Annals of the Institute of Statistical Mathematics (2022) 74:49–68
https://doi.org/10.1007/s10463-021-00787-2

A three-step local smoothing approach for estimating the mean and covariance functions of spatio-temporal Data

Kai Yang[1] · Peihua Qiu[1]

Received: 25 May 2020 / Revised: 5 January 2021 / Accepted: 8 January 2021 / Published online: 20 March 2021

Abstract
Spatio-temporal data are common in practice. Existing methods for analyzing such data often employ parametric modelling with different sets of model assumptions. However, spatio-temporal data in practice often have complicated structures, including complex spatial and temporal data variation, latent spatio-temporal data correlation, and unknown data distribution. Because such data structures reflect the complicated impact of confounding variables, such as weather, demographic variables, life styles, and other cultural and environmental factors, they are usually too complicated to describe by parametric models. In this paper, we suggest a general modelling framework for estimating the mean and covariance functions of spatio-temporal data using a three-step local smoothing procedure. The suggested method can well accommodate the complicated structure of real spatio-temporal data. Under some regularity conditions, the consistency of the proposed estimators is established. Both simulation studies and a real-data application show that our proposed method could work well in practice.

Keywords Bandwidth selection · Consistency · Covariance estimation · Functional data analysis · Local smoothing · Spatio-temporal data

1 Introduction

A large amount of spatio-temporal data has become available to researchers in statistics, epidemiology, geography, oceanography, environmental science, and more. The increasing computing power has made it possible for us to analyze these data with more and more realistic models. This paper aims to propose a general modelling

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-021-00787-2.

✉ Peihua Qiu
pqiu@ufl.edu

[1] Department of Biostatistics, University of Florida, Gainesville, FL 32610, USA

framework for estimating the mean and variance/covariance functions of spatio-temporal data using a local smoothing procedure.

In the statistical literature, there have been many existing spatio-temporal modelling approaches. One such approach suggested by Stroud et al. (2001) is for linear dynamic spatio-temporal modelling (DSTM), in which the mean response at each time point is assumed to be a locally weighted mixture of some pre-specified basis functions, the spatial surfaces at different time points are assumed to follow a linear evolution equation, and the spatial random noise is assumed to be a Gaussian noise process. Some generalized versions of DSTM have been developed to describe nonlinear evolution process (Wikle and Hooten 2010) or model high-dimensional multivariate spatio-temporal data (Bradley et al. 2015). More details about dynamical spatio-temporal models can be found in the books Cressie and Wikle (2011) and Wikle et al. (2019). Another approach for modelling a spatio-temporal point process is based on the log-Gaussian Cox process (LGCP) framework (Møller et al. 1998), where the logarithm of the related spatio-temporal intensity function is assumed to follow a spatio-temporal Gaussian process with a separable or nonseparable covariance structure, and conditional on the intensity function the original point process is assumed to be Poisson-distributed (e.g. Cressie and Huang 1999; Diggle et al. 2013). By using the idea of scale mixing, Fonseca and Steel (2011) suggested a method for modelling non-Gaussian spatio-temporal data, based on the assumptions that the observed data follow a Gaussian process given some scale mixing variables and that the covariance structure has a specific parametric form. To analyze a sequence of large air quality datasets, Datta et al. (2016) developed a dynamic nearest-neighbor Gaussian process model to provide statistical inference by using data information from nearest neighbors. There are some other methods for estimating the spatio-temporal mean and/or covariance structure, including the linear regression modelling using temporal basis functions (cf., Lindström et al. 2015), the kernel smoothing methods with or without considering spatial data correlation (e.g. Kafadar 1996; Yang and Qiu 2018 and Yang and Qiu (2019)), the function estimation methods based on B-splines (e.g. Choi et al. 2013) or LASSO penalized least squares (e.g. Shand and Li 2017), the space-time ANOVA-type methods (Heuvelink and Griffith 2010), and more.

The parametric model assumptions in some existing methods described above could be invalid or difficult to justify in certain applications. Consequently, their performance may not be reliable in such cases. The data covariance structure is not used or well accommodated when estimating the mean function in these or some other existing methods; thus, there is much room for us to improve their effectiveness. In this paper, we propose a general modelling framework for jointly estimating the mean and variance/covariance functions of spatio-temporal data using a three-step local smoothing procedure. The proposed method does not impose restrictive assumptions on the spatio-temporal mean structure, the spatio-temporal covariance structure, and the data distribution. Because it is based on local smoothing, its computation is relatively fast. In the proposed method, the spatio-temporal covariance structure is estimated and accommodated properly when estimating the mean structure. To properly distinguish the mean and covariance structure in model estimation, a new spatio-temporal bandwidth selection procedure is also developed. All these features make the proposed method effective in analyzing spatio-temporal data, which is confirmed by theoretical arguments and numerical studies.

The rest of the article is organized as follows. The proposed nonparametric spatio-temporal regression method is first described in detail in Sect. 2. Some of its theoretical properties are derived in Sect. 3. Its numerical performance is evaluated by some simulation studies in Sect. 4. A real-data application is discussed in Sect. 5, and we conclude with a discussion in Sect. 6. Some technical details, including the proofs of certain theoretical results, can be found in the supporting materials.

2 Local smoothing approach for estimation Spatio-temporal mean and variance/covariance functions

Our proposed method is described in three parts. A nonparametric spatio-temporal regression model is discussed in Sect. 2.1, its three-step model estimation is discussed in Sect. 2.2, and a novel spatio-temporal bandwidth selection procedure is discussed in Sect. 2.3.

2.1 A nonparametric spatio-temporal regression model

Assume that a response variable y is observed at n equally spaced time points $\{t_i = i/n, i = 1, \dots, n\}$ in the time interval [0, 1], and at the ith time point it is observed at m_i spatial locations $\{\boldsymbol{s}_{ij}, j = 1, \dots, m_i\}$ in a spatial region $\Omega \subseteq \mathbb{R}^2$, for $i = 1, \dots, n$. The observed data are assumed to follow the following nonparametric spatio-temporal regression model:

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

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

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

where $t, t' \in [0, 1]$, and $\boldsymbol{s}, \boldsymbol{s}' \in \Omega$. In expressions (1) and (2), 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 times $t_i = i/n \in [0, 1]$, rather than i. This formulation is commonly used in the literature. See, e.g. Robinson (1989), Dahlaus (1997), and Vogt and Linton (2014). The rescaled time is necessary for studying asymptotic properties of the estimated mean and variance/covariance functions in the time domain. Otherwise, the distance between two consecutive observation times would not go to 0 when n increases, and thus the asymptotic properties in the time domain cannot be discussed properly. It should be pointed out that models (1) and (2) are flexible. They do not impose any parametric assumptions on the mean $\lambda(t, \boldsymbol{s})$, the variance/covariance $\sigma^2(t, \boldsymbol{s})$ and $V(t, t'; \boldsymbol{s}, \boldsymbol{s}')$, or the distribution of the response variable $y(t, \boldsymbol{s})$. They even allow the observation locations to be different at different time points. This last feature would make our proposed method applicable in more applications, compared to some existing methods that require the observation locations to be unchanged over time. For example, in infectious disease research, hospitals reporting daily disease incidence data may be different at different time points, which makes the spatial locations of observed daily disease incidence to be different.

2.2 Three-step model estimation

Estimation of models (1) and (2) consists of the following three steps: (i) an initial estimate of the mean function $\lambda(t, \boldsymbol{s})$ is first obtained, (ii) the variance/covariance functions $\sigma^2(t, \boldsymbol{s})$ and $V(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ are then estimated accordingly, and (iii) the final estimate of $\lambda(t, \boldsymbol{s})$ is obtained after the estimated variance/covariance functions are accommodated in the mean estimation. Each of these three steps is described below in detail.

Let $\boldsymbol{Y} = (y(t_1, \boldsymbol{s}_{11}), \ldots, y(t_1, \boldsymbol{s}_{1m_1}), \ldots, y(t_n, \boldsymbol{s}_{nm_n}))^T$ be the long vector of all observations. Then, for given $(t, \boldsymbol{s}) \in [0, 1] \times \Omega$, $\lambda(t, \boldsymbol{s})$ can be estimated by the following weighted local linear kernel (WLLK) smoothing procedure (cf., Qiu 2005, Chapter 2):

