Multiscale Spatio-Temporal Data Monitoring Using Forward Elimination and Scale Information Criterion

Kai Yang[1] and Peihua Qiu[2]

[1]Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI 53226

[2]Department of Biostatistics, University of Florida, Gainesville, FL 32610

Abstract

Spatio-temporal process monitoring has broad applications in disease surveillance, environment monitoring, epidemiology, and many other areas. Most existing spatio-temporal process monitoring methods are developed for analyzing observed data at a single geospatial scale (e.g., state, county, or census tract), and usually impose some parametric assumptions on spatio-temporal data structure. For instance, observed data are often assumed to follow a parametric distribution (e.g., normal or Poisson) and the mean and covariance of a related spatio-temporal model are assumed to have parametric forms, which are rarely valid and/or difficult to justify in many applications. In practice, spatio-temporal data (e.g., incidence rates of infectious diseases) are usually managed at multiple geospatial scales and have complicated structure. In this paper, we develop a new method for multiscale spatio-temporal data monitoring. At each observation time, the new method first detects anomalous regions at each spatial scale. To reduce computing time, a forward elimination procedure is suggested for this purpose. Then, an optimal spatial scale at the current observation time is determined using the detected anomalous regions and a proposed scale information criterion. Finally, a charting statistic is constructed based on the detected anomalous regions at the optimal scale. This method is fully nonparametric in the sense that no parametric assumptions are imposed on the structure of observed spatio-temporal data. It is also scalable and computationally efficient, making it a reliable and effective tool for online multiscale spatio-temporal data monitoring problems.

Keywords: Forward elimination; Geographical scales; Information criterion; Kernel smoothing; Spatio-temporal monitoring; Statistical process control.

1 Introduction

Spatio-temporal data analysis has attracted significant attention in recent years, as such data are increasingly common across various fields, including agriculture, climate science, epidemiology, geography, and medicine (Schoenberg 2003; Shand and Li 2017). In many applications, spatio-temporal data are collected and aggregated over predefined spatial areas or regions and can be organized across multiple geographical scales (e.g., states, counties, and census tracts). In this paper, we focus on such datasets, often referred to as multiscale spatio-temporal areal data (Ferreira et al. 2011; Fonseca and Ferreira 2017; Elkhouly and Ferreira 2021). Our goal is to develop an effective method for detecting changes in the space-time pattern of a multiscale spatio-temporal process.

Our method is motivated by the need to detect infectious disease outbreaks using spatio-temporal surveillance data across multiple geospatial scales. Infectious diseases remain a significant threat to public health. Early detection of disease outbreaks is crucial for minimizing their negative impacts, motivating the development of national and regional surveillance systems that collect disease incidence data over time. In these systems, incidence data are typically reported sequentially from a vast network of medical facilities, aggregated over disjoint administrative regions, and converted between different spatial scales (e.g., from the county level to the state level), resulting in multiscale spatio-temporal data. Consequently, statistical decisions regarding changes in disease incidence patterns must be made prospectively. This is the prospective multiscale disease surveillance problem addressed in this paper. To leverage the multiscale structure of such data, this paper introduces a new monitoring method that identifies an optimal spatial scale for surveillance and detects changes in the spatio-temporal pattern of disease incidence in a timely manner.

To analyze multiscale spatial or spatio-temporal data, several methods have been proposed in the literature, with a focus on regression modeling. These include hierarchical models based on wavelet decomposition (Berliner et al. 1999), approaches for normally distributed or Poisson data based on likelihood decomposition (Kolaczyk and Huang 2001; Louie and Kolaczyk 2006), and dynamic multiscale models using the change-of-resolution Kalman filter and certain state-space equations (Johannesson et al. 2007; Ferreira et al. 2011; Fonseca and Ferreira 2017; Elkhouly and

Ferreira 2021). These methods are developed to model multiscale spatial or spatio-temporal data when all data are collected prior to data analysis. Thus, they are retrospective and not directly applicable to the prospective online data monitoring problem, where new data are continuously collected and sequential decisions about process status are made over time.

For online monitoring of spatio-temporal processes, existing methods focus on analyzing data collected at a single spatial scale. For instance, Knox and Barlett (1964) proposed the Knox statistic to identify irregular space-time interactions, assuming that the total number of event pairs close in space and time follows a Poisson distribution. Kulldorff and Hjalmars (1999) proposed a generalized version to account for population shift bias. These Knox-type statistics are designed for point process data where precise event locations are available and are thus unsuitable for areal data, which provide only aggregated area-level information. Another class of methods is based on scan statistics, defined as a maximum likelihood ratio test statistic over a collection of scanning windows (e.g., Kulldorff 1997; Takahashi et al. 2008; Woodall et al. 2008). These methods require specifying a parametric distribution for observed data (e.g., Poisson, binomial, or normal), an assumption that may be violated in practice. There have been some other methods developed in the framework of statistical process control (SPC) (cf., Qiu 2014). For example, Yan et al. (2018) proposed an online monitoring method based on tensor decomposition, which decomposes observed spatio-temporal data into functional mean, anomaly, and random error components, assuming that random errors are independent and normally distributed. To accommodate complex spatio-temporal data structure, Yang and Qiu (2020) suggested a cumulative sum chart for spatio-temporal process monitoring, and Qiu and Yang (2021) further extended this approach to incorporate covariate information. In both methods, all spatial regions are weighted equally, allowing anomalies to consist of isolated regions scattered across the spatial domain of interest. In practice, however, spatio-temporal anomalies often emerge within small, spatially clustered regions rather than appearing as isolated regions distributed throughout the entire domain. To leverage this spatial feature, Qiu and Yang (2023) proposed an exponentially weighted spatial LASSO procedure for monitoring spatio-temporal processes. To summarize, these existing methods provide useful tools for online monitoring of spatio-temporal processes observed at a single scale. Applying them to multiscale data requires selecting a specific

spatial scale in advance. However, choosing an appropriate scale is often challenging, and reliance on a fixed scale can limit the effectiveness of anomaly detection, as the full multiscale structure of the data may not be adequately captured.

Following the above discussion, there are no methods in the literature that are specifically designed for prospective online monitoring of spatio-temporal data with multiple spatial scales. This paper aims to fill the gap by developing a new multiscale spatio-temporal data monitoring method that is completely nonparametric, computationally efficient, capable of handling large data, and effective for detecting anomalies clustered in small regions. The proposed method first estimates a nonparametric spatio-temporal model to establish a baseline pattern for comparison, using local kernel smoothing based on in-control (IC) data (Fan and Gijbels 1996). Newly observed multiscale spatio-temporal data are then compared with the estimated baseline pattern for sequential anomaly detection. Specifically, at each observation time, anomalous regions are identified at each spatial scale. For applications involving nested spatial scales (e.g., state, county, and census tract levels), a forward elimination procedure is introduced to reduce computational cost by exploiting the hierarchical aggregation structure of multiscale spatio-temporal data. An optimal spatial scale is subsequently selected using a proposed scale information criterion (SIC), and a charting statistic is constructed for online monitoring. With minor modifications, the proposed method can also accommodate non-nested spatial scales, making it applicable to a broad range of spatio-temporal monitoring applications. Theoretical justification and numerical studies demonstrate its effectiveness across a wide range of practical settings.

The remainder of the paper is organized as follows. In Section 2, the proposed method is described in detail, along with some of its theoretical properties. In Section 3, its numerical performance is evaluated using Monte Carlo simulations under several different scenarios. In Section 4, the proposed method is demonstrated using an influenza-like illness dataset collected in Florida. Some concluding remarks are given in Section 5. Estimation of IC spatio-temporal mean and covariance, practical guidelines on implementation of the method, some extra numerical results and proofs of theoretical results are provided in a supplementary file.

2 Proposed Methodology

Our proposed method is described in three parts of this section. Multiscale spatio-temporal data modeling for estimating the baseline pattern of a multiscale spatio-temporal process under monitoring is discussed in Subsection 2.1. Online process monitoring based on the estimated baseline multiscale spatio-temporal pattern is discussed in Subsection 2.2. Some theoretical results of the proposed method are presented in Subsection 2.3.

2.1 Estimation of the baseline multiscale spatio-temporal pattern

Assume that the spatio-temporal process under monitoring is defined in the spatial domain $\Omega \subseteq \mathbb{R}^2$, and a total of L spatial scales are considered. At the lth spatial scale, for $l = 1, \ldots, L$, Ω consists of m_l disjoint regions $\{B_{l1}, \ldots, B_{lm_l}\}$. Namely,

$$\Omega = \cup_{j=1}^{m_l} B_{lj}, \;\; B_{lj_1} \cap B_{lj_2} = \emptyset \;\text{ if } 1 \le j_1 \ne j_2 \le m_l.$$

In cases when the L spatial scales are nested, any region at the $(l+1)$th scale would be a subset of a region at the lth scale, and

$$B_{lj} = \bigcup_{k \in \text{ch}(l,j)} B_{l+1,k}, \quad \text{for } l = 1, \ldots, L-1,$$

where $B_{l+1,k}$ is called a child region of B_{lj}, for each $k \in \text{ch}(l,j)$, and $\text{ch}(l,j)$ is the index set of all child regions of B_{lj} at scale $l+1$. Such nested structure of scales is common. For instance, in the United States, census tracts are nested within counties which in return are nested within states.

To provide intuition for the proposed framework, we illustrate our notation with a simple toy example. Let $\Omega = [0,1] \times [0,1]$ and consider $L = 2$ spatial scales, with $m_1 = 4$ regions at scale 1 and $m_2 = 16$ regions at scale 2. Figure 1 provides a graphical illustration of the corresponding region partitions and their nested structure. As shown, regions B_{21}–B_{24} are nested within B_{11} and are therefore considered child regions of B_{11} (i.e., $\text{ch}(1,1) = \{1,2,3,4\}$).

Assume that, prior to online process monitoring, a dataset has been collected over the baseline time interval $[0,T]$, during which the process has been confirmed to be free of anomalies (e.g.,

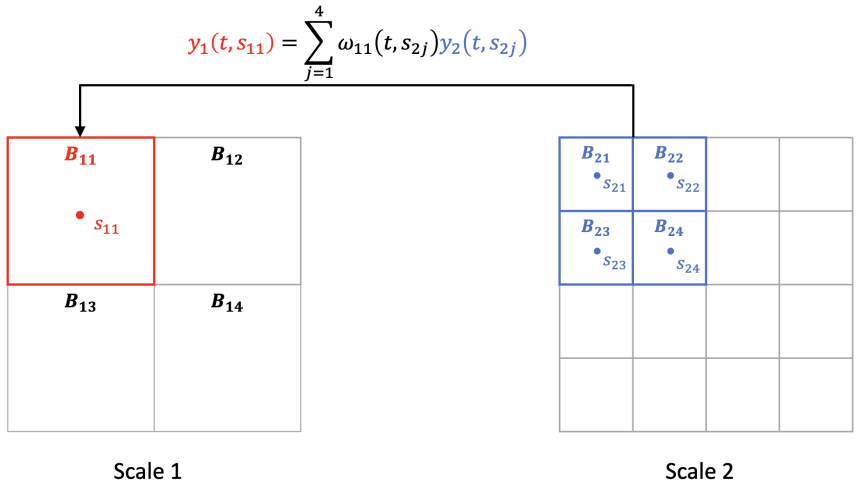


Figure 1: Illustration of nested spatial regions and data aggregation across scales.

no disease outbreaks). This dataset is often called an IC dataset in the SPC literature (e.g., Qiu 2014). Let $\{y_L(t_i, \boldsymbol{s}_{Lj}) : j = 1, \ldots, m_L, i = 1, \ldots, n\}$ denote IC process observations at the finest scale (i.e., scale L), where $t_i \in [0, T]$ is the ith observation time, $\boldsymbol{s}_{Lj}$ is the centroid of region B_{Lj}, and n is the total number of observation times. Although observations at the finest scale provide detailed geographical information about the underlying spatio-temporal process, they often exhibit high data variability. Consequently, spatial aggregation is frequently employed to reduce variability and construct meaningful regions for analysis (Gotway and Young 2002). Let $\boldsymbol{s}_{lj}$ be the centroid of region B_{lj} and $y_l(t_i, \boldsymbol{s}_{lj})$ denote the observation for this region at time t_i, for all j, l, and i. To aggregate data from the finest scale and obtain data at coarser scales 1 through $L-1$, a natural idea is to use the following recursive data aggregation formula:

$$y_l(t_i, \boldsymbol{s}_{lj}) = \boldsymbol{\omega}_{lj}^{\top}(t_i) \boldsymbol{y}_{l+1}(t_i), \ j = 1, \ldots, m_l, \ l = 1, \ldots, L-1, \ i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{y}_{l+1}(t_i) = (y_{l+1}(t_i, \boldsymbol{s}_{l+1,1}), \ldots, y_{l+1}(t_i, \boldsymbol{s}_{l+1,m_{l+1}}))^{\top}$ is the vector of all observations at scale $l+1$, and $\boldsymbol{\omega}_{lj}(t_i) = (\omega_{lj}(t_i, \boldsymbol{s}_{l+1,1}), \ldots, \omega_{lj}(t_i, \boldsymbol{s}_{l+1,m_{l+1}}))^{\top}$ is a vector of pre-specified aggregation weights. The weight $\omega_{lj}(t_i, \boldsymbol{s}_{l+1,k})$ represents the contribution of data from region $B_{l+1,k}$ to the aggregated observation for region B_{lj}. By construction, $\omega_{lj}(t_i, \boldsymbol{s}_{l+1,k}) = 0$ whenever $k \notin \mathrm{ch}(l, j)$, ensuring that only child regions contribute to the aggregation. The formula (1) includes aggregation by total, (weighted) average, and decimation as special cases. For instance, in the context of disease surveillance, if $y_l(t_i, \boldsymbol{s}_{lj})$ represents the disease incidence rate for region B_{lj} at time t_i, a natural choice is $\omega_{lj}(t_i, \boldsymbol{s}_{l+1,k}) = N_{l+1,k}(t_i)/N_{lj}(t_i)$, where $N_{lj}(t_i)$ and $N_{l+1,k}(t_i)$ denote the populations of

region B_{lj} and its child region $B_{l+1,k}$ at time t_i, respectively. In Figure 1, this data aggregation is illustrated by showing how the observed data from child regions B_{21}–B_{24} (i.e., $y_2(t, \boldsymbol{s}_{21})$–$y_2(t, \boldsymbol{s}_{24})$) are combined to form the aggregated observation $y_1(t, \boldsymbol{s}_{11})$ for region B_{11} at a coarser scale.

