0.10	0.03	0.08	0.06	0.02	0.04	0.03
		(0.42)	(0.16)	(0.61)	(0.27)	(0.10)	(0.51)	(0.19)	(0.07)	(0.25)
	(100,500)	0.01	0.13	0.08	0.01	0.07	0.05	0.01	0.04	0.02
		(0.13)	(0.05)	(0.87)	(0.12)	(0.04)	(0.42)	(0.09)	(0.02)	(0.21)
Case 3	(64,200)	0.15	0.40	0.42	0.13	0.25	0.27	0.12	0.14	0.14
		(0.68)	(0.67)	(3.93)	(0.47)	(0.39)	(1.94)	(0.33)	(0.21)	(0.96)
	(100,500)	0.07	0.30	0.60	0.06	0.20	0.35	0.06	0.11	0.18
		(0.48)	(0.20)	(5.22)	(0.46)	(0.17)	(2.51)	(0.36)	(0.09)	(1.31)

TABLE 3 TMASE values and their standard errors (in parentheses) for the proposed method and two competing methods when the true mean function $\lambda(t, \boldsymbol{s})$ is assumed unknown. All numbers in the parentheses are in 1×10^{-2}, and 0.00 means the value is smaller than 5×10^{-5}.

		TMASE(5)			TMASE(10)			TMASE(20)		
	(m,n)	NP	GL	BS	NP	GL	BS	NP	GL	BS
Case 1	(64,200)	0.09	0.06	0.52	0.06	0.03	0.26	0.04	0.02	0.13
		(0.22)	(0.06)	(6.63)	(0.10)	(0.03)	(3.38)	(0.04)	(0.02)	(1.61)
	(100,500)	0.02	0.05	0.56	0.01	0.03	0.28	0.01	0.01	0.14
		(0.05)	(0.02)	(5.17)	(0.03)	(0.01)	(2.61)	(0.02)	(0.00)	(1.32)
Case 2	(64,200)	0.05	0.16	0.11	0.04	0.09	0.07	0.03	0.05	0.04
		(0.32)	(0.20)	(0.29)	(0.19)	(0.12)	(0.14)	(0.18)	(0.06)	(0.07)
	(100,500)	0.02	0.13	0.11	0.01	0.07	0.06	0.01	0.04	0.03
		(0.11)	(0.05)	(0.24)	(0.08)	(0.04)	(0.12)	(0.05)	(0.02)	(0.06)
Case 3	(64,200)	0.17	0.43	0.50	0.15	0.27	0.31	0.13	0.14	0.16
		(1.23)	(0.72)	(5.80)	(1.03)	(0.42)	(2.81)	(0.78)	(0.21)	(1.42)
	(100,500)	0.07	0.31	0.57	0.07	0.20	0.34	0.06	0.11	0.18
		(0.30)	(0.24)	(6.11)	(0.27)	(0.20)	(3.02)	(0.20)	(0.11)	(1.53)

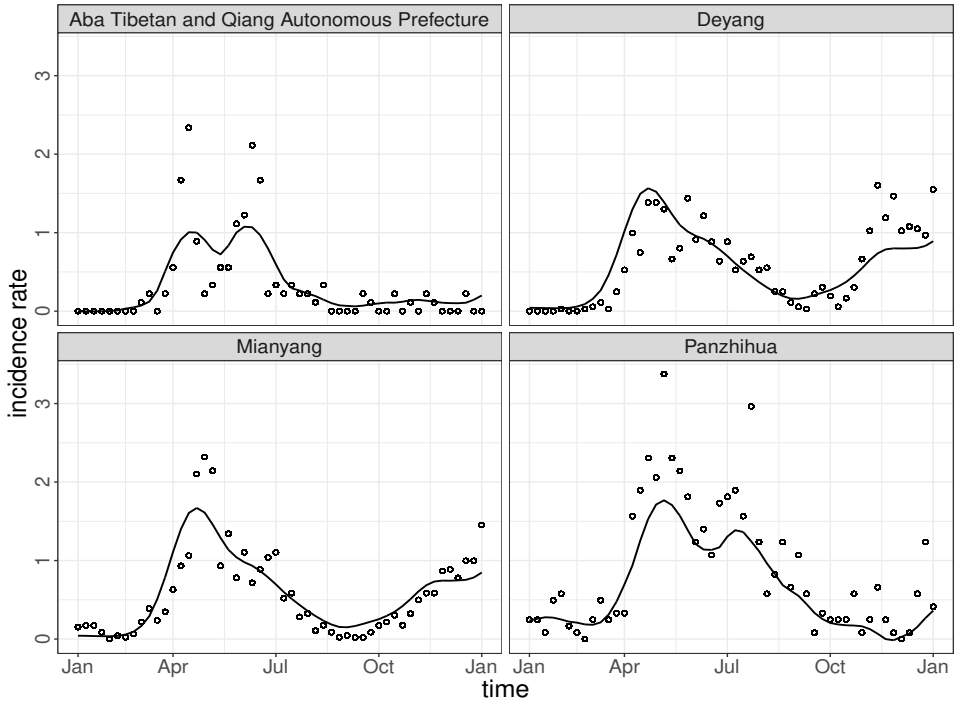

FIGURE 2 Observed incidence rates (little circles) of the hand, foot and mouth disease in four cities of Sichuan province of China and the estimated mean functions (solid curves). The y-axis is in the scale of 10^{-5}.