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^4} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \widehat{\Sigma}_K^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{3}$$

where $\boldsymbol{X} = (X_{11}, \ldots, X_{1m_1}, \ldots, X_{nm_n})^T$ is the design matrix with $X_{ij} = (1, t_i - t, (\boldsymbol{s}_{ij} - \boldsymbol{s})^T)^T$, for $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ is the coefficient vector, $\widehat{\Sigma}_K = D_K^{-1/2} \widehat{\Sigma}_{\mathbf{Y}} D_K^{-1/2}$, $D_K =$ diag $\{w_0(1, 1), \ldots, w_0(1, m_1), \ldots, w_0(n, m_n)\}$,$w_0(i,j) = K_1((t_i - t)/h_{t,0}) K_2(d_E(\boldsymbol{s}_{ij}, \boldsymbol{s})/h_{s,0})$, $h_{t,0}, h_{s,0} > 0$ are two bandwidths, $K_1(\cdot)$ and $K_2(\cdot)$ are two univariate kernel functions, $d_E(\cdot, \cdot)$ is the Euclidean distance in $\mathbb{R}^2$, and $\widehat{\Sigma}_{\mathbf{Y}}$ is the estimated covariance matrix of $\boldsymbol{Y}$. If the estimated covariance function is denoted as $\widehat{V}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$, then $\widehat{V}(t_i, t_k; \boldsymbol{s}_{ij}, \boldsymbol{s}_{kl})$ can be used as the $(\sum_{q=1}^{i-1} m_q + j, \sum_{q=1}^{k-1} m_q + l)$-th element of $\widehat{\Sigma}_{\mathbf{Y}}$, for any $1 \leq i, k \leq n$, $1 \leq j \leq m_i$ and $1 \leq l \leq m_k$. Throughout this paper, the inverse of a matrix refers to the Moore–Penrose generalized inverse. To obtain an initial estimate of $\lambda(t, \boldsymbol{s})$, because $\widehat{V}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ has not been obtained yet, we can simply set $\widehat{\Sigma}_{\mathbf{Y}}$ to be the identity matrix, in which case the spatio-temporal observations are assumed to be independent of each other. Then, the initial estimate of $\lambda(t, \boldsymbol{s})$ is the solution of minimization problem (3) to β_0, and it has the following expression:

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

where $\boldsymbol{e}_1 = (1, 0, 0, 0)^T$. In (3), the two kernel functions are usually chosen to have finite supports. Thus, $\widetilde{\lambda}(t, \boldsymbol{s})$ in (4) is a weighted average of observations in a neighborhood of $(t, \boldsymbol{s})$, the neighborhood size is controlled by the bandwidths $h_{t,0}$ and $h_{s,0}$, and the weights are controlled by the kernel functions $K_1(\cdot)$ and $K_2(\cdot)$. In this paper, all kernel functions are chosen to be the Epanechnikov kernel function because of its good theoretical properties (cf., Epanechnikov 1969). Namely, we choose $K_1(x) = K_2(x) = 0.75(1 - x^2) I(|x| \leq 1)$. Because $K_2(x)$ is used on a two-dimensional spatial region, its normalizing constant should be chosen differently from 0.75 to become a density kernel, but the normalizing constant does not need to be specified correctly because it will be cancelled out in the estimate $\widetilde{\lambda}(t,\boldsymbol{s})$.

The initial estimate $\widetilde{\lambda}(t,\boldsymbol{s})$ in (4) ignores the spatio-temporal data correlation by replacing the covariance matrix $\widehat{\Sigma}_{\mathbf{Y}}$ with the identity matrix. Similar to the theory of generalized estimation equation (Liang and Zeger 1986), it will be shown in Sect. 3 that this estimate is statistically consistent under some regularity conditions. To improve its efficiency, we need to estimate the variance/covariance functions $\sigma^2(t,\boldsymbol{s})$ and $V(t,t';\boldsymbol{s},\boldsymbol{s}')$. To this end, we define residuals $\widetilde{\varepsilon}(t_i,\boldsymbol{s}_{ij}) = y(t_i,\boldsymbol{s}_{ij}) - \widetilde{\lambda}(t_i,\boldsymbol{s}_{ij})$. Then, the variance function $\sigma^2(t,\boldsymbol{s})$ can be estimated by

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

where $w_1(i,j) = K_1\big((t_i-t)/h_{t,1}\big)K_2\big(d_E(\mathbf{s}_{ij},\mathbf{s})/h_{s,1}\big)$. The covariance function $V(t,t';\boldsymbol{s},\boldsymbol{s}')$ can be estimated by

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

where $w_2(i,j,k,l) = K_1\big((t_i-t)/h_{t,1}\big)K_1\big((t_k-t')/h_{t,1}\big)K_2\big(d_E(\boldsymbol{s}_{ij},\boldsymbol{s})/h_{s,1}\big)K_2\big(d_E(\boldsymbol{s}_{kl},\boldsymbol{s}')/h_{s,1}\big)$. In (6), $\widehat{V}(t,t';\boldsymbol{s},\boldsymbol{s}')$ is actually defined as a weighted sample covariance computed from pairs of the residuals in the neighborhoods of $(t,\boldsymbol{s})$ and $(t',\boldsymbol{s}')$, respectively, and the weights are determined by the kernel functions. In (5) and (6), the two bandwidths $(h_{t,1},h_{s,1})$ could be chosen differently from the bandwidths $(h_{t,0},h_{s,0})$ used in computing the initial mean estimate in (4). For the covariance structure, it is often reasonable to assume that $V(t,t';\boldsymbol{s},\boldsymbol{s}')$ is close to 0 when the time lag between t and t' is large. Therefore, to reduce computational burden in covariance estimation, we can simply set the estimated covariance to be 0 when $|t-t'|n$ is larger than a properly chosen threshold value c_n.

After the variance and covariance functions are estimated by (5) and (6), the covariance matrix $\Sigma_{\mathbf{Y}}$ of the observed data $\mathbf{Y}$ can be computed from $\widehat{\sigma}^2(t,\boldsymbol{s})$ and $\widehat{V}(t,t';\boldsymbol{s},\boldsymbol{s}')$. However, the estimated covariance matrix $\widehat{\Sigma}_{\mathbf{Y}}$ may not be a positive definite matrix. To make it to be a symmetric positive definite matrix, we suggest using the modification approach proposed by Higham (1998) 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, $\mathcal{P}$ be the set of all symmetric positive definite matrices with the same dimensions as those of $\widehat{\Sigma}_{\mathbf{Y}}$. Then, the projection of $\widehat{\Sigma}_{\mathbf{Y}}$ on $\mathcal{P}$ in the Frobenius matrix norm is

$$\widetilde{\Sigma}_{\mathbf{Y}} = \arg\min_{P\in\mathcal{P}}\|P-\widehat{\Sigma}_{\mathbf{Y}}\|_F.$$

It has been shown that $\widetilde{\Sigma}_{\mathbf{Y}} = (\widehat{\Sigma}_{\mathbf{Y}} + \widehat{\Sigma}_{\mathbf{Y},\mathbf{p}})/2$ (cf., Higham 1998), where $\widehat{\Sigma}_{\mathbf{Y},\mathbf{p}}$ is the symmetric polar factor of $\widehat{\Sigma}_{\mathbf{Y}}$. The projection $\widetilde{\Sigma}_{\mathbf{Y}}$ can be obtained by using the nearPD() command in R package matrix. Then, we suggest using $\widetilde{\Sigma}_{\mathbf{Y}}$ to replace $\widehat{\Sigma}_{\mathbf{Y}}$ in (3). It can be checked that if $\widehat{\Sigma}_{\mathbf{Y}}$ is symmetric positive definite, then $\widetilde{\Sigma}_{\mathbf{Y}}$ and $\widehat{\Sigma}_{\mathbf{Y}}$ are the same.

Finally, the estimate of $\lambda(t, \boldsymbol{s})$ can be obtained from (3) with $\widehat{\Sigma}_K$ being replaced by $\widetilde{\Sigma}_K = D_K^{-1/2}\widetilde{\Sigma}_{\mathbf{Y}}D_K^{-1/2}$, where D_K is defined immediately after (3). So, the final estimate of $\lambda(t, \boldsymbol{s})$ is defined to be

$$\widehat{\lambda}(t, \boldsymbol{s}) = \boldsymbol{e}_1^T \left(\boldsymbol{X}^T \widetilde{\Sigma}_K^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \widetilde{\Sigma}_K^{-1} \boldsymbol{Y}. \tag{7}$$

In (7), the bandwidths could be chosen to be different from $(h_{t,0}, h_{s,0})$ used in obtaining the initial estimate $\widetilde{\lambda}(t, \boldsymbol{s})$ in (4). The new bandwidths used in (7) are denoted as $(h_{t,2}, h_{s,2})$. From (4)–(7), it is natural to consider the following iterative procedure: the estimates of the variance and covariance functions defined in (5)–(6) can be further updated by replacing the initial mean estimate $\widetilde{\lambda}(t, \boldsymbol{s})$ with the mean estimate $\widehat{\lambda}(t, \boldsymbol{s})$ in (7) when defining the residuals $\{\widetilde{\varepsilon}(t_i, \boldsymbol{s}_{ij})\}$, then the mean estimate in (7) can be further updated by using the updated estimates of the variance and covariance functions, and so forth. However, by theoretical justification and numerical results in Sects. 3 and 4, it is found that no substantial performance gain can be obtained by using such an iterative procedure, but the computational burden of the iterative procedure would be heavy. For these reasons, we suggest using $\widehat{\lambda}(t, \boldsymbol{s})$ in (7) as the final estimate of $\lambda(t, \boldsymbol{s})$.