The observed IC data at the finest scale $\{y_L(t_i, \boldsymbol{s}_{Lj}) : j = 1, \ldots, m_L, i = 1, \ldots, n\}$ are assumed to follow the nonparametric spatio-temporal model:

$$y_L(t, \boldsymbol{s}) = \mu_{L,0}(t, \boldsymbol{s}) + \epsilon_L(t, \boldsymbol{s}), \quad \text{for } (t, \boldsymbol{s}) \in [0, T] \times \Omega, \tag{2}$$

where $\epsilon_L(t, \boldsymbol{s})$ is a zero-mean random error, $\mu_{L,0}(t, \boldsymbol{s})$ is the mean of the IC data $y_L(t, \boldsymbol{s})$, and the subscript "0" denotes "IC". The spatio-temporal correlation is described by the covariance structure:

$$V_{L,0}(t, t'; \boldsymbol{s}, \boldsymbol{s}') = \mathrm{Cov}\left(y_L(t, \boldsymbol{s}), y_L(t', \boldsymbol{s}')\right), \quad \text{for } (t, \boldsymbol{s}), (t', \boldsymbol{s}') \in [0, T] \times \Omega. \tag{3}$$

The model (2)-(3) does not impose any parametric forms on $\mu_{L,0}(t, \boldsymbol{s})$, $V_{L,0}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ and data distribution. So, it is flexible to use in practice. Estimation of $\mu_{L,0}(t, \boldsymbol{s})$ and $V_{L,0}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ has been discussed in Yang and Qiu (2018, 2019). The estimates are denoted as $\widehat{\mu}_{L,0}(t, \boldsymbol{s})$ and $\widehat{V}_{L,0}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$, with their definitions given in Section A of the supplementary file.

For $l = 1, \ldots, L$, let $\mu_{l,0}(t, \boldsymbol{s}) = \mathrm{E}(y_l(t, \boldsymbol{s}))$ and $V_{l,0}(t, t'; \boldsymbol{s}, \boldsymbol{s}') = \mathrm{Cov}(y_l(t, \boldsymbol{s}), y_l(t', \boldsymbol{s}'))$ denote the IC mean and covariance at the lth scale, for $t, t' \in [0, T]$ and $\boldsymbol{s}, \boldsymbol{s}' \in \Omega$. Once $\widehat{\mu}_{L,0}(t, \boldsymbol{s})$ and $\widehat{V}_{L,0}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ are obtained, estimates of the mean and covariance at coarser scales can be derived recursively using the formula (1). Specifically, estimates at scale $L-1$ are computed based on the estimates at scale L using the spatial aggregation formula. Similarly, estimates at scale $L-2$ are derived from the estimates at scale $L-1$, and this process continues down to the coarsest scale. For example, suppose that the estimates at scale $l+1$ have been computed, denoted by $\widehat{\mu}_{l+1,0}(t, \boldsymbol{s})$ and $\widehat{V}_{l+1,0}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$. Based on these, we can obtain the estimated mean vector of $\boldsymbol{y}_{l+1}(t)$ and the estimated covariance matrix between $\boldsymbol{y}_{l+1}(t)$ and $\boldsymbol{y}_{l+1}(t')$, denoted by $\widehat{\boldsymbol{\mu}}_{l+1,0}(t)$ and $\widehat{\mathbf{V}}_{l+1,0}(t, t')$, for $t, t' \in [0, T]$. Then, from Equation (1), at scale l, for any $j, j' = 1, \ldots, m_l$, the estimated mean of $y_l(t, \boldsymbol{s}_{lj})$ and the estimated covariance between $y_l(t, \boldsymbol{s}_{lj})$ and $y_l(t', \boldsymbol{s}_{lj'})$ are given by

$$\begin{cases} \widehat{\mu}_{l,0}(t, \boldsymbol{s}_{lj}) = \widehat{\mathrm{E}}\left(\boldsymbol{\omega}_{lj}^{\top}(t)\boldsymbol{y}_{l+1}(t)\right) = \boldsymbol{\omega}_{lj}^{\top}(t)\widehat{\boldsymbol{\mu}}_{l+1,0}(t), \\ \widehat{V}_{l,0}(t, t'; \boldsymbol{s}_{lj}, \boldsymbol{s}_{lj'}) = \widehat{\mathrm{Cov}}\left(\boldsymbol{\omega}_{lj}^{\top}(t)\boldsymbol{y}_{l+1}(t), \boldsymbol{\omega}_{lj'}^{\top}(t')\boldsymbol{y}_{l+1}(t')\right) = \boldsymbol{\omega}_{lj}^{\top}(t)\widehat{\mathbf{V}}_{l+1,0}(t, t')\boldsymbol{\omega}_{lj'}(t'). \end{cases} \tag{4}$$

It should be pointed out that observations $y_l(t_i, \boldsymbol{s}_{lj})$ could be available at each spatial scale but the aggregation formula (1) may be unknown in some applications. In such cases, instead of using (4), estimates of $\mu_{l,0}(t, \boldsymbol{s})$ and $V_{l,0}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$, for $l = 1, \ldots, L-1$, can be computed at each spatial scale similarly to the procedures given in Section A of the supplementary file. In addition, if observations are missing for a small number of spatial regions in $\{B_{Lj} : j = 1 \ldots, m_L\}$, or the number of spatial regions changes over time, the model estimation procedure described above can still be used. This is because the model estimation mainly depends on accurate estimation of the mean and covariance at scale L, assuming the formula (1) is known. The estimators at scale L are computed using a local kernel smoothing procedure, which involves a weighted average of nearby data and does not require all data points to be available (cf., Yang and Qiu, 2018 and 2019).

2.2 Online spatio-temporal process monitoring

Once the baseline model is estimated from IC data, the regular multiscale spatio-temporal pattern of the process under monitoring can be described by the estimated IC mean and covariance $\widehat{\mu}_{l,0}(t, \boldsymbol{s})$ and $\widehat{V}_{l,0}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$, for $t, t' \in [0, T]$, $\boldsymbol{s}, \boldsymbol{s}' \in \Omega$, and $l = 1, \ldots, L$. In practice, spatio-temporal data often have repeating seasonal patterns. For instance, incidence rate of flu is usually high in the winter and low in the summer. To accommodate such seasonality, it is assumed that the baseline interval $[0, T]$ covers one entire period (e.g., one full year). Future observations without anomalies (i.e., IC) are then expected to roughly follow the same multiscale spatio-temporal pattern, extended periodically in time from $[0, T]$ to $[0, \infty)$ with period T, so that $\mu_{l,0}(t + kT, \boldsymbol{s}) = \mu_{l,0}(t, \boldsymbol{s})$, for any $t \in [0, T]$ and any non-negative integer k, where the subscript "0" denotes "IC" or "baseline". This periodic extension defines the expected IC behavior at all future time points when no anomalies are present. For instance, if a full seasonal cycle in disease incidence is one year, which is the case for many infectious diseases, then $T = 1$ (year) and $\mu_{l,0}(t, \boldsymbol{s})$ repeats across years to reflect the regular annual pattern. Let $\{y_l(t_i, \boldsymbol{s}_{lj}) : l = 1, \ldots, L, j = 1, \ldots, m_l, i \geq n+1\}$ be the observations for online process monitoring. Then, they are assumed to follow the model

$$y_l(t, \boldsymbol{s}) = \mu_{l,0}(t, \boldsymbol{s}) + \delta_l(t, \boldsymbol{s}) + \epsilon_l(t, \boldsymbol{s}), \quad \text{for } l = 1, \ldots, L, \ (t, \boldsymbol{s}) \in (T, \infty) \times \Omega, \tag{5}$$

where $\epsilon_l(t, \boldsymbol{s})$ is a zero-mean random error, and $\delta_l(t, \boldsymbol{s}) = \mathrm{E}(y_l(t, \boldsymbol{s})) - \mu_{l,0}(t, \boldsymbol{s})$ is a deterministic mean shift function. Here, $\delta_l(t, \boldsymbol{s})$ is identifiable because $\mu_{l,0}(t, \boldsymbol{s})$ is fully specified from the IC dataset and extends periodically across seasonal cycles during online monitoring. In Model (5), $\delta_l(t, \boldsymbol{s}) = 0$ when the process under monitoring is IC and becomes non-zero after an anomaly occurs. In SPC terminology, the process is said to be out-of-control (OC) in the latter case.

In practice, anomalies in spatio-temporal processes often start in small spatially clustered regions (Wang et al. 2018). For instance, outbreaks of infectious diseases like flu and COVID-19 usually start in some small communities and then spread to a larger population. In such applications, a natural idea is to first detect such regions with possible anomalies and then focus on the detected regions for online process monitoring. However, it could be challenging to detect anomalous regions accurately at a fine spatial scale, due to the following two main reasons: (i) observed data often display high variability at a fine spatial scale, and (ii) the number of spatial regions under investigation could be large in such cases. In the literature, it has been well studied that selection of an appropriate scale is extremely important when analyzing multiscale spatio-temporal data (cf., Gotway and Young 2002). However, it has not been discussed in the SPC literature on selection of an appropriate spatial scale for online monitoring of a multiscale spatio-temporal process.

To address these challenges, we propose a forward elimination strategy to identify anomalous regions at each spatial scale. First, some notations are introduced. Let ς be the basic time unit which is the largest time unit that all observation times are its integer multiples. This quantity is introduced to handle unequally spaced observation times (Qiu and Xiang 2014). At the current observation time t_i, for $i > n$, let $\mathbf{Y}_{lj}$ be the vector with elements $\{y_l(t_{i'}, \boldsymbol{s}_{lj'}) : j' = 1, \ldots, m_l, d_E(\boldsymbol{s}_{lj}, \boldsymbol{s}_{lj'}) \leq h, n+1 \leq i' \leq i\}$, where $d_E(\boldsymbol{s}_{lj}, \boldsymbol{s}_{lj'})$ is the Euclidean distance between $\boldsymbol{s}_{lj}$ and $\boldsymbol{s}_{lj'}$, and $h > 0$ is a bandwidth. Obviously, $\mathbf{Y}_{lj}$ contains all observations at scale l from the beginning of process monitoring to the current time, contributed by regions whose centroids lie within a Euclidean distance h of the centroid of B_{lj} (i.e., $\boldsymbol{s}_{lj}$). When the monitored process is assumed to be IC, the observed data of the process would follow the baseline pattern described by the estimated IC mean function $\widehat{\mu}_{l,0}(t, \boldsymbol{s})$ and covariance function $\widehat{V}_{l,0}(t, t'; \boldsymbol{s}, \boldsymbol{s}')$, after the periodic extension of both functions as discussed at the beginning of Subsection 2.2. Therefore, the IC mean vector and covariance matrix

of $\mathbf{Y}_{lj}$ can be estimated by the corresponding quantities derived from $\widehat{\mu}_{l,0}(t,\boldsymbol{s})$ and $\widehat{V}_{l,0}(t,t';\boldsymbol{s},\boldsymbol{s}')$, respectively, and the estimates are denoted as $\widehat{\boldsymbol{\mu}}_{lj,0}$ and $\widehat{\mathbf{V}}_{lj,0}$.

