# Spatio-Temporal Incidence Rate Data Analysis By Nonparametric Regression

Kai Yang and Peihua Qiu

Department of Biostatistics

University of Florida

## Abstract

To monitor the incidence rates of cancers, AIDS, cardiovascular diseases, and other chronic or infectious diseases, some global, national and regional reporting systems have been built to collect/provide population-based data about the disease incidence. Such databases usually report daily, monthly, or yearly disease incidence numbers at the city, county, state or country level, and the disease incidence numbers collected at different places and different times are often correlated: with the ones closer in place or time being more correlated. The correlation reflects the impact of various confounding risk factors, such as weather, demographic factors, life styles, and other cultural and environmental factors. Because such impact is complicated and challenging to describe, the spatio-temporal (ST) correlation in the observed disease incidence data has complicated ST structure as well. Furthermore, the ST correlation is hidden in the observed data, and cannot be observed directly. In the literature, there has been some discussion about ST data modeling. But, the existing methods either impose various restrictive assumptions on the ST correlation that are hard to justify, or ignore partially or entirely the ST correlation. This paper aims to develop a flexible and effective method for ST disease incidence data modeling, using nonparametric local smoothing methods. This method can properly accommodate the ST data correlation. Theoretical justifications and numerical studies show that it works well in practice.

*Key Words:* Bandwidth; Consistency; Correlation; Cross-validation; Local smoothing; Residual map; Spatial data; Temporal correlation.

## 1 Introduction

Chronic diseases, including cardiovascular diseases, account for 70% of all deaths in the U.S., which is 1.7 million each year. In 2008 alone, cancer accounted for nearly $1 trillion in economic losses from premature death and disability (Bloom et al. 2011). To monitor the incidence of these deadly diseases, some global, national and regional reporting systems have been established to collect/provide population-based data about the disease incidence. Such reporting systems usually report daily, monthly or yearly disease incidence numbers at the city, county, state or country level. This paper suggests an effective and flexible method for analyzing such spatio-temporal (ST) data.

Figure 1 shows the lung cancer incidence rates in 58 counties of California during August 2005, reported by the Surveillance, Epidemiology, and End Results (SEER) program, with redder color denoting higher incidence rate. Such data usually have the following two features: 1) the exact address of a disease occurrence is unavailable, and 2) disease incidence rate numbers collected at different places and different times are correlated: with the ones closer in place or time being more correlated. This kind of ST correlation, however, is hidden in the observed data, and cannot be observed directly. Further, the ST correlation is related to the impact of various confounding risk factors, such as weather, demographic factors, life styles, and other cultural and environmental factors. Many confounding risk factors may not be included in the database because they are not our interest, or they are difficult to measure, or we are even unaware of their existence. Such complicated and latent impact on the observed data makes the ST correlation especially challenging to describe and estimate.

(put Figure 1 about here)

ST data analysis is an active interdisciplinary research area because of its broad applications in different fields, including geography, meteorology, oceanography, environment, epidemiology, global health, and medicine. There have been much existing research on this topic. Methods based on regression modeling include the one using temporal basis functions and "land use" linear regression that can be implemented using the *R* package *SpatioTemporal* (cf., Lindström et al. 2015), the kernel smoothing methods assuming independence among spatial observations (e.g., Kafadar 1996, Kelsall and Diggle 1998), the function estimation methods based on B-splines (e.g., Choi et al. 2013), and the one assuming ANOVA-type space-time error structure (Heuvelink and Griffith 2010). There are some alternative methods in the literature. One such approach describes the ST data by a log-Gaussian Cox process (LGCP), where the space-time disease incidence rate (see the definition in Section 2) in log scale is assumed to be a Gaussian process with a separable or non-separable covariance structure (e.g., Cressie and Huang 1999, Diggle et al. 2013). Another approach treats the ST data as a time series of spatial process realizations and works in the setting of a dynamic ST model (e.g., Finley et al. 2015). By such methods, the ST data are described using a measurement equation that builds a linear regression between the response variable and some covariates with a space-time varying intercept and serially and spatially uncorrelated zero-centered Gaussian disturbances. The regression coefficients at a given time are assumed to be those at the previous time plus normally distributed random errors, as specified by the transition equations. Because it is challenging to properly describe and accommodate the true ST correlation structure, as described above, almost all these existing methods put various assumptions on their models and/or the ST correlation structure. These assumed

ST data structures are often difficult to verify in practice. When they are valid, the related papers have shown the effectiveness of the related methods. When they are invalid, however, the effectiveness of the methods would be questionable.

In this paper, we propose an alternative ST modeling approach, which does not impose restrictive assumptions on the observation distribution, ST pattern of the disease incidence rate, and ST correlation in the observed data. This method is based on ST local smoothing for correlated data. Its estimated model is proved to be statistically consistent under some regularity conditions. Both simulation studies and real-data examples show that it works well in practice. The proposed method will be described in detail in Section 2. Some of its theoretical properties are discussed in Section 3. Some simulation results are presented in Section 4. Application of the proposed method to the lung cancer data shown in Figure 1 is presented in Section 5. Some concluding remarks are given in Section 6. Proofs of three theorems are given in the supplementary file.

## 2 Nonparametric Spatio-Temporal Modeling

Let $\Omega$ and $[0, T]$ be the region and time interval in which disease incidence trajectory is of our interest. For any $\mathbf{s} \in \Omega$ and $t \in [0, T]$, let $N(t, \mathbf{s}; dt, d\mathbf{s})$ be the number of disease cases in a small region $O(\mathbf{s})$ around $\mathbf{s}$ with area $d\mathbf{s}$ and in the time interval $[t, t + dt]$, and $M(t, O(\mathbf{s}))$ be the population size of the region $O(\mathbf{s})$ at time $t$. Then, $N(t, \mathbf{s}; dt, d\mathbf{s})$ is a measurement of the disease incidence in the region $O(\mathbf{s})$ and in the time interval $[t, t + dt]$, and $N(t, \mathbf{s}; dt, d\mathbf{s})/M(t, O(\mathbf{s}))$ is the corresponding disease incidence rate. Because the disease incidence depends heavily on the population size in the related region, the disease incidence rate is more appropriate to use when comparing disease occurrence in different regions.

However, the disease incidence rate defined above depends on the sizes of the region $O(\mathbf{s})$ and the time interval $[t, t+dt]$. After standardization on these sizes, the *(population) ST disease incidence rate* at time $t$ and location $\mathbf{s}$ is defined as

$$\lambda(t, \mathbf{s}) = \lim_{dt \to 0, d\mathbf{s} \to 0} \frac{E(N(t, \mathbf{s}; dt, d\mathbf{s}))}{M(t, O(\mathbf{s}))dtd\mathbf{s}}, \qquad \text{for } \mathbf{s} \in \Omega, \;\; t \in [0, T]. \tag{1}$$

Obviously, $\lambda(t, \mathbf{s})$ in (1) denotes the expected number of disease cases per population unit, per time unit at $t$, and per area unit at $\mathbf{s}$. This definition is commonly used in the epidemiology literature, although most people simply use its sample version, by replacing $E(N(t, \mathbf{s}; dt, d\mathbf{s}))$ with $N(t, \mathbf{s}; dt, d\mathbf{s})$ in (1) (cf., Last 2001).

## 2.1 The model

Assume that $y(t_i, \mathbf{s}_{ij})$ is the observed incidence rate at time $t_i$ and location $\mathbf{s}_{ij}$, and it follows the model

$$y(t_i, \mathbf{s}_{ij}) = \lambda(t_i, \mathbf{s}_{ij}) + \varepsilon(t_i, \mathbf{s}_{ij}), \qquad \text{for } j = 1, 2, \ldots, m_i, \; i = 1, 2, \ldots, n, \tag{2}$$

where $t_i \in [0, T]$, $\mathbf{s}_{ij} \in \Omega$, $\varepsilon(t_i, \mathbf{s}_{ij})$ is the zero-mean random error, $m_i$ is the number of observation locations at time $t_i$, and $n$ is the number of time points. Let $N = \sum_{i=1}^{n} m_i$ denote the total number of the points considered here. The ST correlation in the observed data can be described by the covariance function

$$V(\boldsymbol{u}; \boldsymbol{v}) = E\left[\varepsilon(\boldsymbol{u})\varepsilon(\boldsymbol{v})\right] = \sigma(\boldsymbol{u})\sigma(\boldsymbol{v})Corr(\varepsilon(\boldsymbol{u}), \varepsilon(\boldsymbol{v})), \qquad \text{for } \boldsymbol{u}, \boldsymbol{v} \in [0, T] \times \Omega, \tag{3}$$

where $\sigma^2(\cdot)$ is the variance function and $Corr(\cdot, \cdot)$ is the correlation function. Models (2)-(3) are general. They allow the number of observations $m_i$ and observation locations $\{\mathbf{s}_{ij}\}$ change over time. They do not impose any parametric forms on $\lambda(t_i, \mathbf{s}_{ij})$, $V(\boldsymbol{u}, \boldsymbol{v})$, and the error distribution. They even allow $V(\boldsymbol{u}, \boldsymbol{v})$ to be a nonparametric function of $\boldsymbol{u}$ and $\boldsymbol{v}$.