2.3 Selection of the bandwidths

The bandwidths $h_{t,0}$ and $h_{s,0}$ are used in obtaining the initial mean estimate $\widetilde{\lambda}(t, \boldsymbol{s})$. In the literature of univariate nonparametric estimation of regression functions from correlated data, it has been well discussed that the bandwidths selected by the conventional cross-validation (CV) procedure (e.g. the leave-one-out CV) would not perform well, because it cannot properly distinguish the data correlation structure from the data mean function (cf., Altman 1990; Brabanter et al. 2011; Opsomer et al. 2001). To overcome this difficulty, Hal95 suggested a block bootstrap procedure for choosing the bandwidth in the univariate nonparametric regression setup when observed data are correlated. This block-bootstrap procedure can be extended to multivariate cases for choosing the bandwidths $(h_{t,0}, h_{s,0})$ used in (4), which is described below. Let $(h_{t,01}, h_{s,01})$ and $(h_{t,02}, h_{s,02})$ be two sets of pre-specified values of $(h_{t,0}, h_{s,0})$, with $(h_{t,01}, h_{s,01})$ chosen relatively small and $(h_{t,02}, h_{s,02})$ chosen relatively large. The corresponding initial estimates of $\lambda(t, \boldsymbol{s})$ are denoted as $\widetilde{\lambda}_1(t, \boldsymbol{s})$ and $\widetilde{\lambda}_2(t, \boldsymbol{s})$, respectively. Let $\widetilde{\varepsilon}_{ij,1} = y(t_i, \boldsymbol{s}_{ij}) - \widetilde{\lambda}_1(t_i, \boldsymbol{s}_{ij})$ be the residuals, $\bar{\widetilde{\varepsilon}}_{1\cdot} = (\sum_{i=1}^n \sum_{j=1}^{m_i} \widetilde{\varepsilon}_{ij,1})/(\sum_{i=1}^n m_i)$ be the mean residual, and $\widetilde{\varepsilon}_{ij,0} = \widetilde{\varepsilon}_{ij,1} - \bar{\widetilde{\varepsilon}}_{1\cdot}$ be the centralized residuals. Assume that b is a pre-specified block size, k_1 is the largest integer that is less than or equal to n/b, and $k_2 = n - b \times k_1$. Then, our suggested block-bootstrap mean average squared error (BB-MASE) criterion for choosing the bandwidths $(h_{t,0}, h_{s,0})$ is described below.

1) Randomly choose a sequence of $k_1 + 1$ integers from $\{1, 2, ..., n - b + 1\}$ with replacement. The selected integers are denoted as $\{i_1, i_2, \ldots, i_{k_1+1}\}$. Then, we define a sequence of n indices as $\{i_1, i_{1+1}, \ldots, i_1 + b - 1, i_2, i_2 + 1, \ldots, i_2 + b - 1, \ldots, i_{k_1}, i_{k_1} + 1, \ldots, i_{k_1} + b - 1, i_{k_{1+1}}, i_{k_{1+1}} + 1, \ldots, i_{k_{1+1}} + k_2 - 1\}ifk_2 > 0$;and $\{i_1, i_1 + 1, \ldots, i_1 + b - 1, i_2, i_2 + 1, \ldots, i_2 + b - 1, \ldots, i_{k_1}, i_{k_1} + 1, \ldots, i_{k_1} + b - 1\}$ if $k_2 = 0$. Let $\{l_i, i = 1, 2, \ldots, n\}$ denote this sequence of n indices. Define the block-bootstrap sample to be $y^*(t_i, \boldsymbol{s}_{l_ij}) = \widetilde{\lambda}_2(t_i, \boldsymbol{s}_{l_ij}) + \widetilde{\varepsilon}_{l_ij,0}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, m_{l_i}$. For given bandwidths $(h_{t,0}, h_{s,0})$, the initial estimate of $\lambda(t, \boldsymbol{s})$ computed from the block-bootstrap sample is denoted as $\widetilde{\lambda}^*(t_i, \boldsymbol{s}_{l_ij})$. Then, we define ASE $(h_{t,0}, h_{s,0}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m_{l_i}} \sum_{j=1}^{m_{l_i}} (\widetilde{\lambda}_2(t_i, \boldsymbol{s}_{l_ij}) - \widetilde{\lambda}^*(t_i, \boldsymbol{s}_{l_ij}))^2$.
2) Step 1) is then repeated for B times, and the average of the B values of ASE $(h_{t,0}, h_{s,0})$ is denoted as BB-MASE $(h_{t,0}, h_{s,0})$.
3) The bandwidths $(h_{t,0}, h_{s,0})$ are determined by minimizing BB-MASE $(h_{t,0}, h_{s,0})$.

Hall et al. (1995) showed theoretically that the block-bootstrap procedure, such as the one described above, should be robust to the two sets of pre-specified values of the bandwidths to choose, which has been illustrated by us through numerical simulations.

For choosing the bandwidths $(h_{t,1}, h_{s,1})$ that are used for estimating the variance and covariance functions in (5) and (6), we suggest a new method based on spatio-temporal prediction, described below. First, define the cross-validation mean squared prediction error (CV-MSPE) by

$$\text{CV-MSPE}\ (h_{t,1}, h_{s,1}) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m_i}\sum_{j=1}^{m_i}\left\{y(t_i, \boldsymbol{s}_{ij}) - \widehat{y}_{-(ij)}(t_i, \boldsymbol{s}_{ij})\right\}^2\right], \tag{8}$$

where $\{\widehat{y}_{-(ij)}(i, \boldsymbol{s}_{ij}), j = 1, \ldots, m_i, i = 1, \ldots, n\}$ are the predicted values obtained by the kriging method (Cressie and Wikle 2011), described below. For each $1 \leq j \leq m_i$ and $1 \leq i \leq n$, let $\widehat{V}_{-(ij)}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ be the estimated covariance function by (6) when the (i, j)-th residual $\widetilde{\varepsilon}(t_i, \boldsymbol{s}_{ij})$ is omitted, $\boldsymbol{\varepsilon}_{-(ij)}$ be the vector with elements $\{\varepsilon(t_k, \boldsymbol{s}_{kl}), l = 1, \ldots, m_k, k = 1, \ldots, n, (k, l) \neq (i, j)\}$, $\boldsymbol{\Sigma}_{ij,-(ij)}$ be the covariance matrix between $\varepsilon(t_i, \boldsymbol{s}_{ij})$ and $\boldsymbol{\varepsilon}_{-(ij)}$, and $\boldsymbol{\Sigma}_{-(ij),-(ij)}$ 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

$$\widehat{y}_{-(ij)}(t_i, \boldsymbol{s}_{ij}) = \widetilde{\lambda}(t_i, \boldsymbol{s}_{ij}) + \widehat{\boldsymbol{\Sigma}}_{ij,-(ij)}^{T} \widehat{\boldsymbol{\Sigma}}_{-(ij),-(ij)}^{-1} \widehat{\boldsymbol{\varepsilon}}_{-(ij)}, \tag{9}$$

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

Note that $\widehat{\boldsymbol{\Sigma}}_{-(ij),-(ij)}$ is a $(\sum_{i=1}^{n} m_i - 1) \times (\sum_{i=1}^{n} m_i - 1)$ matrix. When the total sample size $\sum_{i=1}^{n} m_i$ is large, the computation and storage for the inverse matrix $\widehat{\boldsymbol{\Sigma}}_{-(ij),-(ij)}^{-1}$ 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 $\widehat{y}_{-(ij)}(t_i, \boldsymbol{s}_{ij})$ used in kriging procedure (8)–(9). More specifically, let $\Delta_{ij}(\theta_t, \theta_s) = \{(k, l) : |t_k - t_i| \leq \theta_t, d_E(\boldsymbol{s}_{kl}, \boldsymbol{s}_{ij}) \leq \theta_s, (k, l) \neq (i, j)\}$ be a set of indices around (i, j). Then, when computing $\widehat{\boldsymbol{\Sigma}}_{ij,-(ij)}$ and $\widehat{\boldsymbol{\Sigma}}_{-(ij),-(ij)}$ used in (9), only those points $(t_k, \boldsymbol{s}_{kl})$ whose indices are included in $\Delta_{ij}(\theta_t, \theta_s)$ are used. Also, $\widehat{\boldsymbol{\varepsilon}}_{-(ij)}$ in (9) needs to be replaced by the one that includes the residuals with indices in $\Delta_{ij}(\theta_t, \theta_s)$ only. After this modification, the computation involved in calculating the predicted value $\widehat{y}_{-(ij)}(t_i, \boldsymbol{s}_{ij})$ can be greatly reduced, and the amount of computational reduction is controlled by the parameters θ_t and θ_s. 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 our extensive numerical studies, we suggest choosing θ_t and θ_s such that $\theta_t \geq 5/n$ and $\theta_s \geq 3\theta(i, j)$, where $\theta(i, j) = \min\{d_E(\boldsymbol{s}_{ij}, \boldsymbol{s}_{il}), l = 1, \ldots, m_i, l \neq j\}$.