Let $\mathrm{w}_\delta(i',j';j) = K_h\left(d_E(\boldsymbol{s}_{lj'},\boldsymbol{s}_{lj})\right)(1-\lambda)^{(t_i-t_{i'})/\varsigma}$, $\mathbf{K}_j = (\mathrm{w}_\delta(n+1,1;j),\ldots,\mathrm{w}_\delta(i,m_l;j))^\top$ be the vector of the weights, $\mathbf{D}_j = \mathrm{diag}(1/\mathbf{K}_j)$ be a diagonal matrix, and $\mathbf{W}_{lj} = \left(\sqrt{\widehat{\mathbf{V}}_{lj,0}}\mathbf{D}_j\sqrt{\widehat{\mathbf{V}}_{lj,0}}\right)^+$ be a weight matrix, for $l = 1,\ldots,L$ and $j = 1,\ldots,m_l$, where $K_h(x) = K(x/h)$, the kernel function is chosen to be the Epanechnikov kernel $K(x) = 0.75(1-x^2)\mathrm{I}(|x|\le 1)$ due to its good theoretical properties (Epanechnikov 1969), $\lambda \in (0,1]$ is a weighting parameter, and "+" is the Moore-Penrose inverse. Our forward elimination starts at the coarsest scale. At time t_i, let $\mathcal{M}_1(t_i) = \{j : 1,\ldots,m_1\}$ be the index set of all regions to consider for detecting anomalous regions at scale 1. By combining the ideas of exponentially weighted data smoothing in the time domain and local kernel smoothing in the space domain, we propose to estimate the shift sizes $\{\delta_1(t_i,\boldsymbol{s}_{1j}) : j \in \mathcal{M}_1(t_i)\}$ by the following penalized exponentially weighted kernel smoothing procedure:

$$\arg\min_{\boldsymbol{\delta}_1} \sum_{j\in\mathcal{M}_1(t_i)} \left(\mathbf{Y}_{1j} - \widehat{\boldsymbol{\mu}}_{1j,0} - \mathbf{1}\delta_{1j}\right)^\top \mathbf{W}_{1j}\left(\mathbf{Y}_{1j} - \widehat{\boldsymbol{\mu}}_{1j,0} - \mathbf{1}\delta_{1j}\right) + P\left(\boldsymbol{\delta}_1\right), \tag{6}$$

where $\mathbf{1} = (1,\ldots,1)^\top$, $\boldsymbol{\delta}_1 = (\delta_{11},\ldots,\delta_{1m_1})^\top$, and $P\left(\boldsymbol{\delta}_1\right)$ is a penalty that shrinks estimates to zero in regions without anomalies and encourages spatial clustering of nonzero estimates.

To use the penalized exponentially weighted kernel smoothing procedure (6), the penalty term $P(\boldsymbol{\delta}_1)$ should be chosen properly. There are many different penalty functions suggested in the literature on variable selection, among which the adaptive LASSO penalty (Zou 2006) and the SCAD penalty (Fan and Li 2001) are popular due to their good theoretical properties such as the oracle property (Fan and Li 2001). In this paper, we choose to use a spatial version of the adaptive LASSO penalty because it is convex and the related optimization problem can be solved efficiently. Specifically, the following adaptive spatial LASSO penalty is considered:

$$P\left(\boldsymbol{\delta}_1\right) = \phi_1 \sum_{j\in\mathcal{M}_1(t_i)} \varpi_{1j}\left|\delta_{1j}\right| + \phi_2 \sum_{j\in\mathcal{M}_1(t_i)} \varpi_{2j}\left|\delta_{1j} - \frac{\sum_{j'\in\mathcal{M}_1(t_i)} K_h\left(d_E(\boldsymbol{s}_{1j},\boldsymbol{s}_{1j'})\right)\delta_{1j'}}{\sum_{j'\in\mathcal{M}_1(t_i)} K_h\left(d_E(\boldsymbol{s}_{1j},\boldsymbol{s}_{1j'})\right)}\right|, \tag{7}$$

where $\phi_1,\phi_2 > 0$ are tuning parameters and their selection is discussed in Section B of the supplementary file, and ϖ_{1j} and ϖ_{2j} are the adaptive LASSO weights (cf., Zou 2006). From (7), the penalty consists of two terms. The first term enforces estimated shift sizes in non-anomalous regions to be 0. The second term penalizes the differences among estimated shift sizes in neighboring

regions, making detected anomalous regions to be clustered. When the adaptive spatial LASSO penalty (7) is used, the optimization problem in (6) can be re-formulated as a special case of the generalized LASSO problem discussed in Tibshirani and Taylor (2011). Thus, it can be solved using the efficient path algorithm suggested by Tibshirani and Taylor (2011).

Here, it is important to highlight that the penalized exponentially weighted kernel smoothing procedure (6) does not rely on any data decorrelation algorithms, distinguishing it from the exponentially weighted spatial LASSO proposed by Qiu and Yang (2023). In practice, this procedure would be more computationally efficient and can more accurately detect anomalous regions compared to the exponentially weighted spatial LASSO. The latter method, which incorporates data decorrelation across both space and time, may mistakenly classify some non-anomalous regions as anomalous. This misclassification occurs because the decorrelated data in non-anomalous regions are linear combinations of data from all regions, including anomalous ones. This phenomenon has been confirmed through numerical studies in Subsection 3.2.

The solution to δ_{1j} of Equation (6) is defined to be the estimate of $\delta_1(t_i, \boldsymbol{s}_{1j})$, denoted as $\widehat{\delta}_1(t_i, \boldsymbol{s}_{1j})$, for $j \in \mathcal{M}_1(t_i)$. It can be seen from (6) that $\delta_1(t_i, \boldsymbol{s}_{1j})$ is estimated by a weighted average of all available observations up to the current time and the weights for the observed data that are farther away from $(t_i, \boldsymbol{s}_{1j})$ would be smaller. Note that the dimension of $\mathbf{Y}_{1j}$ is controlled by the bandwidth h that is usually small. Thus, the computation involved in (6) is quite fast. Let $\mathcal{A}_1(t_i) = \{j : \delta_1(t_i, \boldsymbol{s}_{1j}) \neq 0, j \in \mathcal{M}_1(t_i)\}$ be the index set of the true anomalous regions at scale 1. After estimating it by $\widehat{\mathcal{A}}_1(t_i) = \{j : \widehat{\delta}_1(t_i, \boldsymbol{s}_{1j}) \neq 0, j \in \mathcal{M}_1(t_i)\}$, detected anomalous regions at scale 1 are naturally defined to be $\{B_{1j} : j \in \widehat{\mathcal{A}}_1(t_i)\}$.

To detect anomalous regions at scales 2 to L, instead of using a similar approach to the one discussed above at each scale, we suggest using the following recursive procedure to save computing time. Assume that the estimates $\{\widehat{\delta}_{l-1}(t_i, \boldsymbol{s}_{l-1,j}) : j \in \mathcal{M}_{l-1}(t_i)\}$ have been obtained in advance, and we want to detect anomalous regions at scale l, for $l = 2, \ldots, L$. To this end, if $\widehat{\delta}_{l-1}(t_i, \boldsymbol{s}_{l-1,j}) = 0$, then the region $B_{l-1,j}$ is not a detected anomalous region at scale $l-1$ and thus the chance for any of its child regions $\{B_{lk} : k \in \mathrm{ch}(l-1, j)\}$ to be anomalous is small. Based on this observation, we suggest excluding all such child regions from consideration when detecting anomalous regions at

the lth scale. The index set of regions to be considered at scale l is then defined as

$$\mathcal{M}_l(t_i) = \{k : k \in \text{ch}(l-1, j), \widehat{\delta}_{l-1}(t_i, \boldsymbol{s}_{l-1,j}) \neq 0, j \in \mathcal{M}_{l-1}(t_i)\}.$$

The shift sizes $\{\delta_l(t_i, \boldsymbol{s}_{lj}) : j \in \mathcal{M}_l(t_i)\}$ can then be estimated by (6) after all the subscripts "1" are replaced by "l" in that expression. Let $\{\widehat{\delta}_l(t_i, \boldsymbol{s}_{lj}) : j \in \mathcal{M}_l(t_i)\}$ be the estimates and $\mathcal{A}_l(t_i) = \{j : \delta_l(t_i, \boldsymbol{s}_{lj}) \neq 0, j = 1, \ldots, m_l\}$ be the index set of the true anomalous regions at scale l. A natural estimate is $\widehat{\mathcal{A}}_l(t_i) = \{j : \widehat{\delta}_l(t_i, \boldsymbol{s}_{lj}) \neq 0, j \in \mathcal{M}_l(t_i)\}$, and detected anomalous regions at the lth scale are then defined to be $\{B_{lj} : j \in \widehat{\mathcal{A}}_l(t_i)\}$.

So far, we have detected anomalous regions at all spatial scales. Another important question to answer is which spatial scale should be used when monitoring a multiscale spatio-temporal process. Intuitively, an appropriate scale for online monitoring should have the following two properties. First, a related space partition should be fine enough to distinguish areas with and without process shifts. If some regions at a spatial scale contain both regular and anomalous child regions, then a finer partition should be a better choice. Second, observed data at the selected spatial scale should show low variability, implying that a too fine scale should be avoided. Based on these considerations, we define the optimal scale for online process monitoring as

$$\ell(t_i) = \min\left\{ l : \bigcup_{j \in \mathcal{A}_l(t_i)} B_{lj} = \bigcup_{j \in \mathcal{A}_L(t_i)} B_{Lj} \right\}, \quad \text{for } i \geq n+1. \tag{8}$$

The definition in (8) allows the optimal scale $\ell(t_i)$ to vary over time. Simulation results presented in Figure S1 of the supplementary file confirm that the spatial scale defined in (8) is indeed a good choice in many numerical examples considered.

In practice, the true index sets of anomalous regions $\mathcal{A}_l(t_i)$ and $\mathcal{A}_L(t_i)$ used in (8) are often unknown. Thus, $\ell(t_i)$ needs to be estimated from observed data. A naïve estimate can be obtained by replacing $\mathcal{A}_l(t_i)$ and $\mathcal{A}_L(t_i)$ by their respective estimates $\widehat{\mathcal{A}}_l(t_i)$ and $\widehat{\mathcal{A}}_L(t_i)$ determined by the proposed forward elimination procedure. Even though the consistency of detected anomalous regions by the proposed forward elimination procedure is guaranteed under some mild conditions (cf., Theorem 2 in Subsection 2.3), it requires $\mathcal{A}_l(t_i) \approx \widehat{\mathcal{A}}_l(t_i)$, for all l, to ensure the naïve estimate to work well, which may not be realistic in a finite sample setting. To overcome this limitation, we

propose an information criterion for estimating the optimal scale, as described below. The proposed criterion consists of a term measuring the fidelity of the estimated process means to observed data at the finest scale (i.e., scale L) and a penalty term related to the number of detected anomalous regions and the scale level considered. Suppose that scale l is optimal at t_i (i.e., $\ell(t_i) = l$), and anomalous regions at the lth scale are detected precisely. In such cases, from the definition of optimal scale in (8), it can be checked easily that $\bigcup_{j \in \mathcal{A}_L(t_i)} B_{Lj} = \bigcup_{j \in \widehat{\mathcal{A}}_l(t_i)} B_{Lj}$. It follows that $\delta_L(t_i, \boldsymbol{s}_{Lj}) = 0$, for any $1 \le j \le m_L$, if $\boldsymbol{s}_{Lj} \notin B_{lk}$ for all $k \in \widehat{\mathcal{A}}_l(t_i)$. Let

$$\delta_L^{(l)}(t_i, \boldsymbol{s}_{Lj}) = \begin{cases} 0, & \text{if } \boldsymbol{s}_{Lj} \notin B_{lk} \text{ for all } k \in \widehat{\mathcal{A}}_l(t_i), \\ \widetilde{\delta}_L(t_i, \boldsymbol{s}_{Lj}), & \text{otherwise,} \end{cases}$$

where $\widetilde{\delta}_L(t_i, \boldsymbol{s}_{Lj})$ is the estimate of $\delta_L(t_i, \boldsymbol{s}_{Lj})$ that is obtained by (6) after removing the penalty to reduce bias and replacing all subscripts "1" and the index set $\mathcal{M}_1(t_i)$ by "L" and $\{j : \boldsymbol{s}_{Lj} \in B_{lk}$ for one $k \in \widehat{\mathcal{A}}_l(t_i)\}$, respectively. Then, $\sum_{j=1}^{m_L} \left(y_L(t_i, \boldsymbol{s}_{Lj}) - \widehat{\mu}_{L,0}(t_i, \boldsymbol{s}_{Lj}) - \delta_L^{(l)}(t_i, \boldsymbol{s}_{Lj}) \right)^2$ should be small if the optimal scale is l. This quantity is similar to the residual sum of squares in the Bayesian information criterion (BIC, Schwarz 1978) and will be denoted as RSS(l) hereafter.