## 2.2 Model estimation

At a given $(t, \mathbf{s}) \in [0, T] \times \Omega$, we consider the following local linear ST kernel (LLSTK) smoothing procedure for estimating $\lambda(t, \mathbf{s})$:

$$\min_{a,b,c,d \in R} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left\{ y(t_i, \mathbf{s}_{ij}) - [a + b(t_i - t) + c(s_{x,ij} - s_x) + d(s_{y,ij} - s_y)] \right\}^2 \times \quad (4)$$
$$K_1\left(\frac{t_i - t}{h_t}\right) K_2\left(\frac{d_E(\mathbf{s}_{ij}, \mathbf{s})}{h_s}\right),$$

where $\mathbf{s} = (s_x, s_y)$, $\mathbf{s}_{ij} = (s_{x,ij}, s_{y,ij})$, $K_1$ and $K_2$ are two univariate kernel functions, $h_t > 0$ and $h_s > 0$ are two bandwidths, and $d_E(\cdot, \cdot)$ denotes the Euclidean distance. Let $\boldsymbol{X}$ be the design matrix with rows $\{(1, t_i - t, s_{x,ij} - s_x, s_{y,ij} - s_y), j = 1, 2, \ldots, m_i,\ i = 1, 2, \ldots, n\}$, and $\boldsymbol{Y}$ be the corresponding observation vector. Then, expression (4) can be re-written as

$$\min_{\boldsymbol{\beta} \in R^4} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \quad (5)$$

where $\boldsymbol{\beta} = (a, b, c, d)'$, $\boldsymbol{W} = diag\{K(\boldsymbol{H}^{-1}(t_1 - t, s_{x,11} - s_x, s_{y,11} - s_y)'), \ldots, K(\boldsymbol{H}^{-1}(t_n - t, s_{x,nm_n} - s_x, s_{y,nm_n} - s_y)')\}$, $K((t, s_x, s_y)') = K_1(t) K_2(d_E(\mathbf{s}, \mathbf{0}))$, $\boldsymbol{H} = diag\{h_t, h_s, h_s\}$, and $diag\{\boldsymbol{u}\}$ denotes a diagonal matrix with its diagonal elements given in the vector $\boldsymbol{u}$. Then, the estimated incidence rate is defined to be the solution to $a$ of the minimization problem (5) (or (4)), as

$$\widehat{\lambda}(t, \mathbf{s}) = \boldsymbol{e}_1' \left(\boldsymbol{X}' \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}' \boldsymbol{W} \boldsymbol{Y}, \quad (6)$$

where $\boldsymbol{e}_1 = (1, 0, 0, 0)'$. In Section 3, it will be shown that $\widehat{\lambda}(t, \mathbf{s})$ is statistically consistent under some regularity conditions.

## 2.3 Kernel and bandwidth selection

To use the above estimation procedure (4)-(6), the kernel functions $K_1$ and $K_2$ and the bandwidths $h_t$ and $h_s$ should be chosen properly in advance. Selection of these quantities is discussed in this subsection. To choose the kernel functions, let us use the following Mean Average Squared Error ($MASE$) as a metric for evaluating the performance of $\widehat{\lambda}(t, \mathbf{s})$:

$$MASE(h_t, h_s) = E\left\{\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\widehat{\lambda}(t_i, \mathbf{s}_{ij}) - \lambda(t_i, \mathbf{s}_{ij})\right]^2\right\}.$$

Based on this criterion, Theorem 2 in Section 3 says that the Epanechnikov kernel function is the optimal choice for both $K_1$ and $K_2$ under some regularity conditions, where the Epanechnikov kernel function is defined as

$$K_e(u) = \begin{cases} \frac{3}{4}(1-u^2), & \text{if } |u| \leq 1, \\ 0, & \text{if } |u| > 1. \end{cases} \tag{7}$$

To choose the bandwidths $h_t$ and $h_s$, one commonly used method is the leave-one-out cross-validation (CV) procedure. However, when the observed data are correlated, it has been well demonstrated in the literature that the conventional kernel estimators using the regular bandwidth selection procedures, such as the CV procedure, would not generally perform well, because they cannot separate the data correlation structure from the data mean function effectively in such cases (e.g., Altman 1990, Opsomer et al. 2001). To overcome this difficulty, when choosing a bandwidth by CV in the univariate kernel regression setup, Brabanter et al. (2011) suggested using the so-called $\epsilon$-optimal bimodal kernel function defined as

$$\widetilde{K}_\epsilon(u) = \frac{4}{4-3\epsilon-\epsilon^2}\begin{cases} \frac{3}{4}(1-u^2)I(|u| \leq 1), & \text{if } |u| \geq \epsilon; \\ \frac{3(1-\epsilon^2)}{4\epsilon}|u|, & \text{if } |u| < \epsilon, \end{cases} \tag{8}$$

where $\epsilon \in (0, 1)$ is a constant.

In the current problem for estimating $\lambda(t_i, \mathbf{s}_{ij})$, let $\widehat{\lambda}_{-(ij)}(t, \mathbf{s})$ be the estimate of $\lambda(t, \mathbf{s})$ obtained from all observations except the one at $(t_i, \mathbf{s}_{ij})$. Then, the CV score is defined as

$$\text{CV}(h_t, h_s) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{m_i}\sum_{j=1}^{m_i}\left[\widehat{\lambda}_{-(ij)}(t_i, \mathbf{s}_{ij}) - y(t_i, \mathbf{s}_{ij})\right]^2\right\}. \tag{9}$$

The bandwidths selected by the CV procedure are the ones that minimize the CV score $\mathrm{CV}(h_t, h_s)$ defined in (9). By Theorem 3 in Section 3, if $K_1$ and $K_2$ are chosen such that $K_1(0)K_2(0) = 0$ when computing the CV score, then

$$E(\mathrm{CV}(h_t, h_s)) = MASE(h_t, h_s) + \frac{1}{N}\sum_{i=1}^{N} \sigma^2(\boldsymbol{w}_i) + o\left(\frac{1}{N|\boldsymbol{H}|}\right)$$

under some regularity conditions, where $\{\boldsymbol{w}_i, i = 1, 2, \cdots, N\}$ denote the $N$ observation points. Also, from the expression (A.12) in Appendix B, $MASE(h_t, h_s) \sim \frac{1}{N|\boldsymbol{H}|}$. By combining these two results, the CV procedure should work well in choosing the bandwidths (i.e., the minimizers of $\mathrm{CV}(h_t, h_s)$ should be close to the minimizers of $MASE(h_t, h_s)$) when $K_1(0)K_2(0) = 0$ and other regularity conditions are valid. To satisfy the condition that $K_1(0)K_2(0) = 0$, we can choose either $K_1$ to be the $\epsilon$-optimal bimodal kernel function defined in (8), or $K_2$ to be the $\epsilon$-optimal kernel function, or both $K_1$ and $K_2$ to be the $\epsilon$-optimal kernel functions. In the numerical studies presented in Section 4, these three options will be compared in various different scenarios.

## 2.4 Some other technical issues

As discussed in Section 1, most population-based databases report disease incidence rates at the city, county, state, or country level, and the specific locations of individual disease occurrences are not available. In such cases, it is difficult to use the kernel estimation procedure (4) because $\{\mathbf{s}_{ij}\}$ are not well defined. In the literature, one commonly used way to handle this issue is to use major cities or geometric "centroids" to represent the geographic locations of regions, and approximate the distance between two geographical regions by the Euclidean distance between their geographical centroids (e.g., Berke 2004, Christakos and Lai 1997, Cressie 1993, Oliver et al. 1998). However, some authors have pointed out that this distance measurement can introduce substantial errors and produce inaccurate inferences (e.g., Koshizuka and Kurita 1991, Rodriguez-Bachiller 1983). To overcome this difficulty, we propose the following strategy. Without loss of generality, we assume that a database provides monthly disease incidence numbers at the county level. Then for each county, the response in model (2) can be calculated by dividing the case number in the county by the county population and county area. In order to use the estimation procedure (4), we need to define distance between any pair of two counties. To this end, we consider using a grid covering all counties in question. Then, for a pair of two counties A and B, we can calculate the Euclidean distances for all pairs of grid points in A and grid points in B. The average of the pairwise grid point distance is defined as the distance between A and B, denoted as $\widetilde{d}(A, B)$. When the grid is finer and finer, it is easy to check that $\widetilde{d}(A, B)$ converges to

$$d(A,B) = \frac{1}{S_A S_B} \int_{\mathbf{s}^{(1)} \in A} \int_{\mathbf{s}^{(2)} \in B} d_E(\mathbf{s}^{(1)}, \mathbf{s}^{(2)}) \; d\mathbf{s}^{(1)} d\mathbf{s}^{(2)},$$

where $S_A$ and $S_B$ are the areas of $A$ and $B$. So, $\widetilde{d}(A, B)$ will not change much once the grid intensity reaches a certain level. In the above distance definition, if we know the population in each small square of the grid, then we can also use population-weighted distance (e.g., Goovaerts 2006). But, such population information may not be available in practice, and this weighting scheme should not have a substantial impact on the estimator $\widehat{\lambda}(t, \mathbf{s})$ either.