To determine the bandwidths $(h_{t,2}, h_{s,2})$ used in obtaining the final estimate $\widehat{\lambda}(t, \boldsymbol{s})$ in (7), we first notice that the mean of the residual mean squares (RMS) of $\widehat{\lambda}(t, \boldsymbol{s})$ is

$$\begin{aligned} E(\text{ RMS}) =& E\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{m_i}\sum_{j=1}^{m_i}\left(y(t_i, \boldsymbol{s}_{ij}) - \widehat{\lambda}(t_i, \boldsymbol{s}_{ij})\right)^2\right\}\right] \\ =& \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\sigma^2(t_i, \boldsymbol{s}_{ij}) + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}E\left(\lambda(t_i, \boldsymbol{s}_{ij}) - \widehat{\lambda}(t_i, \boldsymbol{s}_{ij})\right)^2 - \\ & \frac{2}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\boldsymbol{e}_1^T\left(X^T\widetilde{\Sigma}_K^{-1}X\right)^{-1}X^T\widetilde{\Sigma}_K^{-1}\Sigma_{\mathbf{Y}}(\zeta_{ij}), \end{aligned} \tag{10}$$

where $\Sigma_{\mathbf{Y}}(\zeta_{ij})$ is the ζ_{ij}th column of the covariance matrix $\Sigma_{\mathbf{Y}}$, and $\zeta_{ij} = \sum_{k=1}^{i-1} m_k + j$. In (10), the first term on the right-hand side is not related to $(h_{t,2}, h_{s,2})$, the second term is the mean square error (MSE) of $\widehat{\lambda}(t, \boldsymbol{s})$ that measures its performance, and the third term is due to data correlation. In our proposed bandwidth selection procedure, we suggest estimating the third term and then choosing the bandwidths by minimizing the sum of the RMS and the estimated third term (Note: the sum should be a good estimate of the sum of the first two terms on the right-hand side of (10)). More specifically, we first define a bias-corrected estimate of the MSE (BCE-MSE) of $\widehat{\lambda}(t, \boldsymbol{s})$ (note: the first term on the right-hand side of (10) has been ignored) to be

$$\begin{aligned} \text{BCE-MSE } (h_{t,2}, h_{s,2}) =& \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{m_i}\sum_{j=1}^{m_i}\left(y(t_i, \boldsymbol{s}_{ij}) - \widehat{\lambda}(t_i, \boldsymbol{s}_{ij})\right)^2\right\} + \\ & \frac{2}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\boldsymbol{e}_1^T\left(X^T\widetilde{\Sigma}_K^{-1}X\right)^{-1}X^T\widetilde{\Sigma}_K^{-1}\widetilde{\Sigma}_{\mathbf{Y}}(\zeta_{ij}), \end{aligned} \tag{11}$$

where $\widetilde{\Sigma}_{\mathbf{Y}}(\zeta_{ij})$ is the ζ_{ij}th column of $\widetilde{\Sigma}_{\mathbf{Y}}$. Then, $(h_{t,2}, h_{s,2})$ are chosen to minimize BCE-MSE $(h_{t,2}, h_{s,2})$.

3 Statistical properties

In this section, we present some statistical properties of the estimates $\widetilde{\lambda}(t,\boldsymbol{s})$, $\widehat{\sigma}^2(t,\boldsymbol{s})$, $\widehat{V}(t,t';\boldsymbol{s},\boldsymbol{s}')$ and $\widehat{\lambda}(t,\boldsymbol{s})$ defined in (4)–(7). In the discussion, it is assumed that $\{m_i, i=1,\dots,n\}$ are in the same order of m. It is further assumed that $\{\boldsymbol{s}_{ij}, i=1,\dots,n, j=1,\dots,m_i\}$ 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 $\{\boldsymbol{s}_{ij}, j=1,\dots,m_i\}$ are independent. Furthermore, $\{\boldsymbol{s}_{ij}, i=1,\dots,n, j=1,\dots,m_i\}$ are assumed to be independent of the random errors $\{\varepsilon(t_i,\boldsymbol{s}_{ij}), j=1,\dots,m_i, i=1,\dots,n\}$ in model (1). All these assumptions on the spatial locations are denoted as *(Assumption-SL)*.

For the random errors $\{\varepsilon(t_i,\boldsymbol{s}_{ij}), i=1,\dots,n, j=1,\dots,m_i\}$, it is assumed that $\varepsilon(t_i,\boldsymbol{s}_{ij})$ consists of two independent components, namely, $\varepsilon(t_i,\boldsymbol{s}_{ij})=\varepsilon_0(t_i,\boldsymbol{s}_{ij})+\varepsilon_1(t_i,\boldsymbol{s}_{ij})$, where $\varepsilon_0(t_i,\boldsymbol{s}_{ij})$ is from a spatially and temporally correlated random process and $\varepsilon_1(t_i,\boldsymbol{s}_{ij})$ is the pure measurement error (i.e. $\{\varepsilon_1(t_i,\boldsymbol{s}_{ij})\}$ are independent at different times and/or locations). Let the strong mixing coefficient of $\{\varepsilon_0(t_i,\boldsymbol{s}_{ij})\}$ in the time domain be defined as

$$\alpha(k)=\sup_{n\ge1,1\le i\le n-k}\ \sup_{A,B}\{|P(AB)-P(A)P(B)|\ :\ A\in\mathcal{F}_1^i, B\in\mathcal{F}_{i+k}^n\},$$

where $\mathcal{F}_{k_0}^{k_1}$ is the σ-algebra generated by $\{\varepsilon_0(t_k,\boldsymbol{s}_{kl}), k_0\le k\le k_1, l=1,\dots,m_k\}$. In addition, for any $(t,\boldsymbol{s})=[0,1]\times\Omega$, define $g_n(\theta;t,\boldsymbol{s})=V(t,t;\boldsymbol{s},\boldsymbol{s})+2\sum_{k=1}^{\tau_n}V(t,t+k/n;\boldsymbol{s},\boldsymbol{s})\cos(2\theta k\pi)$ to be the spectral density function of the time series $\{\varepsilon_0(t_i,\boldsymbol{s}), i=1,\dots,n\}$, where $\theta\in[-1/2,1/2]$, $\tau_n=\lfloor n(1-t)\rfloor$ is the greatest integer less than or equal to $n(1-t)$, and $V(t,t;\boldsymbol{s},\boldsymbol{s})$ is the variance function of $\varepsilon_0(t,\boldsymbol{s})$. Let $g(\theta;t,\boldsymbol{s})=\lim_{n\to\infty}g_n(\theta;t,\boldsymbol{s})$. Next, we will show that $g(\theta;t,\boldsymbol{s})$ is well-defined, and $\widehat{\lambda}(t,\boldsymbol{s})$ defined in (7) is a consistent estimator of $\lambda(t,\boldsymbol{s})$ under some regularity conditions.

Lemma 1 *Assume that there exist two constants* $C_0, C_1>0$ *such that the strong mixing coefficient* $\alpha(k)$ *satisfies the condition that* $\alpha(k)\le C_0\exp(-C_1k)$ *for every* k, *and there are some constants* $\delta>2$ *and* $0\le C_\varepsilon<\infty$ *such that* $E|\varepsilon(t,\boldsymbol{s})|^\delta\le C_\varepsilon$, *for all* t *and* $\boldsymbol{s}$. *Then, for any* $(t,\boldsymbol{s})\in[0,1]\times\Omega$, $g(\theta;t,\boldsymbol{s})$ *is well-defined and nonnegative.*

Lemma 2 *In model* (1), *it is assumed that the design space* Ω *is a compact set in* $\mathbb{R}^2$, *the observation locations* $\{\boldsymbol{s}_{ij}, i=1,\dots,n, j=1,\dots,m_i\}$ *satisfy the assumptions in* (*Assumption-SL*), $f(\boldsymbol{s})$ *is twice continuously differentiable in* Ω *and it has a nonzero 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* $\alpha(k)\le C_0\exp(-C_1k)$, *where* $C_0, C_1>0$ *are two constants,* δ *defined in Lemma* 1 *is larger than* 5, *the kernel functions* $K_1(\cdot)$ *and* $K_2(\cdot)$ *are bounded, symmetric, and Lipschitz*-1 *continuous density functions with finite supports*, $\log(n)^2/(nh_{t,0}^2)=o(1)$, $\log(n)/(mh_{s,0}^2)=O(1)$, $h_{s,0}=o(1)$ *and* $h_{t,0}/h_{s,0}=O(1)$. *Then, we have*

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

Lemma 3 *Besides the conditions in Lemma* 2, *if we further assume that there are some positive constants* C_2, C_3 *and* δ^* *such that* $\Pr(|\varepsilon(t,\boldsymbol{s})| \ge k) \le C_2 k^{\delta^*} \exp(-C_3 k)$, *for all* t, $\boldsymbol{s}$ *and any positive number* k, *the variance function* $\sigma^2(t,\boldsymbol{s})$ *is twice continuously differentiable in* $[0,1]\times\Omega$, *the covariance function* $V(t,t';\boldsymbol{s},\boldsymbol{s}')$ *is also twice continuously differentiable*, $\log(n)^2/(nh_{t,1}^2) = o(1)$, $\log(n)/(mh_{s,1}^2) = O(1)$, $h_{s,1} = o(1)$, *and* $h_{t,1}/h_{s,1} = O(1)$, *then we have*

$$\sup_{(t,\boldsymbol{s})\in[0,1]\times\Omega} \left|\widehat{\sigma}^2(t,\boldsymbol{s}) - \sigma^2(t,\boldsymbol{s})\right| = O_p\Big(h_{t,0}^2 + h_{s,0}^2 + \{\log(n)^2/(nh_{t,0}^2)\}^{1/2} + h_{t,1}^2 + h_{s,1}^2 + \{\log(n)^2/(nh_{t,1}^2)\}^{1/2}\Big), \tag{13}$$

and

$$\sup_{t,t'\in[0,1]} \sup_{\boldsymbol{s},\boldsymbol{s}'\in\Omega} \left|\widehat{V}(t,t';\boldsymbol{s},\boldsymbol{s}') - V(t,t';\boldsymbol{s},\boldsymbol{s}')\right| = O_p\Big(h_{t,0}^2 + h_{s,0}^2 + \{\log(n)^2/(nh_{t,0}^2)\}^{1/2} + h_{t,1}^2 + h_{s,1}^2 + \{\log(n)^2/(nh_{t,1}^2)\}^{1/2}\Big). \tag{14}$$