Then, we propose estimating the optimal scale $\ell(t_i)$ by minimizing the following SIC score:

$$\text{SIC}(l) = m_L \log \left[\text{RSS}(l)/m_L \right] + \mathcal{G}(m_L) \left[\sum_{j=1}^{m_L} \text{I} \left(\delta_L^{(l)}(t_i, \boldsymbol{s}_{Lj}) \neq 0 \right) + l/L \right], \tag{9}$$

where $\mathcal{G}(m_L)$ is an adjustment factor and its selection is discussed in Subsection 2.3. The estimated optimal scale by (9) is denoted as $\widehat{\ell}(t_i)$ throughout the paper. From its construction in (9), it can be seen that SIC performs well as long as anomalous regions at the optimal scale are detected with sufficient accuracy. Both theoretical results in Subsection 2.3 and simulation studies in Section 4 confirm that SIC achieves high selection accuracy when the number of regions at the finest scale is sufficiently large, which is generally the case in practice. These findings indicate that SIC can be applied effectively in real-world settings. As a comparison, the naïve estimation method requires the anomalous regions to be detected accurately at all spatial scales. Thus, it is expected that SIC would perform better than the naïve approach, which is confirmed in Subsection 3.1.

The entire forward elimination procedure is summarized below.

(i) Start at scale 1, compute the estimates $\{\widehat{\delta}_1(t_i, \boldsymbol{s}_{1j}), j \in \mathcal{M}_1(t_i)\}$ by (6), and obtain the index set of the detected anomalous regions $\widehat{\mathcal{A}}_1(t_i) = \{j : \widehat{\delta}_1(t_i, \boldsymbol{s}_{1j}) \neq 0, j \in \mathcal{M}_1(t_i)\}$;

(ii) At scale l, for $l = 2, \ldots, L$, remove regions $\{B_{lj} : j \notin \mathcal{M}_l(t_i)\}$ from consideration, compute $\{\widehat{\delta}_l(t_i, \boldsymbol{s}_{lj}), j \in \mathcal{M}_l(t_i)\}$ by a procedure similar to (6), and obtain the resulting index set of the detected anomalous regions $\widehat{\mathcal{A}}_l(t_i) = \{j : \widehat{\delta}_l(t_i, \boldsymbol{s}_{lj}) \neq 0, j \in \mathcal{M}_l(t_i)\}$;

(iii) Determine the optimal scale by minimizing SIC(l) defined in (9).

Next, we use another toy example to further illustrate the proposed forward elimination procedure. Assume that $\Omega = [0,1] \times [0,1]$, $L = 3$, $m_1 = 4^2$, $m_2 = 8^2$, and $m_3 = 16^2$. Suppose that each region B_{lj} is square and that, at each spatial scale, the centroids of all regions are arranged in an equally spaced grid over Ω. Furthermore, we assume that process shifts occur in region $\Delta = [3/8, 5/8] \times [3/8, 5/8]$, the optimal scale is $\ell(t) = 2$, and $\widehat{\mathcal{A}}_l(t_i) = \mathcal{A}_l(t_i)$. Figure 2 presents a graphical demonstration of the proposed forward elimination procedure in this example. In the figure, the anomalous regions $\{B_{lj} : j \in \mathcal{A}_l(t_i)\}$ are highlighted in red, while the blue lines indicate the regions $\{B_{lj} : j \in \mathcal{M}_l(t_i)\}$. This figure illustrates that the anomalous regions are detected sequentially from scale 1 to scale $L = 3$ and the optimal scale is determined by SIC. Here, we want to mention that only the estimates $\{\widehat{\delta}_l(t_i, \boldsymbol{s}_{lj}) : j \in \mathcal{M}_l(t_i)\}$ need to be computed when using the forward elimination procedure. Note that all the three index sets $\{\mathcal{M}_l(t_i) : l = 1, 2, 3\}$ has 16 elements in this example, which is much smaller than the total number of regions at the finest scale (i.e., $m_3 = 16^2$). Due to this feature, our method is computationally efficient and can handle large multiscale spatio-temporal data with a small or moderate number of anomalous regions.

From the above description, our proposed method is motivated by applications in which spatial scales are hierarchically nested. The proposed forward elimination procedure, however, can still be used in cases without this structure since this structure is not required in its implementation. When scales are not nested, the data aggregation formula (1) and the recursive estimation procedure (4) cannot be used. In such cases, we suggest estimating the mean and covariance at each scale separately by using the local linear kernel smoothing procedure and the weighted moment estimation procedure as discussed in the supplementary file, and then our proposed method can still be used.

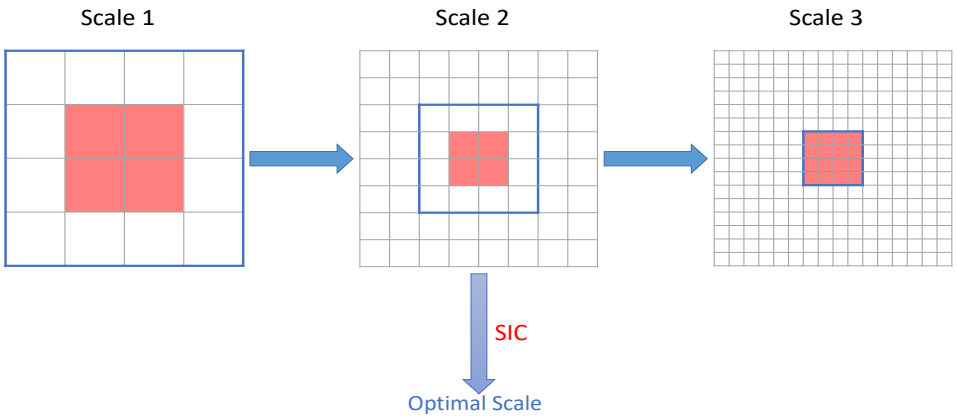


Figure 2: A graphical demonstration of the proposed forward elimination procedure.

After the optimal scale is determined at the current observation time t_i, we suggest using the following control chart to detect mean shifts in a multiscale spatio-temporal process:

$$Q(t_i) = \frac{U(t_i) - \widehat{\mathrm{E}}_0\left(U(t_i)\right)}{\sqrt{\widehat{\mathrm{Var}}_0\left(U(t_i)\right)}} \quad \text{with} \quad U(t_i) = \widehat{\boldsymbol{\delta}}^{\top}(t_i)\widehat{\boldsymbol{\Sigma}}_{\delta,0}^{-1}(t_i)\widehat{\boldsymbol{\delta}}(t_i), \quad \text{for } i \geq n+1, \tag{10}$$

where $\widehat{\boldsymbol{\delta}}(t_i)$ is the vector with elements $\{\widehat{\delta}_{\widehat{\ell}(t_i)}(t_i, \boldsymbol{s}_{\widehat{\ell}(t_i),j}) : j \in \widehat{\mathcal{A}}_{\widehat{\ell}(t_i)}(t_i)\}$, $\widehat{\boldsymbol{\Sigma}}_{\delta,0}(t_i)$ is the estimated IC covariance matrix of $\widehat{\boldsymbol{\delta}}(t_i)$, $\widehat{\mathrm{E}}_0(U(t_i))$ and $\widehat{\mathrm{Var}}_0(U(t_i))$ are the IC estimates of the mean and variance of $U(t_i)$, and all subscripts "0" denote "IC". Because the asymptotic IC distribution of $\widehat{\boldsymbol{\delta}}(t_i)$ is difficult to derive analytically under the penalized procedure in (6), we do not provide closed-form expressions for these estimates. Instead, we estimate them under IC conditions using a block bootstrap procedure. Specifically, we obtain bootstrap samples from IC data, recompute $\widehat{\boldsymbol{\delta}}(t_i)$ and $U(t_i)$ for each bootstrap sample, and then calculate the empirical sample mean and covariance of $\widehat{\boldsymbol{\delta}}(t_i)$ and $U(t_i)$ based on a large number of bootstrap replications. More details are provided in Section B of the supplementary file. The chart in (10) gives a signal if $Q(t_i) > \gamma$, where $\gamma > 0$ is a control limit. This chart is denoted as MST-FE hereafter, where "MST" implies that it is designed for monitoring "multiscale spatio-temporal" data and "FE" denotes "forward elimination".

To evaluate the performance of a control chart such as MST–FE, two standard metrics are commonly used: the IC average run length (ARL), denoted by ARL_0, and the OC ARL, denoted by ARL_1. The ARL_0 is defined as the average number of observation times from the start of process monitoring to the signal time when the process remains IC, whereas ARL_1 is defined as the average

observation times from the occurrence of a process shift to the signal time. These metrics are widely adopted in the SPC literature (Qiu 2014) because they capture not only conventional false positive and false negative rates, but also the timeliness of signal detection. The latter is particularly important in online monitoring applications, where rapid detection of changes is critical.

In practice, a nominal ARL_0 is typically pre-specified as a design parameter, and the control limit of the chart is then calibrated to achieve this target value. For the proposed MST–FE chart, selection of the control limit γ is performed using a block bootstrap procedure that resamples from an IC dataset; implementation details are provided in Section B of the supplementary material. Once γ is determined, the actual ARL_0 and ARL_1 can be evaluated via Monte Carlo simulations. Specifically, multiple IC datasets are generated from Model (2) to estimate ARL_0 and to verify the reliability of the bootstrap calibration, in the sense that the resulting ARL_0 is close to its nominal level. In addition, multiple OC datasets with a specified shift are generated from Model (5) to estimate ARL_1. For a given shift, a chart with a smaller ARL_1 while maintaining the desired ARL_0 is preferred, as it detects anomalies more rapidly.

To use the proposed method MST-FE, the parameters λ, h, ϕ_1, ϕ_2 and γ should be chosen properly. We give practical guidelines on design and implementation of MST-FE, including the selection of these parameters in Section B of the supplementary file. In the ideal case where IC data covering two full periods are available, we recommend using one period of the IC data for baseline model estimation and the remaining period for control limit determination in order to achieve a nominal ARL_0 (e.g., 200). However, when IC data are from a single period only, these data can be used for both purposes. While this is feasible, it would generally result in an ARL_0 value that could deviate more from the nominal level.

2.3 Theoretical Properties

Next, we study the theoretical properties of the proposed method. To this end, the strong mixing coefficient (cf., Ibragimov and Linnik 1971) in the time domain is defined to be

$$\alpha_\kappa = \sup_{i\geq 1} \sup_{A,B} \left\{ |\Pr(AB) - \Pr(A)\Pr(B)| : A \in \mathfrak{F}_1^i, B \in \mathfrak{F}_{i+\kappa}^\infty \right\}, \quad \text{for } \kappa \geq 1,$$

where $\mathfrak{F}_1^i$ is the σ-algebra generated by $\{\epsilon_L(t_{i'}, \boldsymbol{s}_{Lj}) : j = 1, \ldots, m_L, 1 \leq i' \leq i\}$, and $\mathfrak{F}_{i+\kappa}^{\infty}$ is the σ-algebra generated by $\{\epsilon_L(t_{i'}, \boldsymbol{s}_{Lj}) : j = 1, \ldots, m_L, i' \geq i + \kappa\}$.

Additionally, observed areal data are typically generated by aggregating an underlying spatio-temporal point process. For any $t \in [0, \infty)$ and $\boldsymbol{s} \in \Omega$, let $N^*(t, \boldsymbol{s}; dt, O_{\boldsymbol{s}})$ denote the number of events (e.g., disease cases) in a sufficiently small region $O_{\boldsymbol{s}}$ around $\boldsymbol{s}$ with area $d\boldsymbol{s}$ and within the time window $[t, t + dt]$. Define the first-order intensity function to be

$$f_1(t, \boldsymbol{s}) = \lim_{dt \to 0, d\boldsymbol{s} \to 0} \frac{\mathrm{E}(N^*(t, \boldsymbol{s}; dt, O_{\boldsymbol{s}}))}{N(t, \boldsymbol{s}; O_{\boldsymbol{s}})dt}, \quad \text{for any } (t, \boldsymbol{s}) \in [0, \infty) \times \Omega,$$

where $N(t, \boldsymbol{s}; O_{\boldsymbol{s}})$ represents the volume of $O_{\boldsymbol{s}}$ at time t and is used for aggregation of point process data to obtain areal data at a fine scale. Important special cases of $N(t, \boldsymbol{s}; O_{\boldsymbol{s}})$ include (i) $N(t, \boldsymbol{s}; O_{\boldsymbol{s}}) = d\boldsymbol{s}$, and (ii) $N(t, \boldsymbol{s}; O_{\boldsymbol{s}})$ is the population size of $O_{\boldsymbol{s}}$ at time t. When $N(t, \boldsymbol{s}; O_{\boldsymbol{s}}) = d\boldsymbol{s}$, $f_1(t, \boldsymbol{s})$ is the intensity function of a point process (cf., Li and Guan 2014). When $N(t, \boldsymbol{s}; O_{\boldsymbol{s}})$ is the population size, $f_1(t, \boldsymbol{s})$ corresponds to the incidence rate function, which is commonly used in disease surveillance applications (cf., Last 2001; Yang and Qiu 2020). Furthermore, the second-order intensity function is defined to be

$$f_2(t, t'; \boldsymbol{s}, \boldsymbol{s}') = \lim_{\substack{dt \to 0, d\boldsymbol{s} \to 0 \\ dt' \to 0, d\boldsymbol{s}' \to 0}} \frac{E\left[N^*(t, \boldsymbol{s}; dt, O_{\boldsymbol{s}}) N^*(t', \boldsymbol{s}'; dt', O_{\boldsymbol{s}'})\right]}{N(t, \boldsymbol{s}; O_{\boldsymbol{s}}) N(t', \boldsymbol{s}'; O_{\boldsymbol{s}'}) dt dt'}, \text{ for any } (t, \boldsymbol{s}), (t', \boldsymbol{s}') \in [0, \infty) \times \Omega.$$

The following assumptions are needed to facilitate the related theoretical results.