The response $y(t_i, \mathbf{s}_{ij})$ in (2) is the non-negative empirical incidence rate. Its distribution is usually skewed to the right. For rare diseases with small incidence rates, the skewness could be large. In such cases, it is natural to use $y(t_i, \mathbf{s}_{ij})$ in log scale. The resulting estimated model could be more efficient, but its explanation would be more challenging. The model estimation procedure proposed in the current paper is nonparametric. So, it can handle both versions of the model.

In the model estimation procedure (4), the bandwidth $h_s$ is chosen to be the same in the entire spatial space $\Omega$. From Figure 1, it can be seen that some counties in California have much larger areas than others. If a constant bandwith $h_s$ is used in (4), then the neighborhoods of certain counties will contain too few observations if $h_s$ is chosen relatively small, and too many observations if $h_s$ is chosen relatively large. Of course, variable bandwidth selection procedures can be considered here (Loader 1999). But, they are quite complicated to compute. In this paper, we consider the following strategy, representing a compromise between the constant and variable bandwidth selections. At a given $(t, \mathbf{s})$, if the number of observations in the neighborhood $O(\mathbf{s})$ with the bandwidth $h_s$ is 2 or less for at least one observation time point in $[t - h_t, t + h_t]$, then increase $h_s$ at $(t, \mathbf{s})$ to $\rho h_s$, where $\rho \geq 1$ is a parameter. The reason why we use 2 as the threshold for the number of observations in $O(\mathbf{s})$ is because a local spatial plane is estimated in (4) and it needs at least 3 different locations to estimate such a spatial plane properly. It should be pointed out that this bandwidth modification procedure should be used before the CV procedure (9) is applied. The parameter $\rho$ can be chosen by the CV procedure, together with $h_t$ and $h_s$. But, it will take much extra computing time. Based on extensive simulation and real-data studies, we found that $\rho = 1.5$ often gave satisfactory results. So, we recommend using this value in practice.

# 3 Statistical Properties

In this section, we discuss some statistical properties of the estimation procedure described in the previous section. Without loss of generality, assume that the ST data are obtained in the time interval $[0, T] = [0, 1]$ and spatial region $\Omega = [0, 1]^2$, and the observation times and locations are regularly spaced as follows:

$$\left\{ \left( \frac{k}{n}, \frac{j_1}{m_1}, \frac{j_2}{m_2} \right) : k = 1, 2, ..., n; j_1 = 1, 2, ..., m_1; j_2 = 1, 2, ..., m_2 \right\}.$$

In cases when the observation times and locations are unequally spaced, results presented in this section are still valid, as long as the distributions of the observation times and the observation locations have supports $[0, 1]$ and $[0, 1]^2$, respectively. First, we give some assumptions and notations that will be used in presenting major theoretical results.

For the incidence rate $\lambda(\boldsymbol{w})$ and the variance function $\sigma^2(\boldsymbol{w})$, we make the assumption:

(AS.1) The incidence rate $\lambda(\boldsymbol{w})$ is twice continuously differentiable and the variance function $\sigma^2(\boldsymbol{w})$ is continuous in the set $\Gamma = [0, T] \times \Omega$. Since $\Gamma$ is a compact set in $R^3$, $\sigma^2(\boldsymbol{w})$ is uniformly continuous in $\Gamma$.

The following three assumptions are imposed on the kernel functions:

(AS.2.a) For $\boldsymbol{u} = (u_1, u_2, u_3) \in R^3$, $K(\boldsymbol{u}) = K_1(u_1)K_2(\sqrt{u_2^2 + u_3^2})$ is bounded and symmetric about $\mathbf{0}$, $\int_{R^3} K(\boldsymbol{u})d\boldsymbol{u} = 1$, and there exist $d_1 > 0$ and $d_2 > 0$ such that (i) $K_1(u_1) > 0$ if $|u_1| < d_1$ and $K_1(u_1) = 0$ otherwise, and (ii) $K_2(\sqrt{u_2^2 + u_3^2}) > 0$ if $\sqrt{u_2^2 + u_3^2} < d_2$ and $K_2(\sqrt{u_2^2 + u_3^2}) = 0$ otherwise. Define $D_K = \{\boldsymbol{u} \in R^3 : |u_1| \leq d_1 \text{ and } \sqrt{u_2^2 + u_3^2} \leq d_2\}$, $D'_K = \{\boldsymbol{u} \in R^3 : |u_1| \leq 2d_1 \text{ and } \sqrt{u_2^2 + u_3^2} \leq 2d_2\}$, and $D''_K = \{\boldsymbol{u} \in R^3 : |u_1| \leq 3d_1 \text{ and } \sqrt{u_2^2 + u_3^2} \leq 3d_2\}$.

(AS.2.b) For $i = 1, 2, 3$, $\mu_{2,i}(K) = \int u_i^2 K(\boldsymbol{u})d\boldsymbol{u} > 0$. Let $\boldsymbol{\mu}_2(K) = diag\{\mu_{2,1}(K), \mu_{2,2}(K), \mu_{2,3}(K)\}$. Then, we have $\mu_{2,2}(K) = \mu_{2,3}(K)$. In addition, let $\mu(K^2) = \int K^2(\boldsymbol{u})d\boldsymbol{u}$.

(AS.2.c) The function $K(\boldsymbol{u})$ satisfies the Lipschitz-1 continuity condition. Namely, there exists $L \geq 0$ such that $|K(\boldsymbol{u}) - K(\boldsymbol{v})| \leq L\|\boldsymbol{u} - \boldsymbol{v}\|_1$, for any $\boldsymbol{u}, \boldsymbol{v} \in R^3$, where $\| \cdot \|_1$ is the $L_1$-norm in $R^3$.

For a given point $\boldsymbol{w} = (t, \mathbf{s}) \in \Gamma = [0, 1] \times [0, 1]^2$, let $\xi_{\boldsymbol{w},H} = \{\boldsymbol{u} \in R^3 : \boldsymbol{H}^{-1}(\boldsymbol{u} - \boldsymbol{w}) \in D_K\}$, $\xi'_{\boldsymbol{w},H} = \{\boldsymbol{u} \in R^3 : \boldsymbol{H}^{-1}(\boldsymbol{u} - \boldsymbol{w}) \in D'_K\}$, and $\xi''_{\boldsymbol{w},H} = \{\boldsymbol{u} \in R^3 : \boldsymbol{H}^{-1}(\boldsymbol{u} - \boldsymbol{w}) \in D''_K\}$. Then, $\xi_{\boldsymbol{w},H}, \xi'_{\boldsymbol{w},H}$, and $\xi''_{\boldsymbol{w},H}$ are neighborhoods of $\boldsymbol{w}$ with different bandwidths. The point $\boldsymbol{w}$

is an *interior* point if $\xi''_{\boldsymbol{w},H} \subseteq \Gamma$. With these notations, we impose the following assumptions on the random errors:

(AS.3.a) In the neighborhood $\xi''_{\boldsymbol{w},H}$ of a given point $\boldsymbol{w}$, the ST data correlation is homogeneous. Namely, for any $\mathbf{u}, \mathbf{v} \in \xi''_{\boldsymbol{w},H}$, the correlation function has the property that $Corr(\varepsilon(\mathbf{u}), \varepsilon(\mathbf{v})) = \boldsymbol{\rho}_{\boldsymbol{w}}(\mathbf{u}-\mathbf{v})$, where $\boldsymbol{\rho}_{\boldsymbol{w}} : R^3 \to [-1,1]$ is a correlation function.

(AS.3.b) At any $\boldsymbol{w} \in \Gamma$, $C''_{N,\boldsymbol{w}} = \sum_{(\frac{k^*}{nh_t}, \frac{j_1^*}{m_1h_s}, \frac{j_2^*}{m_2h_s}) \in D''_K} |\boldsymbol{\rho}_{\boldsymbol{w}}(\frac{k^*}{n}, \frac{j_1^*}{m_1}, \frac{j_2^*}{m_2})| = o(N|\boldsymbol{H}|)$.