Theorem 1 *Besides the assumptions in Lemma* 3, *if we further assume that* $g(\theta;t,\boldsymbol{s}) > 0$, *for all* $\theta \in [-1/2, 1/2]$ *and* $(t,\boldsymbol{s}) \in [0,1]\times\Omega$, $1/(nh_{t,2}) = o(1)$, $\log(n)/(mh_{s,2}^2) = O(1)$, $h_{t,2} = o(1)$, *and* $h_{s,2} = o(1)$, *then we have*

$$\left|\widehat{\lambda}(t,\boldsymbol{s}) - \lambda(t,\boldsymbol{s})\right| = O_p\Big(h_{t,2}^2 + h_{s,2}^2 + \{1/(nh_{t,2})\}^{1/2}\Big). \tag{15}$$

The proofs of Lemmas 1–3 and Theorem 1 are given in the supplementary materials. In Lemma 3, we assume that $\Pr(|\varepsilon(t,\boldsymbol{s})| \ge k) \le C_2 k^{\delta^*} \exp(-C_3 k)$. This assumption is valid for many commonly used continuous distributions, including normal, exponential and Laplace distributions. In Theorem 1, it is assumed that $g(\theta;t,\boldsymbol{s}) > 0$ for all θ, t, and $\boldsymbol{s}$. It can be checked that this assumption is valid if the temporally correlated random fields $\{\varepsilon_0(t_i,\boldsymbol{s}), i = 1,\dots,n\}$ are generated from some stationary time series models (e.g. the AR models), or some commonly used covariance models, such as the Spherical, Matérn or Gaussian process models. See Choi et al. (2013) for some detailed discussions about these models. Furthermore, it can be checked that this assumption is also valid when $\{\varepsilon_0(t_i,\boldsymbol{s}), i = 1,\dots,n\}$ follow the models mentioned above locally.

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, assume that the observation times are $\{t_i = i/n, i = 1,\dots,n\}$, the observation locations at each time point are equally spaced in $\Omega = [0,1]\times[0,1]$ and they do not change over time, and the number of observation locations is m at each time point. In such cases, the observation locations are denoted as $\{\boldsymbol{s}_j, j = 1,\dots,m\}$. In all simulation examples, (m, n) are chosen to be (36, 50) or (100, 100), and the mean function is chosen to be

$$\lambda(t, \boldsymbol{s}) = 2 + \sin(\pi s_x)\sin(\pi s_y) + \sin(2\pi t),$$

where $\boldsymbol{s} = (s_x, s_y)^T$. Let $\boldsymbol{\varepsilon}(t_i) = \left(\varepsilon(t_i, \boldsymbol{s}_1), \dots, \varepsilon(t_i, \boldsymbol{s}_m)\right)^T$, then $\boldsymbol{\varepsilon}(i)$ is generated from following AR(1) process:

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

where $\left\{\boldsymbol{\eta}(t_i) = (\eta(t_i, \boldsymbol{s}_1), \dots, \eta(t_i, \boldsymbol{s}_m))^T, i = 1, \dots, n\right\}$ are temporally independent spatial processes and $-1 < \phi_t < 1$ is a constant controlling the temporal data correlation. For each $(t_i, \boldsymbol{s}_j)$, $\eta(t_i, \boldsymbol{s}_j)$ is generated from the normal distribution $N(0, \sigma^2)$. At each observation time t_i, the spatial covariance among the elements of $\boldsymbol{\eta}(t_i)$ is described by Cov $\left(\eta(t_i, \boldsymbol{s}_j), \eta(t_i, \boldsymbol{s}_l)\right) = \sigma^2 \rho(d_E(\boldsymbol{s}_j, \boldsymbol{s}_l))$, for any j and l, where $\rho(d) = \exp\{-\phi_s d\}$ and $\phi_s > 0$ is a constant to determine the magnitude of spatial correlation. 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 $V(t_i, t_k; \boldsymbol{s}_j, \boldsymbol{s}_l) = \sigma^2 \phi_t^{n|t_i - t_k|} \rho(d_E(\boldsymbol{s}_j, \boldsymbol{s}_l))$, for $t_i, t_k \in [0, 1]$. In the simulation examples, we choose $\sigma = 0.5$, $\phi_t = 0.3$, 0.6 or 0.9, and $\phi_s = 1$, 3 or 5. To determine the bandwidths $(h_{t,0}, h_{s,0})$ used for computing $\widetilde{\lambda}(t, \boldsymbol{s})$, the block-bootstrap procedure described in Sect. 2.3 is used, where we choose $B = 100$, $(h_{t,01}, h_{s,01}) = (0.05, 0.05)$, and $(h_{t,02}, h_{s,02}) = (0.3, 0.3)$. The block size b is always chosen to be 5, except in cases when we study the effect of block size on the performance of $\widehat{\lambda}(t, \boldsymbol{s})$ below. The threshold value c_n is chosen to be 20 when we compute the covariance function estimate.

First, we compare the proposed method with some representative existing methods: the DSTM and LGCP methods discussed in Sect. 1, the spatially weighted average (SWA) method suggested by Kafadar (1996), and the local linear kernel smoothing (LLKS) method by Yang and Qiu (2018). The DSTM method has been discussed in detail in citeStr01 and can be accomplished using the R-package *spBayes*. In the LGCP method, it is assumed that $y(t_i, \boldsymbol{s}_{ij})\mathcal{N}(t_i, \boldsymbol{s}_{ij})$ follows the distribution $Poisson(R(t_i, \boldsymbol{s}_{ij}))$, for each i and j, where $\mathcal{N}(t_i, \boldsymbol{s}_{ij})$ denotes the population size at time t_i in the design grid cell that contains $\boldsymbol{s}_{ij}$, $R(t_i, \boldsymbol{s}_{ij}) = C_g \mathcal{N}(t_i, \boldsymbol{s}_{ij}) \exp[\xi(t_i, \boldsymbol{s}_{ij})]$, $C_g = 1/m$ is the area of each grid cell, and $\xi(t_i, \boldsymbol{s}_{ij})$ is from a spatio-temporal Gaussian process with a variance parameter σ_ξ^2 and a scale parameter ϕ_ξ. In the simulation studies, $\{\mathcal{N}(t_i, \boldsymbol{s}_{ij})\}$ are fixed to be a constant 10,000. The parameters σ_ξ^2 and ϕ_ξ can be estimated by the Bayesian inference, as discussed in Taylor et al. (2015). This method can be accomplished using the R-package *lgcp*. The proposed initial mean estimate $\widetilde{\lambda}(t, \boldsymbol{s})$ in (4) is denoted as Step1, and the proposed updated mean estimate $\widehat{\lambda}(t, \boldsymbol{s})$ in (7) is denoted as Step2. Intuitively, because the DSTM and LGCP methods require some parametric assumptions on their models and on the data distribution, it is expected that they cannot perform well when these assumptions are violated. The SWA method was developed for smoothing spatial geographical data, and thus it cannot use data information at neighboring time points when estimating the mean function at a given time point. For the LLKS method, although it provides a reliable estimate for the spatio-temporal mean function, the covariance structure is not taken into consideration in its estimate and thus its efficiency has room for improvement. Based on these intuitions, we believe that our proposed method Step2 would perform favorably in comparison with its peers when estimating $\lambda(t, s)$ in cases when the mean and variance structures of the observed spatio-temporal data are complex. For Step1 and Step2, all their bandwidths are chosen by the procedures discussed in Sect. 2.3. The bandwidths used in the method LLKS are chosen according to the modified CV procedure described by Yang and Qiu (2018). For the other three competing methods DSTM, LGCP and SWA, their tuning parameters are chosen to minimize the mean average squared errors (MASE) defined below. To evaluate the performance of each method, the MASE criterion is defined as:

$$\text{MASE } = E\left\{ \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}(\lambda(t_i, s_j) - \hat{\lambda}(t_i, s_j))^2 \right\},$$

which is approximated by the average of 100 ASE values computed from 100 repeated simulations. The results of MASE values of different methods in the above setup are presented in Table 1. From the table, it can be seen that i) the three methods LLKS, Step1 and Step2 have a better overall performance than the methods

Table 1 Estimated MASE values and their standard errors (in parentheses) for six spatio-temporal methods when the sample size (m, n) changes from (36,50) to (100,100) and the spatio-temporal data correlation changes from relatively weak to relatively strong cases. *The standard error numbers in parentheses are in the scale of 1×10^{-3}. In each row, the smallest MASE value is in bold