Assumption 1 (Spatial region and subregions)**.** The spatial region Ω is a bounded and closed set in $\mathbb{R}^2$, and the areas of the subregions at the finest scale satisfy $\max_{1 \leq j \leq m_L} |B_{Lj}| = O(|\Omega|/m_L)$, where $|B_{Lj}|$ and $|\Omega|$ represents the areas of B_{Lj} and Ω, respectively.

Assumption 2 (Data correlation)**.** The spatio-temporal correlation is short-ranged, in the sense that there exists two constants $C_1, C_2 > 0$ such that $\alpha_\kappa \leq C_1 \exp(-C_2 \kappa)$, for $\kappa \geq 1$.

Assumption 3 (Random errors)**.** The random errors have finite fourth moments. Namely, there exists a constant $C_\epsilon > 0$ such that $\mathrm{E}\left(|\epsilon_L(t_i, \boldsymbol{s}_{Lj})|^\psi\right) \leq C_\epsilon$ with $\psi \geq 4$, for all i and j.

Assumption 4 (Kernel function)**.** The kernel function $K(x)$ is a bounded, symmetric, Lipschitz-one continuous density function with a compact support.

Assumption 5 (Continuity)**.** The functions $f_1(t, \boldsymbol{s})$ and $f_2(t, t'; \boldsymbol{s}, \boldsymbol{s}')$ are twice continuously differentiable in their domains.

Assumption 1 requires that the area of the largest subregion at the finest scale is in the order of $|\Omega|/m_L$, ensuring each local neighborhood used when estimating the model (2) by a local kernel smoothing procedure contains enough data. Assumption 2 allows nonstationary spatio-temporal correlation to make our method flexible. But, the data correlation is required to be short-ranged in time, which is a routine requirement in time series analysis and spatio-temporal analysis (e.g., Li et al. 2007). Assumptions 3–4 are standard for studying statistical properties of kernel-based methods. Assumption 5 is commonly used when studying theoretical properties of the related methods for analyzing areal or point process data (cf., Müller et al. 1997; Li and Guan 2014).

Theorem 1. Under Assumptions 1–5, if $\log(n)^2/\left[n\min(h_t^2, g_t^2)\right] = o(1)$, $h_{\max} = o(1)$, $g_{\max} = o(1)$, and $\log(m_L)/\left[m_L \min(h_s^2, g_s^2)\right] = O(1)$, then we have

$$\max_{\substack{i=1,\ldots,n\\ j=1,\ldots,m_l}} |\widehat{\mu}_{l,0}(t_i, \boldsymbol{s}_{lj}) - \mu_{l,0}(t_i, \boldsymbol{s}_{lj})| = O_p\left(h_{\max}^2 + \left[\frac{\log(n)^2}{nh_t^2}\right]^{1/2}\right),$$

$$\max_{\substack{i,i'=1,\ldots,n\\ j,j'=1,\ldots,m_l}} \left|\widehat{V}_{l,0}(t_i, t_{i'}; \boldsymbol{s}_{lj}, \boldsymbol{s}_{lj'}) - V_{l,0}(t_i, t_{i'}; \boldsymbol{s}_{lj}, \boldsymbol{s}_{lj'})\right| = O_p\left(h_{\max}^2 + g_{\max}^2 + \left[\frac{\log(n)^2}{n\min(h_t^2, g_t^2)}\right]^{1/2}\right),$$

for $l = 1, \ldots, L$, where $h_{\max} = \max(h_t, h_s)$ and $g_{\max} = \max(g_t, g_s)$, (h_t, h_s) and (g_t, g_s) are bandwidths defined in Section A of the supplementary file.

Theorem 2. Under all the assumptions in Theorem 1, if the two tuning parameters ϕ_1 and ϕ_2 used in (6) and (7), the bandwidth h, and the adjustment factor $\mathcal{G}(m_L)$ satisfy the conditions that $\max(\phi_1, \phi_2)/\sqrt{m_L h^2}$=o(1), $1/\min(\phi_1, \phi_2) = o(1)$, $h = o(1)$, $1/\mathcal{G}(m_L) = o(1)$, and $\mathcal{G}(m_L)/\sqrt{m_L h^2} = o(1)$, then we have: for $i \geq n + 1$,

(i) $\sqrt{m_L h^2}\left[\widehat{\delta}_l(t_i, \boldsymbol{s}_{lj}) - \delta_l(t_i, \boldsymbol{s}_{lj})\right] \xrightarrow{d} N\left(0, \sigma_{lj}^2(t_i)\right)$, for $j \in \mathcal{A}_l(t_i)$, $l = 1, \ldots, L$;

(ii) $\lim_{m_L \to \infty} \Pr\left(\widehat{\mathcal{A}}_l(t_i) = \mathcal{A}_l(t_i)\right) = 1$, for $l = 1, \ldots, L$;

(iii) $\lim_{m_L \to \infty} \Pr\left(\widehat{\ell}(t_i) = \ell(t_i)\right) = 1$,

where $\xi_k = \int_0^\infty xK^k(x)dx$, for $k = 1$ and 2, and the asymptotic variance in the result (i) is

$$\sigma_{lj}^2(t_i) = \frac{\lambda\left[1-(1-\lambda)^{2(t_i-t_n)/\varsigma}\right]\left[f_2(t_i,t_i;\boldsymbol{s}_{lj},\boldsymbol{s}_{lj}) - f_1^2(t_i,\boldsymbol{s}_{lj})\right]\xi_2}{2\pi\xi_1^2(2-\lambda)\left[1-(1-\lambda)^{(t_i-t_n)/\varsigma}\right]^2}.$$

Theorem 1 establishes the consistency of the offline estimates for the IC model (2) based on data collected before sequential process monitoring, whereas Theorem 2 is focused on the online setting in which new batch of data keep coming over time (i.e., $i = n+1, n+2, \ldots$). Specifically, Result (i) in Theorem 2 gives the asymptotic normality of the estimated shift sizes $\{\widehat{\delta}_l(t_i, \boldsymbol{s}_{lj})\}$, Result (ii) gives the consistency of the detected anomalous regions, and Result (iii) confirms the consistency of the selected spatial scale level by the SIC score (9). Proofs of these results are provided in the supplementary file. It should be noted from Theorem 2 that consistency of the selected spatial scale level is guaranteed as long as the adjustment factor $\mathcal{G}(m_L)$ in (9) satisfies the two conditions that (i) $1/\mathcal{G}(m_L) = o(1)$, and (ii) $\mathcal{G}(m_L)/\sqrt{m_L h^2} = o(1)$. In this paper, it is assumed that $\log(m_L)/\sqrt{m_L h^2} = o(1)$. Then, we can simply choose $\mathcal{G}(m_L) = \log(m_L)$. Since the SIC criterion (9) corresponds to the BIC criterion (Schwarz 1978) that has some good theoretical properties (Yang 2005) when $\mathcal{G}(m_L) = \log(m_L)$, the related SIC criterion should share these properties. Based on extensive numerical studies, we find that the SIC criterion works well in practice.

3 Simulation Studies

In this section, we conduct simulation studies to evaluate the performance of the proposed method MST-FE and compare it with some state-of-the-art approaches. The simulation setup is as follows. Assume that the number of spatial scales is $L = 3$, $\Omega = [0,1] \times [0,1]$, $[0,T] = [0,1]$, and the IC observation times are $\{t_i = i/n : i = 1, \ldots, n\}$. Unless stated otherwise, it is further assumed that $m_1 = m^2, m_2 = 4m^2$, $m_3 = 16m^2$,

$$\begin{aligned}
\{B_{1j}\} =& \{[(j_1-1)/m, j_1/m] \times [(j_2-1)/m, j_2/m] : 1 \le j_1, j_2 \le m\},\\
\{B_{2j}\} =& \{[(j_1-1)/(2m), j_1/(2m)] \times [(j_2-1)/(2m), j_2/(2m)] : 1 \le j_1, j_2 \le 2m\},\\
\{B_{3j}\} =& \{[(j_1-1)/(4m), j_1/(4m)] \times [(j_2-1)/(4m), j_2/(4m)] : 1 \le j_1, j_2 \le 4m\},
\end{aligned}$$

and $\boldsymbol{s}_{lj}$ is the centroid of B_{lj}, for $1 \le l \le L$ and $1 \le j \le m_l$. In the simulation examples, (n, m) are chosen to be $(200, 4)$ unless stated otherwise.

The IC process observations at the finest scale (i.e., scale L) are generated from the baseline model (2), where the IC mean function is assumed to be

$$\mu_{L,0}(t, \boldsymbol{s}) = \sin(2\pi t) + \exp\left(-\frac{\|\boldsymbol{s} - \boldsymbol{s}_c\|^2}{2}\right), \quad \text{for } (t, \boldsymbol{s}) \in [0, 1] \times \Omega,$$

where $\|\mathbf{x}\|$ is the Euclidean norm of $\mathbf{x}$, and $\boldsymbol{s}_c = (0.5, 0.5)^\top$ is the center of Ω. Regarding the random errors, we consider five cases with different data correlation and distribution.

Case 1: The random errors $\{\epsilon_L(t_i, \boldsymbol{s}_{Lj}) : j = 1, \ldots, m_L\}$ are independent and follow $N(0, 1)$.

Case 2: The random errors follow the vector AR(1) model $\boldsymbol{\epsilon}_L(t_i) = 0.5\boldsymbol{\epsilon}_L(t_{i-1}) + \sqrt{1 - 0.5^2}\boldsymbol{\eta}(t_i)$, where $\boldsymbol{\epsilon}_L(t_i) = (\epsilon_L(t_i, \boldsymbol{s}_{L1}), \ldots, \epsilon_L(t_i, \boldsymbol{s}_{Lm_L}))^\top$, and $\boldsymbol{\eta}(t_i) = (\eta(t_i, \boldsymbol{s}_{L1}), \ldots, \eta(t_i, \boldsymbol{s}_{Lm_L}))^\top$ is a realization of a temporally independent Gaussian process with spatial covariance $\rho(\boldsymbol{s}, \boldsymbol{s}') = \exp\{-5||\boldsymbol{s} - \boldsymbol{s}'||\}$, for any $\boldsymbol{s}, \boldsymbol{s}' \in \Omega$.

Case 3: The random errors follow the vector AR(2) model $\boldsymbol{\epsilon}_L(t_i) = 0.6\boldsymbol{\epsilon}_L(t_{i-1}) + 0.3\boldsymbol{\epsilon}_L(t_{i-2}) + \boldsymbol{\eta}(t_i)$, where $\boldsymbol{\eta}(t_i)$ is a realization of a temporally independent Gaussian process with spatial covariance: for $\boldsymbol{s}, \boldsymbol{s}' \in \Omega$, $\rho(\boldsymbol{s}, \boldsymbol{s}') = \exp\left\{-10\left[(\boldsymbol{s} - \boldsymbol{s}')^\top \mathbf{M} (\boldsymbol{s} - \boldsymbol{s}')\right]^{1/2}\right\}$, where $\mathbf{M}$ is a 2×2 matrix with diagonal elements 1 and non-diagonal elements 0.5.

Case 4: The random errors follow the vector AR(1) model $\boldsymbol{\epsilon}_L(t_i) = 0.5\boldsymbol{\epsilon}_L(t_{i-1}) + \sqrt{1 - 0.5^2}\boldsymbol{\eta}(t_i)$, where $\boldsymbol{\eta}(t_i) = \boldsymbol{\eta}_1(t_i) + \boldsymbol{\eta}_2(t_i)$, $\{\boldsymbol{\eta}_1(t_i)\}$ are generated from a temporally independent Gaussian process with the same spatial covariance as that in Case 2, and all elements in $\{\boldsymbol{\eta}_2(t_i)\}$ are independent and follow the standardized $\chi^2(2)$ distribution.

Case 5: The random errors for each region B_{Lj} are generated from the AR(1) model $\epsilon_L(t_i, \boldsymbol{s}_{Lj}) = \rho_1(\boldsymbol{s}_{Lj})\epsilon_L(t_{i-1}, \boldsymbol{s}_{Lj}) + \sqrt{1 - \rho_1^2(\boldsymbol{s}_{Lj})}\eta(t_i, \boldsymbol{s}_{Lj})$, for $j = 1, \ldots, m_L$, where $\rho_1(\boldsymbol{s}) = \|\boldsymbol{s} - \boldsymbol{s}_c\|$, $\boldsymbol{s}_c = (0.5, 0.5)^\top$, $\{\boldsymbol{\eta}(t_i) = (\eta(t_i, \boldsymbol{s}_{L1}), \ldots, \eta(t_i, \boldsymbol{s}_{Lm_L}))^\top\}$ are generated from a temporally independent Gaussian process with the spatial covariance $\rho_2(\boldsymbol{s}, \boldsymbol{s}') = \exp\{-10||\boldsymbol{s} - \boldsymbol{s}'||^2\}$.