(AS.3.b)' At any $\boldsymbol{w} \in \Gamma$, $C''_{N,\boldsymbol{w}} = O(1)$, $D''_{N,\boldsymbol{w}} = \sum_{(\frac{k^*}{nh_t}, \frac{j_1^*}{m_1h_s}, \frac{j_2^*}{m_2h_s}) \in D''_K} (\frac{|k^*|}{nh_t} + \frac{|j_1^*|}{m_1h_s} + \frac{|j_2^*|}{m_2h_s}) |\boldsymbol{\rho}_{\boldsymbol{w}}(\frac{k^*}{n}, \frac{j_1^*}{m_1}, \frac{j_2^*}{m_2})| = o(1)$ and $C''_{N,\boldsymbol{w}} - C_{N,\boldsymbol{w}} = o(1)$, where $C_{N,\boldsymbol{w}} = \sum_{(\frac{k^*}{nh_t}, \frac{j_1^*}{m_1h_s}, \frac{j_2^*}{m_2h_s}) \in D_K} |\boldsymbol{\rho}_{\boldsymbol{w}}(\frac{k^*}{n}, \frac{j_1^*}{m_1}, \frac{j_2^*}{m_2})|$. Let $Q_{N,\boldsymbol{w}} = \sum_{(\frac{k^*}{nh_t}, \frac{j_1^*}{m_1h_s}, \frac{j_2^*}{m_2h_s}) \in D_K} \boldsymbol{\rho}_{\boldsymbol{w}}(\frac{k^*}{n}, \frac{j_1^*}{m_1}, \frac{j_2^*}{m_2})$. Then, we have $Q_{N,\boldsymbol{w}} = O(1)$ since $|Q_{N,\boldsymbol{w}}| \leq C''_{N,\boldsymbol{w}}$. We further assume that $\lim_{N\to\infty} Q_{N,\boldsymbol{w}} = Q(\boldsymbol{w})$ exists.

For the bandwidths, we need the following assumptions:

(AS.4) $h_t = o(1)$, $h_s = o(1)$, $\frac{1}{nh_t} = o(1)$, $\frac{1}{m_1h_s} = o(1)$, and $\frac{1}{m_2h_s} = o(1)$.

The assumption (AS.1) requires that both $\lambda(\boldsymbol{w})$ and $\sigma^2(\boldsymbol{w})$ are smooth, which should be reasonable in most applications. The assumptions (AS.2.a)-(AS.2.c) on the two kernel functions are conventional. Basically, (AS.2.a) requires that both $K_1$ and $K_2$ are non-negative and have compact supports, (AS.2.b) requires that they are non-zero, and (AS.2.c) requires that they are Lipschitz-1 continuous. In (AS.3.a), we assume that the ST data correlation is homogeneous locally, which is much more flexible than the conventional assumption in the literature that the ST data correlation is homogeneous globally (e.g., Choi et al. 2013). In (AS.3.b), we assume that $C''_{N,\boldsymbol{w}} = o(N|\boldsymbol{H}|)$, and in (AS.3.b)' we assume that $C''_{N,\boldsymbol{w}} = O(1)$. From the definition of $C''_{N,\boldsymbol{w}}$, we can see that it is a summation of all absolute values of the correlations $\{\boldsymbol{\rho}_{\boldsymbol{w}}(\frac{k^*}{n}, \frac{j_1^*}{m_1}, \frac{j_2^*}{m_2})\}$ when $(\frac{k^*}{nh_t}, \frac{j_1^*}{m_1h_s}, \frac{j_2^*}{m_2h_s}) \in D''_K$. In the literature, the spatial or temporal correlations are often assumed to be exponentially decaying when two time points or two locations move away (e.g., in cases with $AR(1)$ time series model). In such cases, both assumptions mentioned above would be valid. The assumptions in (AS.4) on the two bandwidths are commonly used in kernel regression. Basically, they require that the neighborhood $\xi_{\boldsymbol{w},H}$ at a given point $\boldsymbol{w}$ has smaller and smaller size, but the number of observations in it gets larger and larger, when the sample size $N$ increases. Under the assumptions stated above, we have several theorems given below about our proposed method for analyzing ST data.

**Theorem 1** *Under the assumptions (AS.1), (AS.2.a)-(AS.2.c), (AS.3.a), (AS.3.b) and (AS.4), we have*

$$\lim_{n,m_1,m_2\to\infty} a_N^2 E\left(\widehat{\lambda}(t,\mathbf{s})-\lambda(t,\mathbf{s})\right)^2 = 0, \qquad \textit{for any } (t,\mathbf{s})\in(0,1)\times(0,1)^2, \tag{10}$$

*where* $a_N = \min\{\frac{1}{h_t^{2-\delta}}, \frac{1}{h_s^{2-\delta}}, (\frac{N|\boldsymbol{H}|}{C''_{N,\boldsymbol{w}}})^{0.5-\delta}\}$ *and* $0<\delta<0.5$.

Theorem 1 establishes the $L_2$ strong consistency for the estimator $\widehat{\lambda}(t,\mathbf{s})$. The two kernel functions used in (4) should both be chosen the Epanechnikov kernel defined in (7), which is confirmed in Theorem 2 below.

**Theorem 2** *Under the assumptions (AS.1), (AS.2.a)-(AS.2.c), (AS.3.a), (AS.3.b)' and (AS.4), the MASE criterion $MASE(h_t,h_s)$ defined in Section 2.3 reaches the minimum when both $K_1$ and $K_2$ are chosen to be the Epanechnikov kernel defined in (7).*

Finally, regarding the CV score defined in (9), we have the following result.

**Theorem 3** *Under the assumptions (AS.1), (AS.2.a)-(AS.2.c), (AS.3.a), (AS.3.b)' and (AS.4), if the two kernel functions $K_1$ and $K_2$ are chosen such that $K_1(0)K_2(0)=0$, then* *we have*

$$E[CV(h_t, h_s)] = MASE(h_t, h_s) + \frac{1}{N}\sum_{i=1}^{N}\sigma^2(\boldsymbol{w}_i) + o\left(\frac{1}{N|\boldsymbol{H}|}\right), \tag{11}$$

where $\{\boldsymbol{w}_i, i = 1, 2, \cdots, N\}$ denote the $N$ points in $\Gamma$ at which observations of $\lambda(t, \mathbf{s})$ are obtained.

# 4 Simulation Study

In this section, we present some simulation results about the numerical performance of our proposed method described in the previous sections. For simplicity, the number of spatial locations at each time point is chosen to be the same. The following four types of kernel functions are considered in the LLSTK local smoothing procedure (4) as discussed in Subsection 2.3:

$$K^{(1)}(\boldsymbol{u}) = \widetilde{K}_{0.1}(u_1)\widetilde{K}_{0.1}(\sqrt{u_2^2+u_3^2}), \qquad K^{(2)}(\boldsymbol{u}) = \widetilde{K}_{0.1}(u_1)K_e(\sqrt{u_2^2+u_3^2}),$$
$$K^{(3)}(\boldsymbol{u}) = K_e(u_1)\widetilde{K}_{0.1}(\sqrt{u_2^2+u_3^2}), \qquad K^{(4)}(\boldsymbol{u}) = K_e(u_1)K_e(\sqrt{u_2^2+u_3^2})$$

The simulation is organized in four parts. In the first part, the proposed method and selection of its kernel functions are studied in different cases when the correlation level of the noise, the number of spatial locations at a given time point, and the number of observation times all change. In the second part, the number of observation times and the number of spatial locations at a given time point are both fixed, but the noise is generated in a different way than the one in the first part. By this example, we would like to check whether the overall conclusions about kernel selection depend on how the random noise is generated. In the third part, the number of observation times, the correlation level of noise and the number of spatial locations are all fixed. But, the spatial locations are generated in a different way than the one in the first part. This part aims to check whether the overall conclusions about kernel selection depend on how the spatial locations are distributed. Finally, in the fourth part, we compare the proposed method with several representative existing methods.

## 4.1 Performance of the proposed method

In the first part of the simulation study, we consider the following setup. In model (2), assume that $\{t_i, i = 1, 2, \ldots, n\}$ are equally spaced in $[0, 1]$. At each observation time $t_i$, the observation locations $\{\mathbf{s}_{ij}, j = 1, 2, \ldots, m\}$ are assumed to be equally spaced in $\Omega = [0, 1] \times [0, 1]$. The true incidence rate function is chosen to be

$$\lambda(t, \mathbf{s}) = 0.5 + 0.3 \sin\left(\frac{\pi}{2} + \pi s_x\right) \sin\left(\frac{\pi}{2} + \pi s_y\right) + 0.15 \cos\left(\frac{3\pi}{2} + 2\pi t\right),$$

where $\mathbf{s} = (s_x, s_y)'$. For the ST noise $\varepsilon(t_i, \mathbf{s}_{ij})$, we first use the R package *neuRosim* and *deSolve* to generate the spatially correlated noise $\{\widetilde{\varepsilon}(\mathbf{s}_j), i = 1, 2, ..., m\}$, and then use the AR(1) model to generated the temporally correlated noise $\{\widetilde{\varepsilon}(t_i), j = 1, 2, ..., n\}$. Then, the ST noise is defined as $\{\varepsilon(t_i, \mathbf{s}_j) = \widetilde{\varepsilon}(t_i)\widetilde{\varepsilon}(\mathbf{s}_j), i = 1, 2, ..., n, j = 1, 2, ..., m\}$. The following 36 cases are considered in the simulation: $n = 50$ or 100, $m = 100$ or 225, the parameter for controling the temporal correlation level $\rho_t = 0.5, 0.3,$ or 0.05, and the parameter for controling the spatial correlation level $\rho_{\mathbf{s}} = 0.5, 0.3,$ or 0.05. In each case, the simulation is repeated 100 times, and the MASE value of the estimator $\widehat{\lambda}(t, \mathbf{s})$ can be approximated by the sample mean of the 100 Average Squared Error (ASE) values.