(m, n)	(ϕ_t, ϕ_s)	DSTM	LGCP	SWA	LLKS	Step1	Step2
(36,50)	(0.3,1)	0.218(3.40*)	0.235(2.33)	0.237(3.41)	0.054(1.81)	0.050(1.73)	**0.048**(1.85)
	(0.3,3)	0.175(1.73)	0.231(1.55)	0.215(1.78)	0.044(0.99)	0.044(0.96)	**0.042**(1.12)
	(0.3,5)	0.152(1.13)	0.228(1.17)	0.200(1.20)	0.038(0.72)	0.037(0.67)	**0.036**(0.84)
	(0.6,1)	0.238(4.55)	0.241(4.64)	0.239(4.52)	0.117(3.50)	0.086(3.01)	**0.077**(3.12)
	(0.6,3)	0.202(2.40)	0.230(2.14)	0.216(2.36)	0.094(1.86)	0.076(1.69)	**0.064**(1.82)
	(0.6,5)	0.185(1.62)	0.221(1.12)	0.201(1.57)	0.079(1.23)	0.068(1.15)	**0.054**(1.25)
	(0.9,1)	0.246(8.54)	0.255(8.12)	0.238(8.50)	0.197(8.38)	0.184(8.13)	**0.166**(7.96)
	(0.9,3)	0.237(4.97)	0.243(4.65)	0.217(4.86)	0.170(4.66)	0.156(4.58)	**0.130**(4.41)
	(0.9,5)	0.230(3.61)	0.237(3.65)	0.201(3.38)	0.139(3.14)	0.135(3.15)	**0.109**(3.19)
(100,100)	(0.3,1)	0.225(2.43)	0.233(1.78)	0.231(2.43)	0.041(1.10)	0.034(1.02)	**0.030**(1.03)
	(0.3,3)	0.192(1.26)	0.225(1.12)	0.202(1.26)	0.034(0.56)	0.029(0.72)	**0.026**(0.81)
	(0.3,5)	0.168(0.82)	0.208(0.67)	0.179(0.82)	0.026(0.34)	0.024(0.33)	**0.022**(0.47)
	(0.6,1)	0.233(3.26)	0.239(2.78)	0.231(3.25)	0.095(2.32)	0.064(2.02)	**0.047**(1.94)
	(0.6,3)	0.211(1.69)	0.224(1.74)	0.202(1.68)	0.071(1.12)	0.054(1.01)	**0.039**(1.11)
	(0.6,5)	0.194(1.09)	0.203(0.91)	0.179(1.07)	0.060(0.70)	0.046(0.64)	**0.033**(0.78)
	(0.9,1)	0.251(7.27)	0.244(6.21)	0.238(7.29)	0.194(7.04)	0.161(6.62)	**0.129**(6.25)
	(0.9,3)	0.241(3.61)	0.238(3.20)	0.206(3.58)	0.169(3.43)	0.133(3.20)	**0.102**(3.21)
	(0.9,5)	0.235(2.27)	0.231(2.11)	0.181(2.19)	0.137(2.07)	0.113(1.96)	**0.083**(2.06)

DSTM, LGCP and SWA, ii) Step1 outperforms LLKS uniformly, iii) Step2 outperforms Step1 uniformly and iv) the performance of Step1 and Step2 is improved when the sample size (m, n) is increased from (36, 50) to (100, 100). The last conclusion is consistent with the consistency results in Lemma 1 and Theorem 1 in Sect. 3. The conclusion iii) says that the update from Step1 to Step2 by accommodating the estimated variance and covariance functions (cf., (4)–(7)) is helpful for estimating the mean function $\lambda(t, s)$. The conclusion ii) confirms that the block-bootstrap bandwidth selection procedure described in Sect. 2.3 is effective, since the major difference between Step1 and LLKS is that the bandwidths in Step1 are selected by the block-bootstrap procedure, while the bandwidths in LLKS are selected by an alternative procedure. The conclusion i) confirms the benefits to use the nonparametric spatio-temporal modelling methods LLKS, Step1 and Step2, in which the spatio-temporal data correlation is accommodated, while the related models are kept flexible. As a comparison, the methods DSTM and LGCP both require some restrictive model assumptions, and the method SWA ignores the spatio-temporal data correlation completely.

To better perceive the performance improvement of the proposed method Step2 in comparison with alternative methods, we also present the percentage of improvement in MASE (PIMASE) values in Table 2, where PIMASE of an alternative method is defined as

$$\text{PIMASE} = \frac{\text{MASE of an alternative method} - \text{MASE of Step2}}{\text{MASE of Step2}}.$$

If the value of PIMASE is positive, then Step2 performs better than the alternative method in question in terms of MASE. From Table 2, it can be seen that i) Step2 is better than all five alternative methods uniformly, ii) PIMASE is uniformly larger when the sample size (m, n) is larger, and iii) the improvements of Step2 over the five alternative methods are quite substantial in most cases considered.

Table 2 The percentage of improvement in MASE (PIMASE) values when comparing the proposed method Step2 with five competing methods in cases considered in Table 1

	(m,n)=(36,50)					(m,n)=(100,100)				
(ϕ_t, ϕ_s)	DSTM	LGCP	SWA	LLKS	Step1	DSTM	LGCP	SWA	LLKS	Step1
(0.3,1)	354%	390%	394%	13%	4%	650%	677%	670%	37%	13%
(0.3,3)	316%	450%	412%	5%	5%	638%	765%	677%	31%	12%
(0.3,5)	322%	533%	456%	6%	3%	664%	845%	714%	18%	9%
(0.6,1)	209%	213%	210%	52%	12%	396%	409%	391%	102%	36%
(0.6,3)	216%	259%	238%	47%	19%	441%	474%	418%	82%	38%
(0.6,5)	243%	309%	272%	46%	26%	488%	515%	442%	82%	39%
(0.9,1)	48%	54%	43%	19%	11%	95%	89%	84%	50%	25%
(0.9,3)	82%	87%	67%	31%	20%	136%	133%	102%	66%	30%
(0.9,5)	111%	117%	84%	28%	24%	183%	178%	118%	65%	36%

In the previous simulation examples, the block length b used in the block-bootstrap procedure described in Sect. 2.3 for determining the bandwidths $(h_{t,0}, h_{s,0})$ is fixed at 5. In this part, we study the impact of b on the performance of the mean estimate $\widehat{\lambda}(t, \boldsymbol{s})$. In the setup of Table 1, let $(m, n) = (36, 50)$ and b change among 1, 3, 5, 7 and 9. The difference of average squared errors (DASE) of $\widehat{\lambda}(t, \boldsymbol{s})$ when two different block sizes are used is defined as

$$\text{DASE}\,(b_1, b_2) = \text{ ASE}\,(b_1) - \text{ ASE}\,(b_2),$$

where b_1 and b_2 are two different block sizes, and ASE (b_1) and ASE (b_2) are the ASE values of $\widehat{\lambda}(t, \boldsymbol{s})$ when block sizes b_1 and b_2 are used, respectively. To present the results in a concise way, we use $b = 5$ as a baseline case and compare other choices of b to this case. Namely, the values of DASE $(b_1, 5)$ are computed for $b_1 = 1, 3, 7$ or 9. The results of DASE $(b_1, 5)$ based on 100 replicated simulations are shown in Fig. 1 by box plots. From the plots, it can be seen that i) the results would not be good if we choose $b = 1$, ii) the results are good when $b = 3, 5$ or 7, and iii) the results would not be good if b is chosen too large. The conclusion i) confirms that block-bootstrap with block size $b > 1$ is necessary to accommodate data correlation. The conclusions ii) and iii) confirm that $b = 5$ is a reasonable choice.

As mentioned in Sect. 2.2, the mean estimate $\widehat{\lambda}(t, \boldsymbol{s})$ in (7) by three-step estimation procedure (3)–(7) can actually be further updated in an iterative way. In this part, we study whether the performance of the mean estimate can be improved substantially by using more iterations. More specifically, besides Step1 and Step2 considered in Table 1, we also consider Step3 and Step4, where Step3 denotes the mean estimate obtained by (7) after the estimates of the variance and covariance functions are updated by (5) and (6) with the initial mean estimate $\widetilde{\lambda}(t, \boldsymbol{s})$ replaced by the mean estimate $\widehat{\lambda}(t, \boldsymbol{s})$ of Step2 when defining the residuals, and Step4 denotes the mean estimate after another iteration. As in the previous example, to compare two methods, we use the DASE metric, with Step2 as the baseline method (i.e. DASE is defined to be ASE of an alternative method minus ASE of Step2). In the setup of

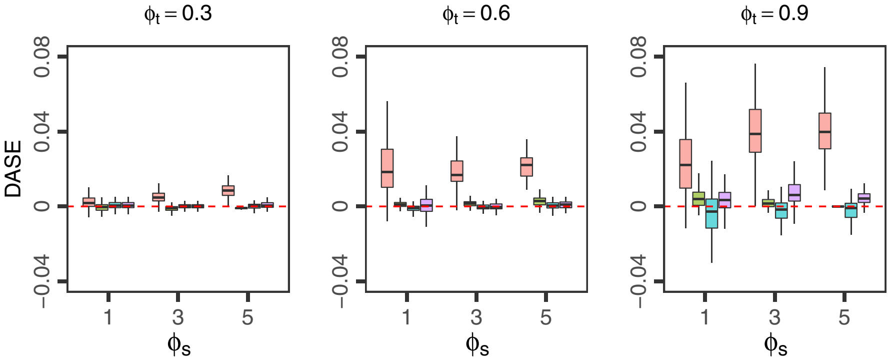

Fig. 1 Boxplots of the DASE values DASE $(b_1, 5)$, where block length $b_2 = 5$ is used as the baseline for comparing different values of the block size b used in the block-bootstrap procedure described in Sect. 2.3, in cases when $(m, n) = (36, 50)$, $\phi_s = 1, 3$ or 5, and $\phi_t = 0.3, 0.6$ or 0.9. In each plot, for given values of (ϕ_t, ϕ_s), the four boxes are for cases when $b_1 = 1$, 3, 7 and 9, respectively

Table 1 when $(m, n) = (36, 50)$, the results based on 100 replicated simulations are shown in Fig. 2 by box plots. From the plots of Fig. 2, it can be seen that i) Step2 is much better than Step1 (i.e. the first iteration is very helpful), and ii) Step3 and Step4 improve Step2 only marginally. Because of the computational burden with more iterations, we recommend using $\widehat{\lambda}(t, \boldsymbol{s})$ in (7) (i.e. the one obtained after the first iteration) as the final estimate of $\lambda(t, \boldsymbol{s})$.