In all five cases, the random errors are centralized and scaled properly so that their means are zero and their variances are one. From the above description, it can be seen that the random errors

in the five cases have different spatio-temporal data correlation and distributions. Specifically, they are independent and identically distributed with the standard normal distribution in Case 1. The data correlation is stationary in both time and space in Case 2, and nonstationary in the space domain in Case 3. Case 4 is similar to Case 2, except that the noises follow a mixture of the normal and χ^2 distributions that is skewed. In Case 5, the spatio-temporal covariance is non-separable, as the temporal correlation varies across different regions and thus the covariance cannot be written as a product of a spatial correlation function and a temporal correlation function.

Figure 3 displays the random errors from a single simulation realization, where the three rows correspond to times t_1, t_2, and t_3, and the five columns correspond to Cases 1–5. From the figure, we observe that: (i) the independently generated random errors in Case 1 are highly noisy and exhibit no discernible spatial pattern; (ii) in Case 3, the errors at the three times display similar spatial patterns, whereas such similarity is not evident in Case 2, reflecting the stronger temporal AR(2) dependence in Case 3; (iii) Case 4 exhibits more extreme values due to the skewed mixture distribution; and (iv) Case 5 shows the clearest spatial patterns, which evolve more rapidly over time near the center of Ω and more slowly near the corners, consistent with the assumed nonseparable spatio-temporal correlation structure.

In the model (5), the following three types of shifts in the mean of the process are considered: Type (I): $\delta_L(t, \boldsymbol{s}) = 0.25\nu \mathrm{I}(\boldsymbol{s} \in \Delta)$; Type (II): $\delta_L(t, \boldsymbol{s}) = [0.5 + 0.1\nu \cos(2\pi t)]\, \mathrm{I}(\boldsymbol{s} \in \Delta)$; Type (III): $\delta_L(t, \boldsymbol{s}) = \left[0.5 + 0.1\cos(2\pi t) + 0.3\nu||\boldsymbol{s} - \boldsymbol{s}_c||^2\right] \mathrm{I}(\boldsymbol{s} \in \Delta)$, where the parameter ν controls the shift size, its value changes among $\{1, 2, 3, 4\}$, and Δ is the true region with process shifts. In the examples, unless stated otherwise, $\Delta = [3/8, 5/8] \times [3/8, 5/8]$. So, the optimal scale defined in (8) is 2 when $m = 4$. It is obvious that the three types of mean shifts correspond to a constant shift, a time-varying shift, and a space-time-varying shift, respectively.

So far, we have described how to generate IC and OC process observations at the finest scale L. They can be aggregated by the recursive aggregation formula (1) to obtain observed data at scales 1 to $L - 1$. Since each parent region contains four child regions with equal areas, we choose all weights $\omega_{lj}(t_i, \boldsymbol{s}_{l+1,k})$ in Equation (1) to be 0.25, for all $l, j, k \in \mathrm{ch}(l, j)$ and i.

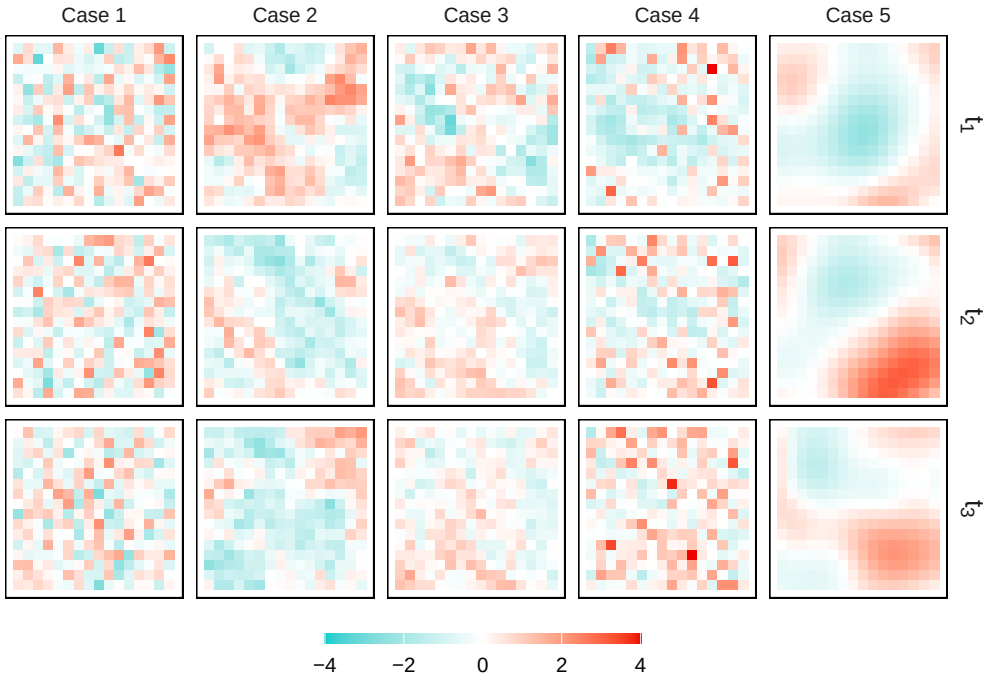


Figure 3: Simulated random errors at times t_1, t_2, and t_3 (rows 1–3) for Cases 1–5 (columns 1–5).

In the simulation, the IC mean, variance and covariance are assumed to be unknown and need to be estimated from the IC dataset $\{y_l(t_i, \boldsymbol{s}_{lj}) : 1 \leq l \leq L, 1 \leq j \leq m_l, 1 \leq i \leq n\}$. To determine the control limit γ of the chart (10), a second IC dataset of the same size is generated, and the block bootstrap procedure discussed in Section B of the supplementary file is used to calibrate γ, with the nominal ARL_0 fixed at 200. Together, the two IC datasets reflect the case study setting discussed in Section 4 where two years of IC data are available. Given the estimated baseline model and calibrated control limit, the actual ARL is computed by monitoring a sequence of process observations until a signal occurs, yielding a run length for each replication. This online monitoring procedure is repeated 1,000 times, and the mean of these run lengths is used as an estimate of ARL. Note that this estimated ARL depends on the randomly generated IC data used for baseline model estimation and control limit calibration. To reduce variability due to these IC datasets, the entire simulation process—including generation of the two IC datasets, estimation of the baseline model, determination of the control limit, and computation of the 1,000-run ARL—is repeated 100 times. The final reported ARL is then obtained as the average of the 100 ARL estimates, and the corresponding standard error is also computed.

3.1 Performance of scale selection and anomaly detection

Performance of scale selection. To evaluate the numerical performance of the proposed SIC criterion, we consider cases when $L = 3$, $\lambda = 0.1$, the nominal $\text{ARL}_0 = 200$, $n = 200$ and m changes from 2 to 6. Selection accuracy of the SIC criterion is measured by the proportion of correct estimation of the optimal scale by $\widehat{\ell}(t)$ among 1,000 replicated simulations. The results of selection accuracy in Cases 1-5 when detecting Type (III) shifts with $\nu = 1$ are shown in the right panel of Figure 4. As a comparison, the left panel of Figure 4 shows the corresponding results of the naïve method for choosing the optimal spatial scale discussed in the paragraph below Equation (8). It can be seen from the figure that (i) SIC performs much better than the naïve method in this example, (ii) the selection accuracy of SIC gets better when m is larger, implying the consistency verified theoretically in Theorem 2, and (iii) by comparing the results in Cases 1 and 2, the selection accuracy gets worse when data correlation is stronger, which is consistent with our intuition. In summary, the simulation results indicate the effectiveness and robustness of SIC across all scenarios considered, frequently selecting the optimal scale more often than the naïve approach.

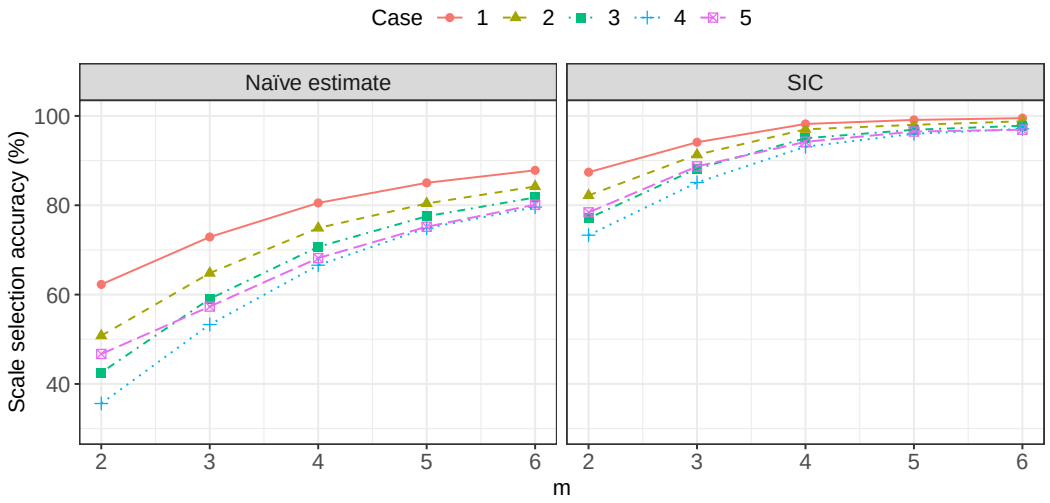


Figure 4: Spatial scale selection accuracy by the proposed SIC criterion (right panel) and the naïve method (left panel) in Cases 1-5 when detecting Type (III) shift with $\nu = 1$ by the proposed chart MST-FE with $\lambda = 0.1$.

Performance of anomaly detection. Next, we study the performance of MST-FE in detecting anomalous regions. To this end, let us consider scenarios when m can change among 2, 4 and

6. Other setup is the same as that in Figure 4. The proportion of times each region is detected as an anomalous region by the proposed MST-FE method when detecting Type (III) shift with $\nu = 1$ is computed based on 1,000 replicated simulations. The proportions of all regions in Cases 1-5 when $m = 2, 4$ and 6 are presented in Figure 5, where the boundary of the true anomalous region is denoted by the blue line in each plot. From the figure, we can make two main conclusions: (i) MST-FE is reliable in detecting the anomalous regions, and (ii) its performance gets better when m increases. The second conclusion is consistent with the theoretical results in Theorem 2. These results show that MST-FE achieves high accuracy in detecting anomalous regions, highlighting its strong performance and reliability under varying conditions.

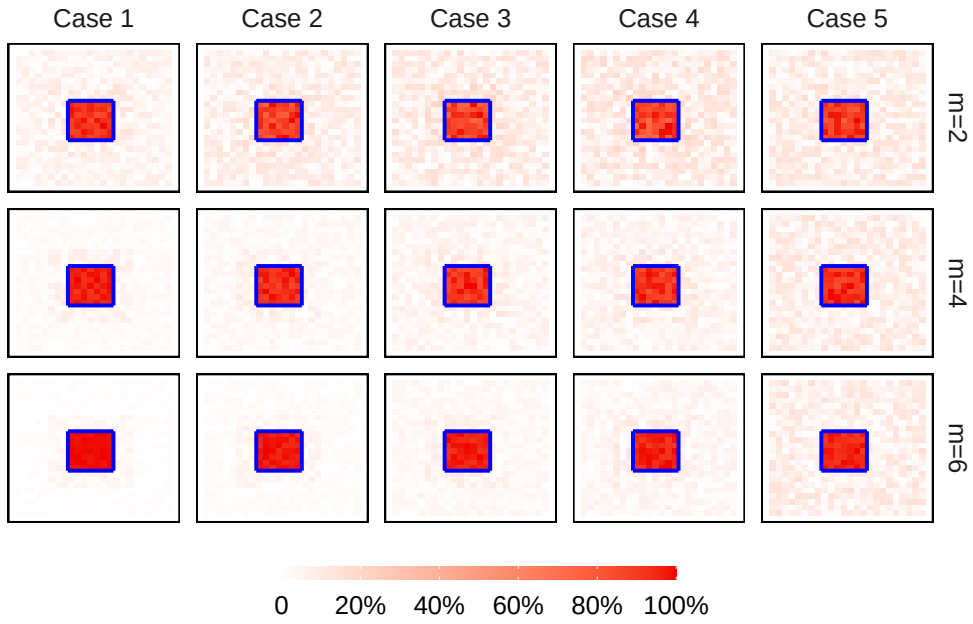


Figure 5: Proportion of times each region is detected as an anomalous region by MST-FE based on 1,000 replicated simulations in Cases 1-5 when detecting Type (III) shift with $\nu = 1$. In each plot, the blue line highlights the boundary of the true anomalous region.