The two bandwidths $h_t$ and $h_s$ can be chosen by minimizing the MASE. The resulting bandwidths are denoted as $h_{t,opt}$ and $h_{s,opt}$, respectively. In practice, because $\lambda(t, \mathbf{s})$ is unknown, MASE cannot be calculated. So, the optimal values of the bandwidths cannot be calculated either. Instead, we can choose the bandwidths using the CV procedure (9).

In Table 1, the number in the first line of each entry is the minimum MASE value when $(n, m) = (50, 100)$ and when the optimal bandwidths $h_{t,opt}$ and $h_{s,opt}$ and the kernel functions specified by $K^{(1)}$, $K^{(2)}$, $K^{(3)}$, or $K^{(4)}$ are used in model estimation. The second line gives the optimal bandwidths $(h_{t,opt}, h_{s,opt})$. The third line gives the MASE value when the bandwidths are chosen by CV, and the corresponding bandwidths are given in the fourth line. The fifth line gives the MASE value when the bandwidths are chosen by CV (i.e., those in the fourth line) but the kernel functions used in estimating $\lambda(t, \mathbf{s})$ are those specified by $K^{(4)}$ (i.e., the regular Epanechnikov kernel functions). The corresponding results when $(n, m) = (50, 225)$, $(100, 100)$, and $(100, 225)$ are presented in Tables S.1-S.3 of the supplementary file, respectively. From these tables, we can make the following conclusions. i) If both the spatial and temporal correlations are relatively strong (e.g., $(\rho_t, \rho_{\mathbf{s}}) = (0.5, 0.5)$), then the kernel functions specified by $K^{(1)}$ should be used in the CV procedure, because the selected bandwidths are closer to the optimal bandwidths in most such cases, compared to the results with other kernel functions, and the related MASE value is often better as well. ii) In cases when the temporal correlation is relatively weak (e.g., $(\rho_t, \rho_{\mathbf{s}}) = (0.05, 0.5)$), the kernel functions in $K^{(1)}$ and $K^{(2)}$ would be good choices. iii) In cases when the spatial correlation is relatively weak (e.g., $(\rho_t, \rho_{\mathbf{s}}) = (0.5, 0.05)$), the kernel functions in $K^{(1)}$ and $K^{(3)}$ would be good choices. iv) If both the spatial and the temporal correlations are relatively weak, then the ones in $K^{(4)}$ could be a good choice. v) After the bandwiths are chosen properly by CV, it seems that it is always a good idea to use the conventional Epanechnikov kernel functions in estimating $\lambda(t, \mathbf{s})$. vi) When $m$ or $n$ increases, the MASE results will generally be better. *Based on these results, we suggest choosing the bandwidths by CV using the kernel functions specified in $K^{(1)}$ when we think that the ST correlation is an issue. Otherwise, the conventional Epanechnikov kernel functions can be used in bandwidth selection by CV. After the bandwidths are chosen, the Epanechnikov kernel functions can be used for* *estimating* $\lambda(t, \mathbf{s})$.

To further demonstrate the simulation results, the estimation errors, defined as differences between the fitted values and the true incidence rates, when the optimal bandwidths and the bandwidths chosen by CV with kernels in $K^{(1)}$, $K^{(2)}$, $K^{(3)}$ and $K^{(4)}$ are shown in the five columns of Figure 2, respectively, in cases when $t = 0.25$ (1st row), 0.5 (2nd row) and 0.75 (3rd row). In this figure, we choose $(n, m) = (100, 100)$, $(\rho_t, \rho_{\mathbf{s}}) = (0.3, 0.3)$, and the simulation with the median ASE value among 100 replicated simulations when the bandwidths chosen by CV with the kernels in $K^{(1)}$ are used. From the figure, it can be seen that i) the estimation errors are small when the optimal bandwidths are used, as expected, ii) the results with the bandwidths chosen by CV using the kernels in $K^{(1)}$ are close to those with the optimal bandwidths, and iii) the estimation errors are relatively large when the other three sets of bandwidths are used.

(put Table 1 about here)

(put Figure 2 about here)

## 4.2 Different ways of ST random error generalization

In the previous part, for simplicity, the ST random error is generated by a product of a sequence of temporally correlated random noise and a set of spatially correlated random noise. In that sense, the ST data correlation structure is simplified in that example. As a matter of fact, generalization of general ST correlated random numbers is a challenging research topic itself (Welvaert et al. 2011). In this part, we generate the 3-D random errors $\{\varepsilon(t_i, \mathbf{s}_{ij})\}$ directly by the *neuRosim* and *deSolve* packages in cases when $n = 100$ and $m = 100$. By this approach to generate ST correlated random errors, the ST data correlation is controlled by a single parameter $\rho_{t,\mathbf{s}}$, which is fixed at 0.05, 0.3 and 0.5 to simulate cases with different correlation. The results in the same setup as that of Table 1 are presented in Table S.4 in the supplementary file. From the table, it can be seen that the overall conclusions made earlier from Tables 1 and S.1-S.3 are still valid in the current example.

## 4.3 Cases with random design points

In the previous examples, the design points $\{(t_i, \boldsymbol{s}_{ij}), i = 1, 2, \ldots, n, j = 1, 2, \ldots, m\}$ are deterministic and regularly spaced in $\Gamma = [0, 1] \times [0, 1]^2$. In this part, we consider cases when the spatial locations are randomly distributed in $[0, 1]^2$, to investigate whether different types of design points would change the performance of $\widehat{\lambda}(t, \mathbf{s})$ and the strategies of kernel and bandwidth selection. To this end, let us consider cases when $n = 100$, $m = 100$, $\rho_t = 0.3$ and $\rho_{\mathbf{s}} = 0.3$. Unlike the setup in Subsection 4.1, the observation locations $\{\mathbf{s}_j, j = 1, 2, \ldots, 100\}$ are generated in two steps as follows. First, 10 $x$-axis values $\{\mathbf{s}_{x,j_1}, j_1 = 1, 2, ..., 10\}$ are generated from the distribution Unif$[0, 1]$, and 10 $y$-axis values $\{\mathbf{s}_{y,j_2}, j_2 = 1, 2, ..., 10\}$ are generated from the same distribution. Then, the spatial locations are defined as $\{(\mathbf{s}_{x,j_1}, \mathbf{s}_{y,j_2}), j_1 = 1, 2, ..., 10, j_2 = 1, 2, ..., 10\}$. The results in the same setup as that of Table 1 are presented in the first part of Table S.5 in the supplementary file with the label “Random Design”. For convenience of comparison, the results in Table S.2 when $(\rho_t, \rho_{\mathbf{s}}) = (0.3, 0.3)$ are included in the second part of Table S.5 with the label “Fixed Design”. It can be seen that the two sets of results have similar patterns, and we should choose the bandwiths using $K^{(1)}$ in both scenarios. We tried other cases with random design, and the findings are similar.

### 4.4 Comparison with some existing methods

In this part, we compare our proposed method, denoted as LLSTK, with three representative ST methods: DSTM and LGCP that are discussed in Section 1, and the weighted average smoothing (WAS) method by Kafadar (1996). For the proposed LLSTK method, we use $K^{(1)}$ in CV bandwidth selection, and $K^{(4)}$ in estimating $\lambda(t, \mathbf{s})$ after the bandwidths are selected. For the three existing methods, because the corresponding CV procedures are still unavailable, their parameters are chosen to minimize the MASE criterion. Therefore, this comparison is in the advantage of the three existing methods. We consider cases when $n = 100$, $m = 100$, and $\rho_t$ and $\rho_{\mathbf{s}}$ change among 0.5, 0.3, and 0.05. The calculated MASE values based on 100 replicated simulations are presented in Table 2. From the table, we can see that: i) LLSTK outperforms all three existing methods in all cases with quite large margins, and ii) all methods perform the best when the ST data correlation is the weakest (i.e., the case when $(\rho_t, \rho_{\mathbf{s}}) = (0.05, 0.05)$). Figure S.1 in the supplementary file presents the estimation errors using of the four methods at $t$ =0.25 (1st row), 0.5 (2nd row) and 0.75 (3rd row), when we choose $(n, m) = (100, 100)$, $(\rho_t, \rho_{\mathbf{s}}) = (0.3, 0.3)$, and the simulation that LLSTK has the median ASE value among 100 replicated simulations when the bandwidths chosen by CV with the kernels in $K^{(1)}$ are used. From the plots, we can see that LLSTK indeed has much smaller estimation errors, compared to the other three methods.