In all previous examples, the design points $\{(t_i, \boldsymbol{s}_{ij}), i = 1, \ldots, n, j = 1, \ldots, m_i\}$ are deterministic and regularly spaced in $[0, 1] \times [0, 1]^2$. In this part, we consider cases when the spatial locations are generated randomly in $[0, 1]^2$, to investigate whether different types of design points would change the performance of the mean estimates. To this end, let us consider the same setup as that of Table 1 when $(m, n) = (36, 50)$, except that the observation locations $\{\boldsymbol{s}_j, j = 1, \ldots, m\}$ here are generated from the two-dimensional uniform distribution in $[0, 1]^2$. In such cases, the calculated MASE values of the six competing methods based on 100 replicated simulations are presented in Table 3, along with their standard errors. It can be seen that the results are similar to those in Table 1 with a fixed design.

At the end of this section, we would like to point out that the computation of the proposed method Step2 is relatively simple because $\lambda(t, \boldsymbol{s})$ and $V(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ are both estimated by local smoothing procedures in small neighborhoods. More specifically, when estimating $\lambda(t, \boldsymbol{s})$ by (7), only the observations in the neighborhood of $(t, \boldsymbol{s})$ (i.e. $\mathcal{N}(t, \boldsymbol{s}) = \{y(t_i, \boldsymbol{s}_{ij}) : |t_i - t| \leq h_{t,2}, d_E(\boldsymbol{s}_{ij}, \boldsymbol{s}) \leq h_{s,2}\}$) are used. Assume that there are N^* observations in $\mathcal{N}(t, \boldsymbol{s})$. Then, when we take the Moore–Penrose generalized inverse of the matrix $\widetilde{\Sigma}_K$, we only need to calculate the inverse of a related small submatrix of $\widetilde{\Sigma}_K$ with the dimension $N^* \times N^*$, because all other elements in $\widetilde{\Sigma}_K$ are 0. Note that the number N^* would be much smaller than the total number of observations mn. In the simulation examples when $(m, n) = (100, 100)$ and $(\phi_t, \phi_s) = (0.6, 3)$, the average computation time (ACT) for estimating the mean function $\lambda(t, \boldsymbol{s})$ at a given time and a given location is about 0.08 second using a Mac desktop with a 2.9 GHz Intel Core i5

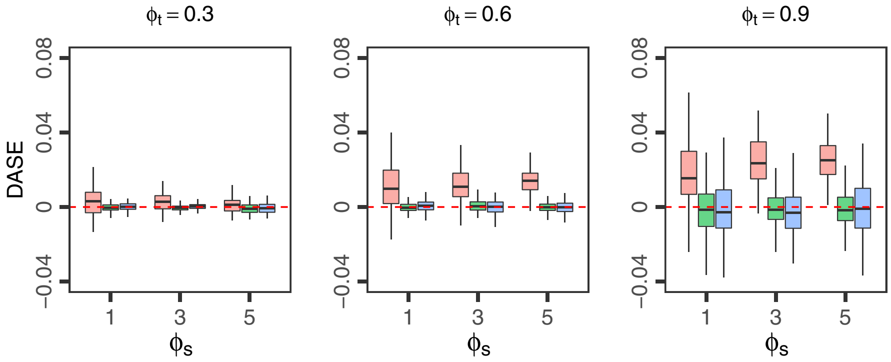

Fig. 2 Boxplots of the DASE values to compare mean estimates obtained after the first several iterations of the iterative model estimation of (1) in cases when $(m, n) = (36, 50)$, $\phi_s = 1, 3$ or 5, and $\phi_t = 0.3, 0.6$ or 0.9. For specific values of (ϕ_t, ϕ_s) in each plot, the three boxes are for comparing Step1, Step3 and Step4 with Step2, respectively, with Step2 as a baseline method

Table 3 Estimated MASE values and their standard errors (in parentheses) for six spatio-temporal methods when $(m, n) = (36, 50)$ and all other setups are the same as those in Table 1, except that the spatial locations here are generated from a uniform distribution in $[0, 1]^2$. *The standard error numbers in parentheses are in the scale of 1×10^{-3}. In each row, the smallest MASE values is in bold

(ϕ_t, ϕ_s)	DSTM	LGCP	SWA	LLKS	Step1	Step2
(0.3,1)	0.220(3.38*)	0.231(2.30)	0.233(3.41)	0.060(1.76)	0.051(1.71)	**0.046**(1.77)
(0.3,3)	0.181(1.77)	0.225(1.63)	0.194(1.81)	0.054(0.99)	0.049(0.98)	**0.043**(1.18)
(0.3,5)	0.160(1.20)	0.218(1.09)	0.167(1.26)	0.041(0.71)	0.041(0.70)	**0.040**(0.87)
(0.6,1)	0.233(4.55)	0.237(4.24)	0.235(4.52)	0.117(3.38)	0.091(3.05)	**0.079**(3.12)
(0.6,3)	0.206(2.46)	0.230(2.25)	0.195(2.41)	0.105(1.91)	0.083(1.76)	**0.066**(1.79)
(0.6,5)	0.190(1.71)	0.224(1.32)	0.168(1.70)	0.100(1.37)	0.076(1.25)	**0.061**(1.39)
(0.9,1)	0.246(8.60)	0.245(8.42)	0.233(8.57)	0.203(8.36)	0.193(8.19)	**0.171**(7.86)
(0.9,3)	0.237(5.20)	0.239(4.85)	0.196(5.03)	0.184(5.12)	0.182(4.99)	**0.152**(4.94)
(0.9,5)	0.231(3.91)	0.228(3.85)	0.172(3.79)	0.183(3.92)	0.169(3.78)	**0.139**(3.90)

processor, where ACT $= \{mn\}^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m}$ CT$_{ij}$ and CT$_{ij}$ is the computation time for computing $\widehat{\lambda}(t_i, \boldsymbol{s}_{ij})$. To further investigate the impact of (m, n) on ACT, consider cases when $m = 36, 64, 100, 225$ or 400, $n = 50, 100, 200, 300, 400$ or 500, and $(\phi_t, \phi_s) = (0.6, 3)$. The computed ACT values are presented in Figure S.1 of the supplementary file. From Figure S.1, it can be seen that ACT gets larger when m or n becomes larger and it requires about 3.28 seconds to compute the estimate $\widehat{\lambda}(t, \boldsymbol{s})$ when $(m, n) = (400, 500)$.

5 Application to a hand, foot and mouth disease dataset

In this section, we present an application of our proposed method to a hand, foot and mouth disease (HFMD) dataset that contains weekly incidence rates of HFMD in 21 cities of Sichuan Province in China during 2009-2010 (52 weeks). To implement the proposed method, the threshold value c_n is chosen to be 20, the block size b used in the block-bootstrap procedure is chosen to be 5, and the parameters $h_{t,01}$ and $h_{t,02}$ are chosen to be 0.05 and 0.3, respectively, as in the simulation examples. For the parameters $h_{s,01}$ and $h_{s,02}$, we first compute the largest distance between any two different cities and then choose $h_{s,01}$ and $h_{s,02}$ to be 0.05 and 0.3 times of that largest distance. The estimated mean function $\widehat{\lambda}(t, \boldsymbol{s})$ for 4 cities (Deyang, Luzhou, Neijiang and Ngawa Autonomous Prefecture) is presented in Fig. 3, along with the observed data.

Besides the mean estimate by Step2, we also consider the five alternative methods DSTM, LGCP, SWA, LLKS and Step1 that are discussed in Sect. 4. Since the true mean function $\lambda(t, \boldsymbol{s})$ is unknown in real-data applications, the performance metric MASE cannot be used in such cases. Instead, we use the following mean square prediction error (MSPE), defined as

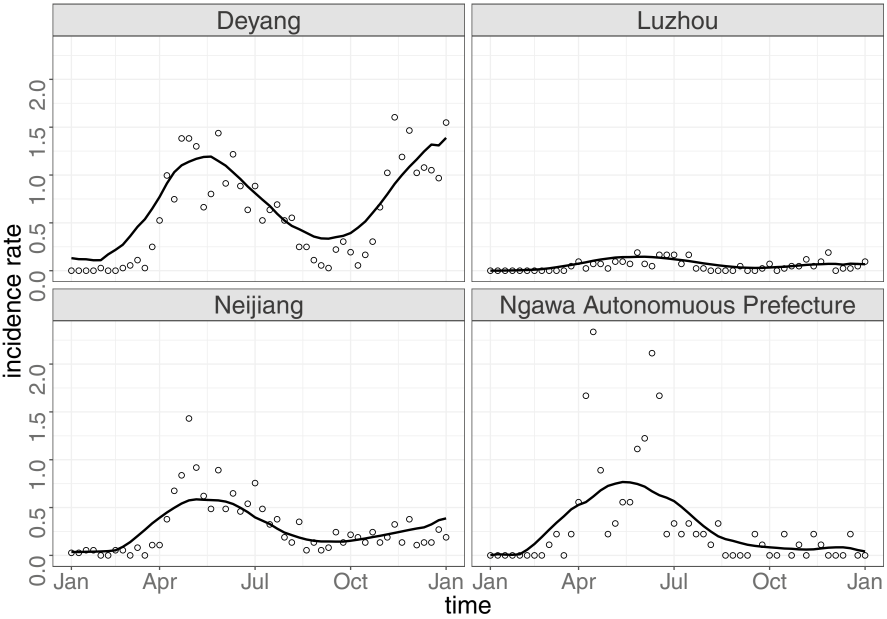

Fig. 3 Observed incidence rates (little circles) of the hand, foot and mouth disease in four cities of Sichuan Province in China, and the estimated mean incidence rate functions (solid curves). All incidence rates are in the scale of 10^{-5}