3.2 Method Comparison

In this part, we compare the numerical performance of the proposed method MST-FE with some representative alternative methods. As discussed in Section 1, multiscale spatio-temporal process monitoring is understudied in the literature, and almost all existing methods are designed for

monitoring spatio-temporal data at a single spatial scale. Here, we consider three existing methods used to monitor observed data at the optimal spatial scale $\ell(t)$, where $\ell(t)$ keeps unchanged over time in all simulation examples considered in this part. Specifically, based on the idea of smooth sparse decomposition, Yan et al. (2018) proposed an online monitoring method for detecting anomalies in a spatio-temporal process, denoted as ST-SSD. This method can handle data with space-time-varying means and is capable of detecting sparse anomalies effectively. However, it assumes that observed data are independent over space and normally distributed. By using a kernel smoothing procedure, Zhao et al. (2011) suggested a spatio-temporal change detection method, denoted as ST-KS. This method can accommodate space-time-varying mean structures in a spatio-temporal process. But, it assumes that the temporal correlation can be described by an AR(2) time series model, observations for different regions are independent, and observed data are normally distributed. Based on the exponentially weighted spatial LASSO (EWSL) procedure, Yang and Qiu (2023) proposed a spatial variable-selection control chart for monitoring spatio-temporal processes, denoted as ST-EWSL. This method does not impose any parametric assumptions on space-time data variation, spatio-temporal data correlation, or data distribution. It can identify anomalous regions that are spatially clustered. However, it is computationally intensive when the number of regions is large. In addition, since this method relies on data decorrelation, a non-anomalous region could be mistakenly identified as anomalous. This occurs because the decorrelated observation in a non-anomalous region is a linear combination of data from all regions, including those from anomalous regions.

Next, we present the ARL_1 values of the four methods for detecting the three types of shifts in Cases 1–5. To this end, we consider scenarios when $L = 3$, $(n, m) = (200, 4)$, and the nominal ARL_0 is fixed at 200. To make the comparison fair, the control limits of the four methods, especially ST-SSD and ST-KS, have been adjusted so that their actual ARL_0 values equal to 200. In addition, different methods have different procedure parameters involved (e.g., λ in MST-FE). If these parameters are set at given values, the related methods may not be comparable (cf., Qiu 2008). To address this issue, we compare their optimal OC performance, in the sense that their parameters are chosen such that the smallest ARL_1 values are obtained for detecting a given shift. The optimal ARL_1 values of the four methods are shown in Figure 6. From the figure, we can make three conclu-

sions. First, ST-SSD has a better OC performance than ST-KS in Case 1 when observed data are i.i.d. and normal, and the two methods have similar performance in Cases 2-5 when observed data are correlated. Second, MST-FE and ST-EWSL have much better OC performance than ST-SSD and ST-KS in all cases considered because the former methods are effective for detecting clustered anomalous regions and can accommodate spatio-temporal data correlation and other complex data structure. Third, MST-FE performs better than ST-EWSL in all cases considered since the former can detect anomalous regions more accurately. We have similar results when spatial scales are non-nested or shifts occur in multiple regions. See Figure S4 in the supplementary file for more details. Moreover, we compare the IC performance of the four methods, as well as their computing times. The results are presented in Table S1 and Figure S5 of the supplementary file. We observe that MST-FE has satisfactory IC performance and its computing times increase slowly with m. Hence, MST-FE is capable of handling cases when the number of regions is large. Overall, these results show that MST-FE has the best performance among the four methods, offering superior anomaly detection, reliable IC performance, and efficient computation.

4 Case Study

In this section, we demonstrate the proposed method using the influenza-like illness (ILI) data provided by the Florida Electronic Surveillance System for the Early Notification of Community-based Epidemics (FL-ESSENCE). Here, ILI is defined to be an acute respiratory infection with a fever of 100°F or higher and cough or sore throat. In the United States, there are approximately 36,000 ILI-associated deaths and more than 200,000 ILI-associated hospitalizations each year (Thompson et al. 2004). Due to tremendous damage by ILI pandemics, the Florida Department of Health (FDOH) established the FL-ESSENCE system to collect near real-time ILI data from 229 emergency departments and 35 care centers that are distributed across the entire state. Researchers can have access to the Florida ILI incidence data that are organized at spatial scales of geographic region, county, and county subdivision, respectively, after a proper registration through the FDOH webpage (https://www.floridahealth.gov). The Florida state has 8 geographic regions, including Northwest,

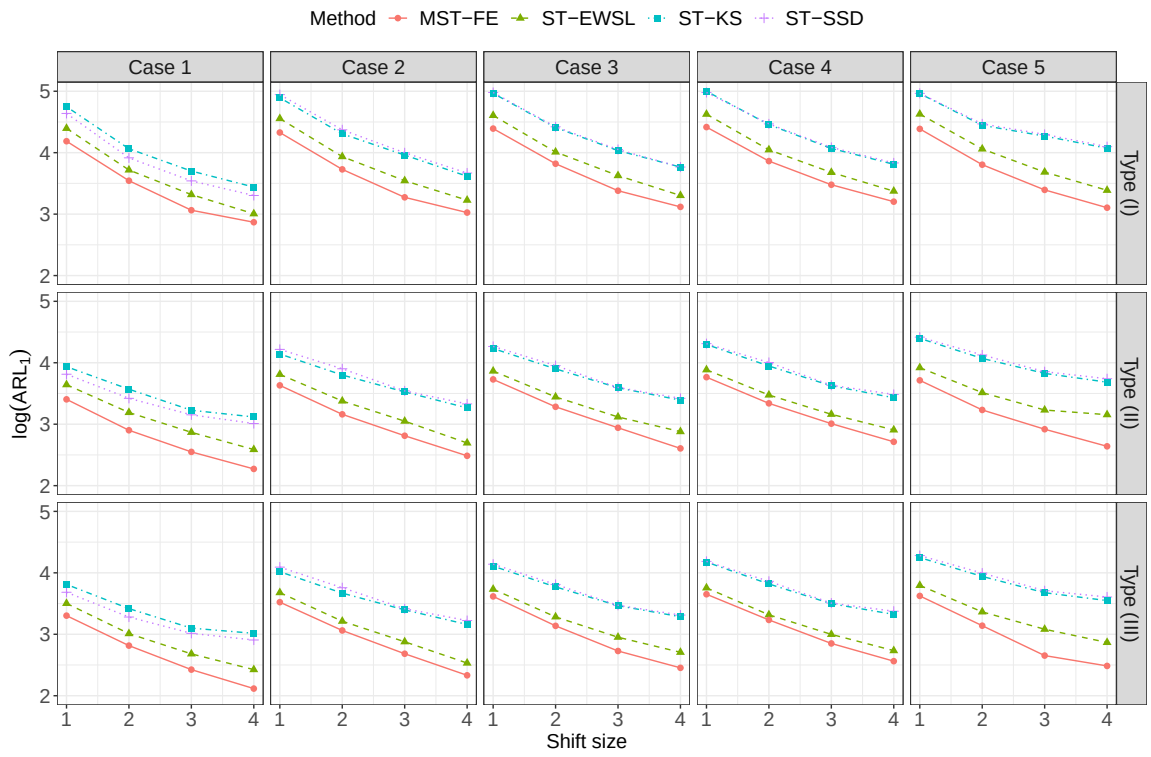


Figure 6: Optimal ARL_1 values in the log scale of the four methods for detecting the three types of shifts in Cases 1-5 when $L = 3$, $(n, m) = (200, 4)$, and $\mathrm{ARL}_0 = 200$.

North Central, Northeast, East Central, West Central, Southeast, Southwest, and South regions. Each geographic region contains several counties, and there are a total of 67 counties in Florida. Then, each county consists of one or more county subdivisions, and Florida consists of 316 county subdivisions. So, in this example, the three spatial scales are nested in a hierarchical way. Namely, county subdivisions are nested within counties, and counties are nested within geographic regions.

The Florida ILI data used in this paper were collected daily during the three years 2012-2014. The observed data on December 15 (a winter time) of the 3 years at the scales of geographic region, county, and county subdivision are shown in Figure 7. From the figure, it can be seen that the ILI incidence rates in the winter of 2014 seem much higher than those in the winters of 2012 and 2013. In addition, the bottom-left panel in Figure 7 shows that the Northwest region has a moderately high incidence rate in December 2014. But, the other two plots in the third row of Figure 7 suggest that some counties (e.g., Calhoun county) and county subdivisions in the Northwest Florida have quite low ILI incidence rates in that winter. Therefore, the spatial scale of geographic region may

not be a good choice for detecting disease outbreaks in this example.

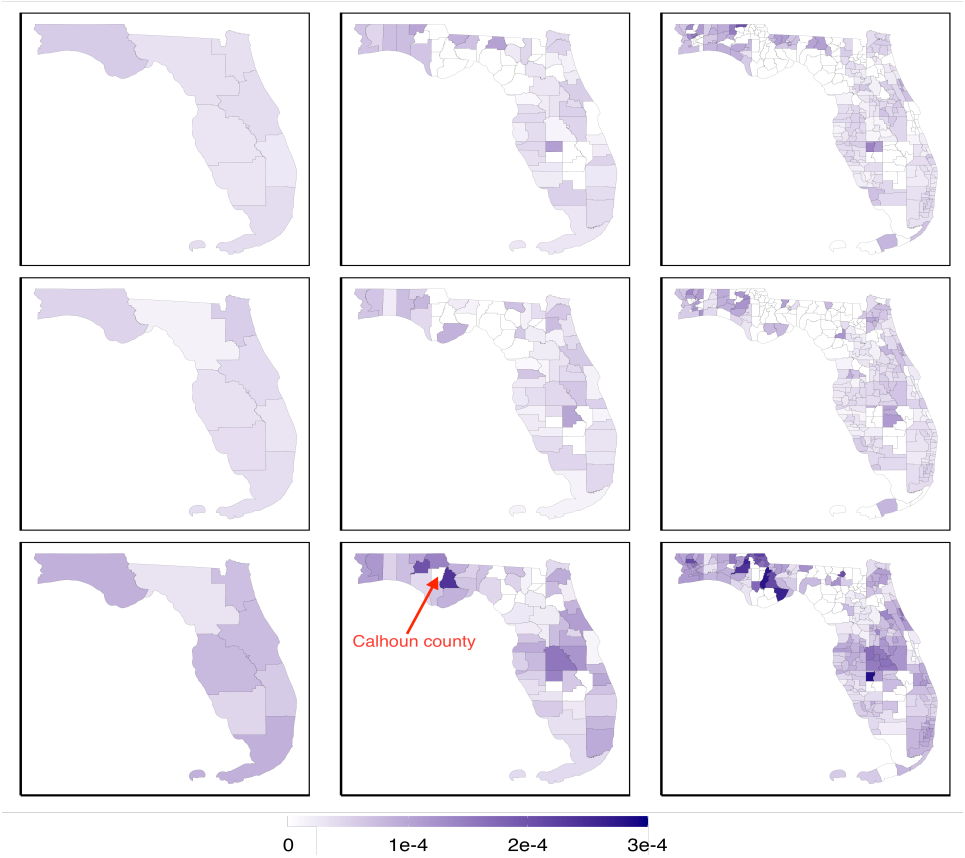


Figure 7: Observed ILI incidence rates on December 15 in years 2012 (1st row), 2013 (2nd row) and 2014 (3rd row) that are organized in geographic regions (1st column), counties (2nd column) and county subdivisions (3rd column).

To find an appropriate scale for effective detection of disease outbreaks, we apply our proposed method MST-FE to this dataset. To this end, the observed ILI incidence rates in 2013 and the observed data in 2012 are used as IC data for estimating the baseline model (2) and for determining the control limit γ of the chart MST-FE, respectively. In this study, we use data from 2013 rather than 2012 to estimate the baseline pattern, as the more recent data are likely to better reflect current disease dynamics. Moreover, the 2013 dataset provides a well-structured and sufficiently large sample at the county-subdivision scale. Specifically, the IC data from 2013 include observations from 316 county subdivisions over 365 days, yielding a total of 115,340 ILI incidence rates. This large dataset enables reliable estimation of both the annual seasonal pattern and spatio-temporal correlation using a local linear kernel smoothing procedure, under the assumption that the mean

and covariance functions are continuous in both space and time.