(put Table 2 about here)

## 5 Application to the Lung Cancer Dataset

In this section, we demonstrate the proposed nonparametric ST modeling approach LLSTK using the lung cancer dataset that is briefly described in Section 1 (cf., Figure 1) and can be downloaded from the Surveillance, Epidemiology, and End Results (SEER) system (http://seer.cancer.gov/). The dataset contains monthly incidence rates of the lung cancer in 58 counties of California during 2000–2011. Besides the proposed method LLSTK, we also consider the three existing methods LGCP, DSTM and WAS that are discussed in Subsection 4.4. Because the true ST function $\lambda(t, \mathbf{s})$ is unknown in real-data applications, the performance metric MASE is not well defined. Instead, we consider using the root predictive error mean square (RPEMS), defined as

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

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

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

where $\widehat{\lambda}_{-(ij)}(t_i, \mathbf{s}_{ij})$ is the leave-one-out estimator of $\lambda(t_i, \mathbf{s}_{ij})$ that is considered in the CV score in (9). Because the computation of the method LGCP is heavy, $\widehat{\lambda}_{-(ij)}(t_i, \mathbf{s}_{ij})$ is replaced by $\widehat{\lambda}(t_i, \mathbf{s}_{ij})$ in both RPEMS and MAPE for this method, which is in its advantage because the quantities $\{|\widehat{\lambda}_{-(ij)}(t_i, \mathbf{s}_{ij}) - y(t_i, \mathbf{s}_{ij})|\}$ are usually larger than $\{|\widehat{\lambda}(t_i, \mathbf{s}_{ij}) - y(t_i, \mathbf{s}_{ij})|\}$. For the proposed method LLSTK, its bandwidths are selected by the CV procedure (9) with the kernels specified in $K^{(1)}$, and the kernels specified in $K^{(4)}$ are used when estimating $\lambda(t, \mathbf{s})$. For the other three methods, their parameters are chosen to minimize the RPEMS. The calculated values of RPEMS for the methods LGCP, DSTM, WAS and LLSTK are 7.30, 5.56, 4.92 and 4.75, respectively, and their MAPE values are 3.68, 3.70, 2.80, and 2.67. It can be seen that the proposed method LLSTK has the smallest values of RPEMS and MAPE among all four methods.

To further investigate the performance of the four methods, their residual maps for the data in March 2007, October 2010 and December 2011 are shown in columns 2-5 of Figure

3, and the observed data in these three months are shown in the first column. For each county, the residual at a given month is defined to be the difference between the observed and estimated disease incidence rates. The residual maps at other months look similar. From the residual maps, it can be seen that the first two competing methods generate relatively large residuals at some counties, and the residuals of the proposed method and WAS are relatively small. Meanwhile, we choose four counties to show the observed temporal data in Figure S.2 in the supplementary file, along with the estimated functions of $\lambda(t, \mathbf{s})$ by the four competing methods. From the plots in that figure, it can be seen that the proposed method provides a good estimate of $\lambda(t, \mathbf{s})$ in terms of both bias and variance of the estimate. As a comparison, estimates by the three competing methods have either quite large bias or quite large variance.

(put Figure 3 about here)

## 6 Concluding Remarks

We have presented a nonparametric modeling approach for analyzing ST disease incidence data. This approach does not require restrictive assumptions on the observation distribution, ST pattern of the disease incidence rate, and ST correlation in the observed data. It has been shown by both theoretical arguments and numerical studies that it is effective in practice. However, there are still several issues that need to be addressed in the future research. For instance, there could be covariates that have a substantial impact on the disease incidence rate, which have not been accommodated by the current method yet. Also, the estimated model needs to be evaluated more rigorously about its goodness-of-fit. More graphical tools are needed in this regard. We will continue to work on these and some other related problems and make the proposed LLSTK method more effective and powerful for analyzing ST data.

# References

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

Berke, O. (2004), “Exploratory disease mapping: kriging the spatial risk function from regional count data,” *International Journal of Health Geographics*, **3**, 1–18.

Bloom, D.E., Cafiero, E.T., Jané-Llopis, E., Abrahams-Gessel, S., Bloom, L.R., Fathima, S., Feigl, A.B., Gaziano, T., Mowafi, M., Pandya, A., Prettner, K., Rosenberg, L., Seligman, B., Stein, A.Z., and Weinstein, C. (2011), *The Global Economic Burden of Noncommunicable Diseases*. Geneva: World Economic Forum.

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

Chandra, T.K., and Ghosal, S. (1996), “The strong law of large numbers for weighted averages under dependence assumptions,” *Journal of Theoretical Probability*, **9**, 797–809.

Chen, K., and Jin, Z. (2005), “Local polynomial regression analysis of clustered data,” *Biometrika*, **92**, 59–74.

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

Christakos, G., Lai, J.J. (1997), "A study of the breast cancer dynamics in North Carolina," *Social Science and Medicine*, **45**, 1503–1517.

Cressie, N. (1993), *Statistics for Spatial Data,* New York: John Wiley & Sons.

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

Diggle, P.J., Moraga, P., Rowlingson, B., and 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.

Finley, A.O., Banerjee, S., and Gamerman, D. (2015), "spBayes for large univariate and multivariate point-referenced spatio-temporal data models," *Journal of Statistical Software*, **63(13)**, 1–28.

Goovaerts, P. (2006), "Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging," *International Journal of Health Geographics*, **5**, 1–31.

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

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

Koshizuka, T., and Kurita, O. (1991), "Approximate formulas of average distances associated with regions and their applications to location problems," *Mathematical Programming*, **52**, 99–123.

Last, J.M., ed. (2001), *A Dictionary of Epidemiology (4 ed.)*, New York, NY: Oxford University Press.

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

Loader, C.R. (1999), "Bandwidth selection: classical or plug-in?" *The Annals of Statistics*, **27**, 415–438.

McLeish, D.L. (1975), "A maximal inequality and dependent strong laws," *The Annals of Probability*, **3**, 829–839.

Oliver, M.A., Webster, R., Lajaunie, C., Muir, K.R., Parkes, S.E., Cameron, A.H., Stevens, M.C., and Mann, J.R. (1998), "Binomial cokriging for estimating and mapping the risk of childhood cancer," *IMA Journal of Mathematics Applied in Medicine and Biology*, **15**, 279–297.

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

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

Rodriguez-Bachiller, A. (1983), "Errors in the measurement of spatial distances between discrete regions," *Environment and Planning A*, **15**, 781–799.

Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., and Rosseel, Y. (2011), "**neuRosim**: An R package for generating fMRI data," *Journal of Statistical Software*, **44**, 1–18.

Xiang, D., Qiu, P., and Pu, X. (2013), "Local polynomial regression analysis of multivariate longitudinal data," *Statistica Sinica*, **23**, 769–789.

Table 1: In each entry, line 1 gives the MASE value and its standard error (in parenthesis) when $(h_{t,opt}, h_{s,opt})$ are used. Line 2 gives $(h_{t,opt}, h_{s,opt})$. Line 3 gives the MASE value and its standard error (in parenthesis) when the bandwidths are chosen by CV. Line 4 gives the bandwidths by CV. Line 5 gives the MASE value and its standard error (in parenthesis) when the bandwidths are chosen by CV and the conventional Epanechnikov kernel functions are used in estimating $\lambda(t, \mathbf{s})$. Cases when $(n, m) = (50, 100)$.