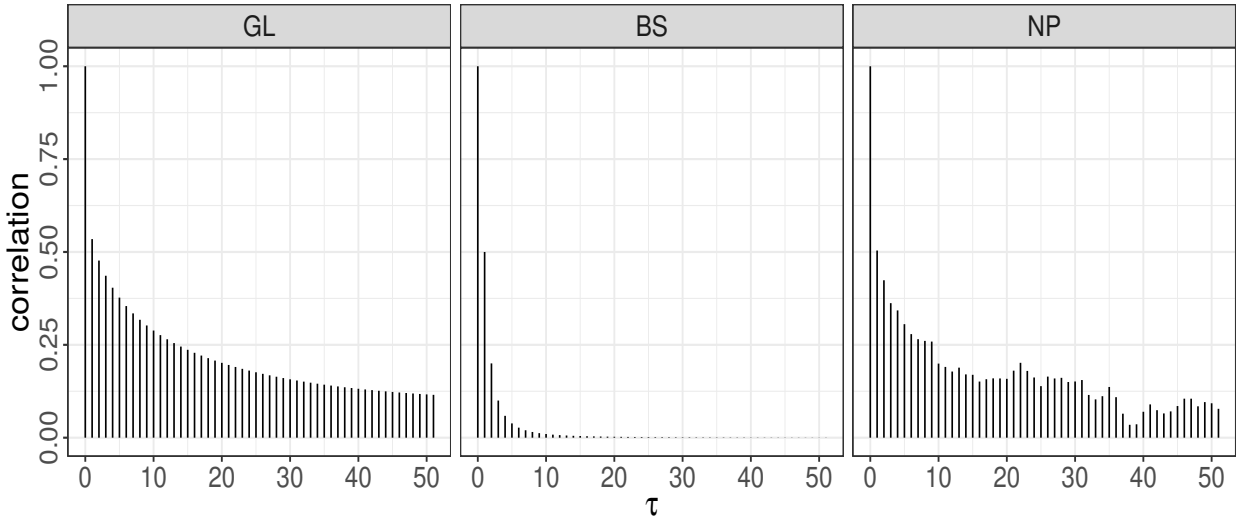

FIGURE 3 Estimated temporal correlations by the proposed method NP and two competing methods GL and BS.

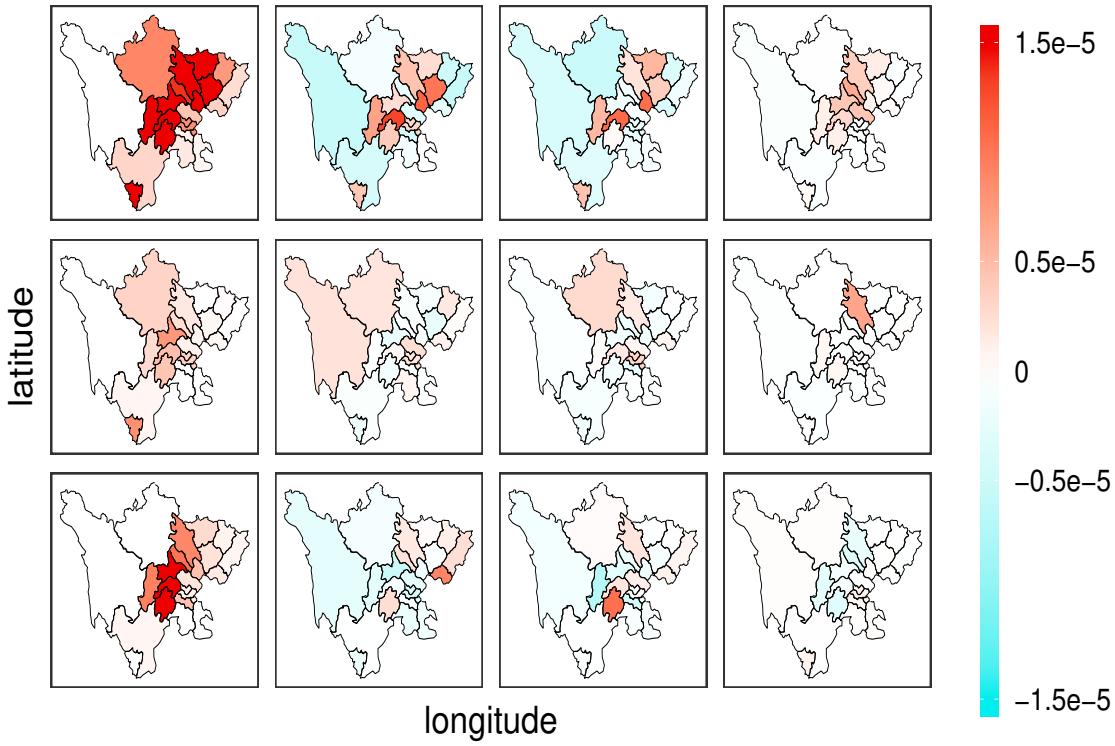

FIGURE 4 Observed incidence rates of the hand, foot and mouth disease (1st column) in 21 cities of Sichuan province in China, and prediction errors of GL (2nd column), BS (3rd column), and NP (4th column), at the 16th week (1st row), 32nd week (2nd row) and 48th week (3rd row) of the year 2009.