In the chart MST-FE, we set $\lambda = 0.1$ and $\text{ARL}_0 = 200$, as routinely used in SPC. The control limit γ is determined by the block bootstrap procedure discussed in Section B of the supplementary file with $N_b = 1,000$ bootstrap replications and $\tau = 5$. After implementing MST-FE, we find that the selected optimal scale by SIC is at the county level at most observation times, which is consistent with the observation from Figure 7 that is discussed above. Besides MST-FE, the three alternative charts ST-EWSL, ST-KS and ST-SSD discussed in Subsection 3.2 are also considered here for comparison purposes and applied to county-scale data. For ST-EWSL, its weighting parameter is chosen to be 0.1. The control limits of the three alternative methods are determined in the same way as that in Subsection 3.2. The charting statistics of the four charts for monitoring the observed ILI incidence rates in 2014 are shown in Figure 8. From the figure, it can be seen that (i) ST-KS and ST-SSD give signals as early as January 2014 and a large amount of alarms is given by them, and (ii) MST-FE gives its first signal of disease outbreak on September 17, 2014, which is 19 days earlier than the first signal of ST-EWSL on October 6, 2014. Because both ST-KS and ST-SSD require restrictive assumptions on observed data, their signals may not be reliable to use, as demonstrated by the simulation results in Table S1 of the supplementary file. In addition, their frequent signals would cause waste of human/financial resource and/or neglect of related alarms (cf., Shmueli and Burkom 2010). As a comparison, MST-FE and ST-EWSL would be more reliable to use. Between the two methods, this example confirms that MST-FE is more effective for monitoring the multiscale ILI incidence rates since its first signal is 19-day earlier than that of ST-EWSL.

5 Concluding Remarks

Multiscale spatio-temporal process monitoring is an important research problem with broad applications. In this paper, we introduce a new method to solve this problem based on the proposed forward elimination strategy and SIC criterion. The proposed method is fully nonparametric, effective for detecting anomalous regions that are spatially clustered, and capable of handling large spatio-temporal data. Both theoretical justifications and numerical studies have shown that the pro-

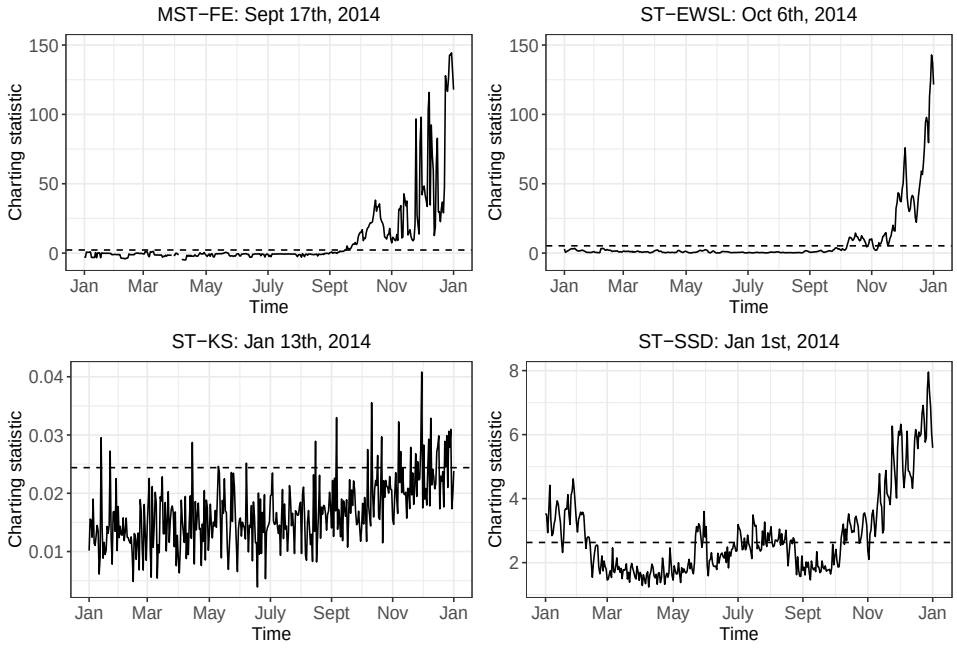


Figure 8: Four charts MST-FE, ST-EWSL, ST-KS, and ST-SSD for monitoring the Florida ILI incidence rates in 2014. In each plot, the dashed horizontal line represents the control limit of the chart, and the date given in the title of the plot is its first signal time.

posed method provides a useful tool for sequentially monitoring multiscale spatio-temporal data.

The proposed method is designed to detect deviations from a baseline spatio-temporal pattern, regardless of whether these deviations arise from dynamic seasonality (i.e., evolving seasonal patterns) or other types of process shifts. Although the method does not distinguish between changes caused by dynamic seasonality and those unrelated to it, it remains valuable in many practical settings. First, in numerous applications, timely detection of departures from the baseline is critical, even when such deviations are driven by dynamic seasonality. For example, shifts in seasonal disease patterns—such as changes in peak timing or unusually high incidence—may be associated with climate change or altered social behavior, and early detection is essential for enabling prompt interventions. Second, in some applications, historical IC data from multiple periods may be unavailable, particularly for emerging diseases with limited records (e.g., COVID-19). In such cases, the proposed method remains applicable and effective for detecting changes using IC data from only one or two periods. Third, the proposed method can serve as a foundation for more advanced monitoring systems that explicitly account for dynamic seasonality. When the focus is primarily on

changes unrelated to seasonal dynamics, the baseline model in the current method can be modified to accommodate evolving seasonal patterns. In this context, methods suggested by Chen et al. (2014) and So and Chung (2014) may be useful for model generalization. Estimating such generalized models would require IC data from multiple periods (e.g., several years in disease surveillance applications). This generalization will be explored in future work.

There are some other issues about the proposed method that need to be considered carefully. First, the method uses the penalized exponentially weighted kernel smoothing procedure to estimate a process shift under the assumption that the shift is smooth in the space domain. This may cause difficulties in detecting scattered or isolated shifts. To address this issue, a robust online monitoring method that are effective for detecting both clustered and isolated shifts would be useful. In addition, the penalty term in the penalized exponentially weighted kernel smoothing procedure (6)–(7) is chosen to be the L_1 adaptive spatial LASSO penalty, due to its good computational and theoretical properties. In the literature, there are some alternative penalties, such as the smoothly clipped absolute deviation (Fan and Li 2001), and the elastic net (Zou and Hastie 2005). It would be interesting to study the impact of the selected penalty function on the performance of the method. Moreover, the proposed method assumes that IC pattern is periodic in the time domain with a known period, which is reasonable in some applications, such as infectious disease surveillance and environmental monitoring. In some other applications, however, the period may be unknown and needs to be estimated properly. In such cases, the nonparametric estimation method proposed by Vogt and Linton (2014) and the cross-validation-based procedure suggested by Sun et al. (2012) should be useful. All these issues need to be studied carefully in future research.

Acknowledgments: The authors would like to thank the Editor, the Associate Editor, and two anonymous referees for their insightful comments and constructive suggestions, which improved the quality of the paper greatly.

Data Availability Statement: The authors confirm that the data supporting the findings of this study are available in the supplementary materials.

Disclosure Statement: The authors report there are no competing interests to declare.

References

Berliner, L.M., Wikle, C.K., and Milliff, R.F. (1999), "Multiresolution wavelet analyses in hierarchical Bayesian turbulence models," in *Bayesian inference in wavelet based models*, eds. P. Müller and B. Vidakovic, New York: Springer-Verlag, pp. 341–359.

Chen, Y.C., Cheng, M.Y., and Wu, H.T. (2014), "Non-parametric and adaptive modelling of dynamic periodicity and trend with heteroscedastic and dependent errors," *Journal of the Royal Statistical Society (Series B)*, **76**, 651–682.

Elkhouly, M., and Ferreira, M.A.R. (2021), "Dynamic multiscale spatiotemporal models for multivariate Gaussian data," *Spatial Statistics*, **41**, 100475.

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

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications,* London, UK: Chapman and Hall.

Fan, J., and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, **96**, 1348–1360.

Ferreira, M.A.R., Holan, S.H., and Bertolde, A.I. (2011), "Dynamic multiscale spatio-temporal models for Gaussian areal data," *Journal of the Royal Statistical Society (Series B)*, **73**, 663–688.

Fonseca, T.C.O., and Ferreira, M.A.R. (2017), "Dynamic multiscale spatiotemporal models for Poisson data," *Journal of the American Statistical Association*, **112**, 215–234.

Ibragimov, I. A., and Linnik, Y. V. (1971), *Independent and Stationary Sequences of Random Variables*, Gronigen, The Netherlands: Wolters-Noordhoff Publishing.

Johannesson, G., Cressie, N., and Huang, H. (2007), "Dynamic multi-resolution spatial models," *Environmental and Ecological Statistics*, **14**, 5–25.

Knox, E., and Bartlett, M. (1964), "The detection of space-time interactions," *Journal of the Royal Statistical Society (Series C)*, **13**, 25–30.

Kolaczyk, E.D., and Huang, H. (2001), "Multiscale statistical models for hierarchical spatial aggregation," *Geographical Analysis*, **33**, 95–118.

Kulldorff, M. (1997), "A spatial scan statistic," *Communications in Statistics–Theory and Methods*, **26**, 1481–1496.

Kulldorff, M., and Hjalmars, U. (1999), "The Knox method and other tests for space-time interactions," *Biometrics*, **55**, 544–552.

Last, J.M. (2001), *A Dictionary of Epidemiology (4th edition)*, Oxford University Press.

Li, B., Genton, M.G., and Sherman, M. (2007), "A nonparametric assessment of properties of space–time covariance functions," *Journal of the American Statistical Association*, **102**, 736–744.

Li, Y., and Guan, Y. (2014), "Functional principal component analysis of spatiotemporal point processes with applications in disease surveillance," *Journal of the American Statistical Association*, **109**, 1205–1215.

Louie, M.M., and Kolaczyk, E.D. (2006), "A multiscale method for disease mapping in spatial epidemiology," *Statistics in Medicine*, **25**, 1287–1306.

Müller, H.G, Stadtmüller, U., and Tabnak, F. (1997), "Spatial smoothing of geographically aggregated data, with application to the construction of incidence maps," *Journal of the American Statistical Association*, **92**, 61–71.

Qiu, P. (2008), "Distribution-free multivariate process control based on log-linear modeling," *IIE Transactions*, **40**, 664–677.

Qiu, P. (2014), *Introduction to Statistical Process Control*, Boca Raton, FL: CRC Press.

Qiu, P., and Xiang, D. (2014), "Univariate dynamic screening system: An approach for identifying individuals with irregular longitudinal behavior," *Technometrics*, **56**, 248–260.

Qiu, P., and Yang, K. (2021), "Effective disease surveillance by using covariate information," *Statistics in Medicine*, **40**, 5725–5745.

Qiu, P., and Yang, K. (2023), "Spatio-temporal process monitoring using exponentially weighted spatial LASSO", *Journal of Quality Technology*, **55**, 163–180.

Schwarz, G. (1978),"Estimating the dimension of a model," *The Annals of Statistics*, **6**, 461–464.

Schoenberg, F. P. (2003), "Multidimensional Residual Analysis of Point Process Models for Earthquake Occurrences," *Journal of the American Statistical Association*, **98**, 789–795.

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

Shmueli, G., and Burkom, H. (2010), "Statistical challenges facing early outbreak detection in biosurveillance," *Technometrics*, **52**, 39–51.

So, M.K.P, and Chung, R.S.W. (2014), "Dynamic seasonality in time series," *Computational Statistics and Data Analysis*, **70**, 212–226.

Sun, Y., Hart, J.D., and Genton, M.G. (2012), "Nonparametric inference for periodic sequences," *Technometrics*, **54**, 83–96.

Takahashi, K., Kulldorff, M., Tango, T., and Yih, K. (2008), "A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring," *International Journal of Health Geographics*, **7**, article 14.

Thompson, W.W., Shay, D.K.,Weintraub, E., Brammer, L., Bridges, C.B., Cox, N.J., and Fukuda, K. (2004), "Influenza-associated hospitalizations in the United States," *Journal of the American Medical Association*, **292**, 1333–1340.

Tibshirani, R., and Taylor, J. (2011), "The solution path of the generalized LASSO," *The Annals of Statistics*, **39**, 1335–1371.

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

Wang, A., Xian, X., Tsung, F., and Liu, K. (2018), "A spatial-adaptive sampling procedure for online monitoring of big data streams," *Journal of Quality Technology*, **50**, 329–343.

Woodall, W.H., Marshall, J.B., Joner, M.D., Jr, Fraker, S.E., and Abdel-Salam, A.S.G. (2008), "On the use and evaluation of prospective scan methods for health-related surveillance," *Journal of the Royal Statistical Society (Series A)*, **171**, 223–237.

Yan, H., Paynabar, K., and Shi, J. (2018), "Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition," *Technometrics*, **60**, 181–197.

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

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

Yang, K., and Qiu, P. (2020), "Online sequential monitoring of spatio-temporal disease incidence rates," *IISE Transactions*, **52**, 1218–1233.

Yang, Y. (2005), "Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation," *Biometrika*, **92**, 937–950.

Zhao, Y., Zeng, D., Herring, A.H., Ising, A., Waller, A., Richardson, D., and Kosorok, M.R. (2011), "Detecting disease outbreaks using local spatiotemporal methods," *Biometrics*, **67**, 1508–1517.

Zou, H. (2006), "The adaptive LASSO and its oracle properties," *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou, H., and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society (Series B)*, **67**, 301–320.