| $(\rho_t, \rho_\mathbf{s})$ | $K^{(1)}$ | $K^{(2)}$ | $K^{(3)}$ | $K^{(4)}$ |
|---|---|---|---|---|
| (0.5,0.5) | $1.10 \times 10^{-3}(6.14e-5)$ | $1.04 \times 10^{-3}(6.21e-5)$ | $1.03 \times 10^{-3}(5.70e-5)$ | $9.76 \times 10^{-4}(5.75e-5)$ |
| | (0.16,0.32) | (0.16,0.32) | (0.16,0.32) | (0.16,0.32) |
| | $2.05 \times 10^{-3}(1.12e-4)$ | $3.15 \times 10^{-3}(1.61e-4)$ | $2.58 \times 10^{-3}(1.31e-4)$ | $3.55 \times 10^{-3}(1.72e-4)$ |
| | (0.05,0.21) | (0.05,0.11) | (0.03,0.15) | (0.03,0.11) |
| | $1.85 \times 10^{-3}(1.05e-4)$ | $2.81 \times 10^{-3}(1.44e-4)$ | $2.71 \times 10^{-3}(1.43e-4)$ | $3.55 \times 10^{-3}(1.72e-4)$ |
| (0.5,0.3) | $8.38 \times 10^{-4}(3.99e-5)$ | $7.71 \times 10^{-4}(4.04e-5)$ | $7.87 \times 10^{-4}(3.76e-5)$ | $6.88 \times 10^{-4}(3.31e-5)$ |
| | (0.15,0.24) | (0.15,0.24) | (0.15,0.24) | (0.14,0.32) |
| | $1.15 \times 10^{-3}(5.40e-5)$ | $2.68 \times 10^{-3}(1.26e-4)$ | $2.13 \times 10^{-3}(9.71e-5)$ | $3.03 \times 10^{-3}(1.35e-4)$ |
| | (0.07,0.23) | (0.05,0.11) | (0.03,0.15) | (0.03,0.11) |
| | $9.90 \times 10^{-4}(5.39e-5)$ | $2.40 \times 10^{-3}(1.14e-4)$ | $2.03 \times 10^{-3}(1.01e-4)$ | $3.03 \times 10^{-3}(1.35e-4)$ |
| (0.3,0.5) | $8.74 \times 10^{-4}(4.60e-5)$ | $8.32 \times 10^{-4}(4.62e-5)$ | $7.89 \times 10^{-4}(4.23e-5)$ | $7.49 \times 10^{-4}(4.22e-5)$ |
| | (0.15,0.26) | (0.15,0.28) | (0.15,0.25) | (0.15,0.27) |
| | $9.60 \times 10^{-4}(5.28e-5)$ | $1.88 \times 10^{-3}(9.78e-5)$ | $2.15 \times 10^{-3}(1.07e-4)$ | $2.44 \times 10^{-3}(1.23e-4)$ |
| | (0.12,0.23) | (0.07,0.12) | (0.03,0.15) | (0.03,0.14) |
| | $8.17 \times 10^{-4}(4.94e-5)$ | $1.55 \times 10^{-3}(8.58e-5)$ | $2.26 \times 10^{-3}(1.16e-4)$ | $2.44 \times 10^{-3}(1.23e-4)$ |
| (0.5,0.05) | $6.09 \times 10^{-4}(1.96e-5)$ | $5.03 \times 10^{-4}(1.98e-5)$ | $5.83 \times 10^{-4}(1.99e-5)$ | $4.76 \times 10^{-4}(2.12e-5)$ |
| | (0.13,0.33) | (0.12,0.32) | (0.13,0.32) | (0.13,0.26) |
| | $6.13 \times 10^{-4}(2.09e-5)$ | $2.20 \times 10^{-3}(9.70e-5)$ | $5.91 \times 10^{-4}(2.15e-5)$ | $2.49 \times 10^{-3}(1.04e-4)$ |
| | (0.13,0.31) | (0.05,0.11) | (0.13,0.30) | (0.03,0.11) |
| | $4.77 \times 10^{-4}(1.91e-5)$ | $1.97 \times 10^{-3}(8.80e-5)$ | $4.77 \times 10^{-4}(1.97e-5)$ | $2.49 \times 10^{-3}(1.04e-4)$ |
| (0.05,0.5) | $6.80 \times 10^{-4}(3.15e-5)$ | $6.44 \times 10^{-4}(3.14e-5)$ | $5.82 \times 10^{-4}(2.89e-5)$ | $5.46 \times 10^{-4}(2.92e-5)$ |
| | (0.15,0.24) | (0.15,0.26) | (0.14,0.24) | (0.14,0.25) |
| | $6.80 \times 10^{-4}(3.15e-5)$ | $6.47 \times 10^{-4}(3.27e-5)$ | $1.66 \times 10^{-3}(8.05e-5)$ | $1.45 \times 10^{-3}(7.57e-5)$ |
| | (0.15,0.24) | (0.15,0.24) | (0.03,0.15) | (0.03,0.22) |
| | $5.48 \times 10^{-4}(2.91e-5)$ | $5.48 \times 10^{-4}(2.91e-5)$ | $1.73 \times 10^{-3}(8.77e-5)$ | $1.45 \times 10^{-3}(7.57e-5)$ |
| (0.3,0.3) | $6.62 \times 10^{-4}(2.74e-5)$ | $5.98 \times 10^{-4}(2.73e-5)$ | $6.00 \times 10^{-4}(2.58e-5)$ | $5.38 \times 10^{-4}(2.60e-5)$ |
| | (0.14,0.26) | (0.14,0.27) | (0.14,0.25) | (0.14,0.25) |
| | $6.92 \times 10^{-4}(3.04e-5)$ | $2.02 \times 10^{-3}(1.03e-4)$ | $1.78 \times 10^{-3}(7.87e-5)$ | $2.13 \times 10^{-3}(9.75e-5)$ |
| | (0.12,0.24) | (0.05,0.12) | (0.03,0.15) | (0.03,0.12) |
| | $5.57 \times 10^{-4}(2.86e-5)$ | $1.56 \times 10^{-3}(7.66e-5)$ | $1.70 \times 10^{-3}(8.20e-5)$ | $2.13 \times 10^{-3}(9.75e-5)$ |
| (0.3,0.05) | $5.13 \times 10^{-4}(1.53e-5)$ | $4.12 \times 10^{-4}(1.60e-5)$ | $4.77 \times 10^{-4}(1.47e-5)$ | $3.73 \times 10^{-4}(1.51e-5)$ |
| | (0.13,0.30) | (0.13,0.25) | (0.13,0.30) | (0.13,0.25) |
| | $5.13 \times 10^{-4}(1.53e-5)$ | $1.96 \times 10^{-3}(9.30e-5)$ | $4.78 \times 10^{-4}(1.70e-5)$ | $1.51 \times 10^{-3}(6.70e-5)$ |
| | (0.13,0.30) | (0.05,0.11) | (0.13,0.25) | (0.05,0.11) |
| | $3.85 \times 10^{-4}(1.36e-5)$ | $1.51 \times 10^{-3}(6.70e-5)$ | $3.73 \times 10^{-4}(1.51e-5)$ | $1.51 \times 10^{-3}(6.70e-5)$ |
| (0.05,0.3) | $5.28 \times 10^{-4}(1.96e-5)$ | $4.73 \times 10^{-4}(2.00e-5)$ | $4.52 \times 10^{-4}(1.79e-5)$ | $3.98 \times 10^{-4}(1.83e-5)$ |
| | (0.14,0.24) | (0.13,0.25) | (0.13,0.24) | (0.13,0.24) |
| | $5.28 \times 10^{-4}(1.96e-5)$ | $4.75 \times 10^{-4}(2.02e-5)$ | $1.37 \times 10^{-3}(5.86e-5)$ | $1.30 \times 10^{-3}(6.14e-5)$ |
| | (0.14,0.24) | (0.14,0.24) | (0.03,0.15) | (0.03,0.15) |
| | $4.01 \times 10^{-4}(1.79e-5)$ | $4.01 \times 10^{-4}(1.79e-5)$ | $1.30 \times 10^{-3}(6.14e-5)$ | $1.30 \times 10^{-3}(6.14e-5)$ |
| (0.05,0.05) | $4.23 \times 10^{-4}(1.19e-5)$ | $3.35 \times 10^{-4}(1.16e-5)$ | $3.71 \times 10^{-4}(1.17e-5)$ | $2.83 \times 10^{-4}(1.10e-5)$ |
| | (0.13,0.25) | (0.13,0.24) | (0.13,0.24) | (0.12,0.24) |
| | $4.23 \times 10^{-4}(1.19e-5)$ | $3.37 \times 10^{-4}(1.24e-5)$ | $3.71 \times 10^{-4}(1.23e-5)$ | $3.01 \times 10^{-4}(1.37e-5)$ |
| | (0.13,0.25) | (0.13,0.23) | (0.12,0.24) | (0.09,0.23) |
| | $2.91 \times 10^{-4}(9.99e-6)$ | $2.86 \times 10^{-4}(1.12e-5)$ | $2.83 \times 10^{-4}(1.10e-5)$ | $3.01 \times 10^{-4}(1.37e-5)$ |

Table 2: MASE values of the four ST modeling approaches in cases when $n = 100$, $m = 100$, and $\rho_t$ and $\rho_{\mathbf{s}}$ change among 0.5, 0.3, and 0.05.