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{m_i}\sum_{j=1}^{m_i}\left(\widehat{\lambda}_{-(ij)}(t_i,\boldsymbol{s}_{ij}) - y(t_i,\boldsymbol{s}_{ij})\right)^2\right\},$$

and the mean absolute prediction error (MAPE), defined as

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{m_i}\sum_{j=1}^{m_i}\left|\widehat{\lambda}_{-(ij)}(t_i,\boldsymbol{s}_{ij}) - y(t_i,\boldsymbol{s}_{ij})\right|\right\},$$

where $\widehat{\lambda}_{-(ij)}(t_i,\boldsymbol{s}_{ij})$ is the leave-one-out estimate of $\lambda(t_i,\boldsymbol{s}_{ij})$ with the (i,j)th observation $y(t_i,\boldsymbol{s}_{ij})$ deleted when estimating $\lambda(t_i,\boldsymbol{s}_{ij})$. The calculated values of MSPE of DSTM, LGCP, SWA, LLKS, Step1 and Step2 are 6.26×10^{-11}, 4.58×10^{-11}, 4.66×10^{-11}, 2.87×10^{-11}, 2.24×10^{-11} and 1.91×10^{-11}, respectively, and their MAPE values are 5.61×10^{-6}, 3.97×10^{-6}, 4.06×10^{-6}, 3.18×10^{-6}, 2.85×10^{-6} and 2.61×10^{-6}. It can be seen that Step2 outperforms all 5 alternative methods quite substantially in terms of both MSPE and MAPE in this example.

To further investigate the six methods, the maps of the absolute residual values for the data in the 20th, 30th and 40th weeks are shown in columns 1–6 of Fig. 4. For each city, the residual at a given week is defined to be the absolute difference between the observed and estimated incidence rates. From these maps, it can be seen that the first four methods have relatively large residuals at some cities, while the residuals of Step1 and Step2 are relatively small.

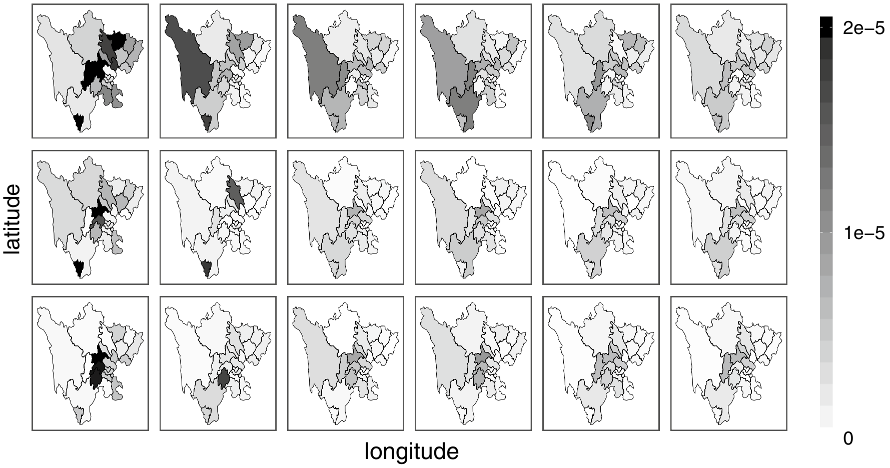

Fig. 4 Maps of the absolute residual values of the method DSTM (1st column), LGCP (2nd column), SWA (3rd column), LLKS (4th column), Step1 (5th column) and Step2 (6th column) in 21 cities of Sichuan Province in China at the 20th (1st row), 30th (2nd row) and 40th (3rd row) week in the year 2009

6 Concluding remarks

Spatio-temporal data are common in practice with many applications. Most existing methods either impose various restrictive model assumptions or ignore partially or completely the spatio-temporal data correlation during data modelling and estimation. In the previous several sections, we have presented a new three-step local smoothing procedure for jointly estimating the mean and variance/covariance functions of spatio-temporal data, in which both the mean structure and the variance/covariance structure are kept flexible. Because the proposed method is based on local smoothing, its computation is relatively simple. Effective bandwidth selection procedures are also developed for implementing the proposed method. Both the theoretical arguments and numerical studies have confirmed that the proposed method can work well in practice. There are still a number of issues that need to be addressed in our future research. For instance, bandwidths used in the proposed method are constants across the entire design time interval and space. In practice, curvature of the mean function $\lambda(t, \boldsymbol{s})$ and the data variability could be quite different in both time and space. Thus, variable bandwidths might be more reasonable to use. Also, there could be covariates (e.g. weather conditions) that affect the main response variable y in practice. In such cases, these covariates should be taken into account when modelling the observed data. To accommodate the covariate effect, the current spatio-temporal model could be generalized to a semiparametric model or a generalized additive model. Such possible generalizations are not straightforward and will be studied carefully in our future research. We will continue to work on these and some other research problems to make the proposed method more effective and powerful for analyzing spatio-temporal data.

Acknowledgements The authors would like to thank the editor, the associate editor, and two anonymous referees for their constructive comments and suggestions, which improved the quality of the paper greatly. This research was supported in part by the NSF Grant DMS-1914639 in USA.

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association, 85,* 749–759.

Brabanter, K. D., Brabanter, J. D., Suykens, J. A. K., De Moor, B. (2011). Kernel regression in the presence of correlated errors. *Journal of Machine Learning Research, 12,* 1955–1976.

Bradley, J. R., Holan, S. H., Wikle, C. K. (2015). Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *The Annals of Applied Statistics, 9,* 1761–1791.

Choi, I., Li, B., Wang, X. (2013). Nonparametric estimation of spatial and space-time covariance function. *Journal of Agricultural, Biological, and Environmental Statistics, 18,* 611–630.

Cressie, N., Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association, 94,* 1330–1340.

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

Dahlaus, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics, 25,* 1–37.

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A., Schaap, M. (2016). Nonseparable dynamic nearest-neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *The Annals of Applied Statistics, 10,* 1286–1316.

Diggle, P. J., Moraga, P., Rowlingson, B., Taylor, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science, 28,* 542–563.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications, 14,* 153–158.

Fonseca, T. C. O., Steel, M. F. J. (2011). Non-Gaussian spatiotemporal modelling through scale mixing. *Biometrika, 98,* 761–774.

Hall, P., Lahiri, S. N., Polzehl, J. (1995). On bandwidth choice in nonparametric regression with both short- and long-range dependent errors. *The Annals of Statistics, 23,* 1921–1936.

Heuvelink, G. B. M., Griffith, D. A. (2010). Space-time geostatistics for geography: a case study of radiation monitoring across parts of Germany. *Geographical Analysis, 42,* 161–179.

Higham, N. J. (1998). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications, 103,* 103–118.

Kafadar, K. (1996). Smoothing geographical data, particularly rates of disease. *Statistics in Medicine, 15,* 2539–2560.

Liang, K. Y., Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13–22.

Lindström, J., Szpiro, A., Sampson, P.D., Bergen, S., Sheppard, L. (2015), "Spatiotemporal: an R package for spatio-temporal modelling of air-pollution," https://cran.rproject.org/web/packages/SpatioTemporal/index.html.

Møller, J., Syversveen, A. R., Waagepetersen, R. P. (1998). Log-gaussian cox processes. *Scandinavian Journal of Statistics, 25,* 451–482.

Opsomer, J., Wang, Y., Yang, Y. (2001). Nonparametric regressin with correlated errors. *Statistical Science, 16,* 134–153.

Qiu, P. (2005). *Image Processing and Jump Regression Analysis*. New York: John Wiley.

Robinson, P. M. (1989). *Nonparametric estimation of time-varying parameters* (pp. 253–264). Statistical Analysis and Forecasting of Economic Structural Change, Berlin: Springer.

Shand, L., Li, B. (2017). Modeling nonstationarity in space and time. *Biometrics, 73,* 759–768.

Stroud, J. R., Müller, P., Sansó, B. (2001). Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society: Series B, 63,* 673–689.

Taylor, B. M., Davies, T. M., Rowlingson, B. S., Diggle, P. J. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-Gaussian Cox processes in R. *Journal of Statistical Software, 63,* 1–48.

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

Wikle, C. K., Hooten, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. *Test, 19,* 417–451.

Wikle, C. K., Zammit-Mangion, A., Cressie, N. (2019). *Spatio-Temporal Statistics with R*. Boca Raton, FL: Chapman Hall/CRC.

Yang, K., Qiu, P. (2018). Spatiotemporal incidence rate data analysis by nonparametric regression. *Statistics in Medicine, 37,* 2094–2107.

Yang, K., Qiu, P. (2019). Nonparametric estimation of the spatio-temporal covariance structure. *Statistics in Medicine, 38,* 4555–4565.