| $(\rho_t, \rho_{\mathbf{s}})$ | DSTM | LGCP | WAS | LLSTK |
|---|---|---|---|---|
| (0.5,0.5) | $5.95 \times 10^{-3}(1.19e-4)$ | $7.07 \times 10^{-3}(2.13e-4)$ | $4.85 \times 10^{-3}(1.86e-4)$ | $1.73 \times 10^{-3}(8.48e-5)$ |
| (0.5,0.3) | $5.26 \times 10^{-3}(1.55e-4)$ | $6.90 \times 10^{-3}(1.77e-4)$ | $3.84 \times 10^{-3}(1.25e-4)$ | $9.24 \times 10^{-4}(4.39e-5)$ |
| (0.5,0.05) | $4.71 \times 10^{-3}(1.31e-4)$ | $6.69 \times 10^{-3}(1.57e-4)$ | $1.12 \times 10^{-3}(3.80e-5)$ | $3.33 \times 10^{-4}(1.43e-5)$ |
| (0.3,0.5) | $5.34 \times 10^{-3}(1.75e-4)$ | $7.11 \times 10^{-3}(2.07e-4)$ | $4.87 \times 10^{-3}(1.81e-4)$ | $6.85 \times 10^{-4}(3.81e-5)$ |
| (0.3,0.3) | $4.54 \times 10^{-3}(1.22e-4)$ | $6.91 \times 10^{-3}(1.65e-4)$ | $3.85 \times 10^{-3}(1.18e-4)$ | $3.74 \times 10^{-4}(1.88e-5)$ |
| (0.3,0.05) | $3.91 \times 10^{-3}(9.33e-5)$ | $6.72 \times 10^{-3}(1.45e-4)$ | $1.12 \times 10^{-3}(3.68e-5)$ | $2.55 \times 10^{-4}(9.93e-6)$ |
| (0.05,0.5) | $4.75 \times 10^{-3}(1.58e-4)$ | $7.15 \times 10^{-3}(2.07e-4)$ | $4.89 \times 10^{-3}(1.82e-4)$ | $3.41 \times 10^{-4}(2.06e-5)$ |
| (0.05,0.3) | $3.86 \times 10^{-3}(9.82e-5)$ | $6.95 \times 10^{-3}(1.62e-4)$ | $3.86 \times 10^{-3}(1.16e-4)$ | $2.55 \times 10^{-4}(1.26e-5)$ |
| (0.05,0.05) | $3.18 \times 10^{-3}(6.69e-5)$ | $6.75 \times 10^{-3}(1.41e-4)$ | $1.13 \times 10^{-3}(3.64e-5)$ | $1.90 \times 10^{-4}(6.78e-6)$ |

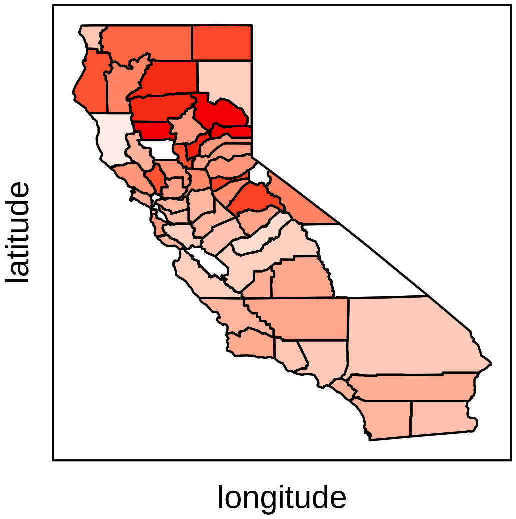

Figure 1: Lung cancer incidence rates in 58 counties of California in Feburary, 2006. The redder the color, the higher the incidence rate.

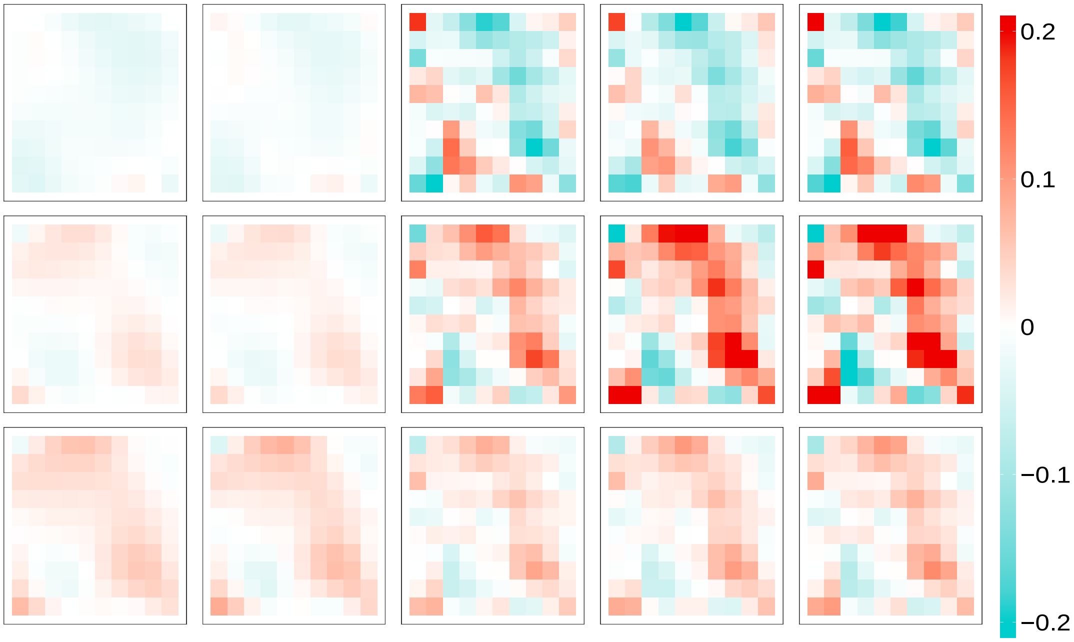

Figure 2: Estimation errors of $\widehat{\lambda}(t, \mathbf{s})$ when using the optimal bandwidth (1st column), the bandwidths found by CV with kernels in $K^{(1)}$ (2nd column), $K^{(2)}$ (3rd column), $K^{(3)}$ (4th column) and $K^{(4)}$ (5th column), at $t$ =0.25 (1st row), 0.5 (2nd row) and 0.75 (3rd row). In the plots, $(n, m) = (100, 100)$, $(\rho_t, \rho_{\mathbf{s}}) = (0.3, 0.3)$, and we choose the simulation with the median ASE value among 100 replicated simulations when the bandwidths chosen by CV using the kernels in $K^{(1)}$ are used.

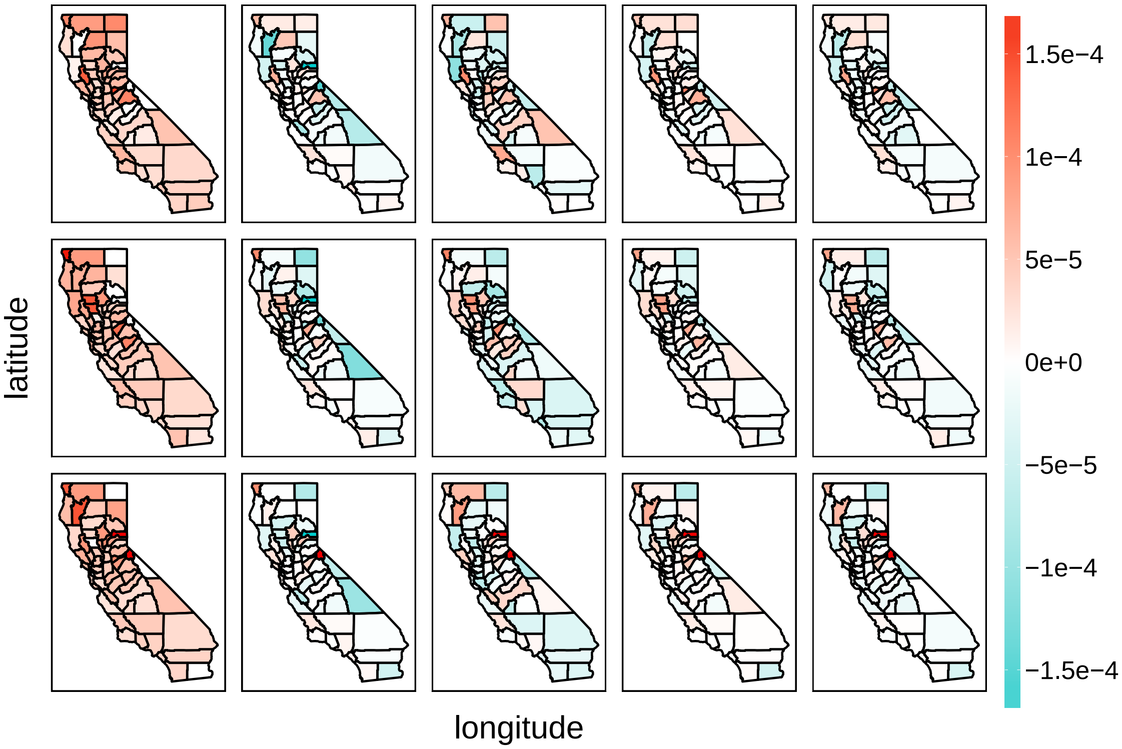

Figure 3: Observed lung cancer data (1st column) in 58 counties of California in March 2007 (1st row), October 2010 (2nd row) and December 2011 (3rd column), and the residual plots of the method LGCP (2nd column), DSTM (3rd column), WAS (4th column) and LLSTK (5th column).