Journal of Nonparametric Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gnst20

Nonparametric estimation and inference for spatiotemporal epidemic models

Yueying Wang, Myungjin Kim, Shan Yu, Xinyi Li, Guannan Wang & Li Wang

To cite this article: Yueying Wang, Myungjin Kim, Shan Yu, Xinyi Li, Guannan Wang & Li Wang (2021): Nonparametric estimation and inference for spatiotemporal epidemic models, Journal of Nonparametric Statistics, DOI: 10.1080/10485252.2021.1988084

To link to this article: https://doi.org/10.1080/10485252.2021.1988084

View supplementary material

Published online: 20 Nov 2021.

Submit your article to this journal

Article views: 152

View related articles

View Crossmark data

JOURNAL OF NONPARAMETRIC STATISTICS
https://doi.org/10.1080/10485252.2021.1988084

Taylor & Francis
Taylor & Francis Group

Nonparametric estimation and inference for spatiotemporal epidemic models

Yueying Wang[a], Myungjin Kim[a], Shan Yu[b], Xinyi Li[c], Guannan Wang[d] and Li Wang[a]

[a]Department of Statistics, Iowa State University, Ames, IA, USA; [b]Department of Statistics, University of Virginia, Charlottesville, VA, USA; [c]School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC USA; [d]Department of Mathematics, William & Mary College, Williamsburg, VA, USA

ABSTRACT

Epidemic modelling is an essential tool to understand the spread of the novel coronavirus and ultimately assist in disease prevention, policymaking, and resource allocation. In this article, we establish a state-of-the-art interface between classic mathematical and statistical models and propose a novel space-time epidemic modelling framework to study the spatial-temporal pattern in the spread of infectious diseases. We propose a quasi-likelihood approach via the penalised spline approximation and alternatively reweighted least-squares technique to estimate the model. The proposed estimators are consistent, and the asymptotic normality is established for the constant coefficients. Utilizing spatiotemporal analysis, our proposed model enhances the dynamics of the epidemiological mechanism and dissects the spatiotemporal structure of the spreading disease. We evaluate the numerical performance of the proposed method through a simulation example. Finally, we apply the proposed method in the study of the devastating COVID-19 pandemic.

ARTICLE HISTORY
Received 28 October 2020
Accepted 27 September 2021

KEYWORDS
Coronavirus; nonparametric modelling; partially linear models; spatial epidemiology; splines; varying coefficient models

AMS SUBJECT CLASSIFICATIONS
62G05; 62G08; 62G20

1. Introduction

Since the beginning of the reported cases in December 2019, the outbreak of COVID-19 has spread globally within weeks. To assist in prevention efforts and ultimately stop the pandemic, it is crucial to identify the vulnerable communities and why they are more likely to be infected. One way to answer these questions is through scientific modelling (Gog 2020; Vespignani et al. 2020). Several attempts have been made to model and forecast the spread and mortality of COVID-19; for example, see Kucharski et al. (2020) and Sun et al. (2020).

The fundamental concept of infectious disease epidemiology is the investigation of how infections spread. Mathematical methods, such as the class of susceptible-infectious-recovered (SIR) models (Brauer, Van den Driessche, and Wu 2008; Pfeiffer et al. 2008; Lawson, Banerjee, Haining, and Ugarte 2016), are widely used in epidemics to capture the

CONTACT Li Wang lilywang@iastate.edu Department of Statistics, Iowa State University, Ames, IA 50011, USA

Supplemental data for this article can be accessed here. https://doi.org/10.1080/10485252.2021.1988084

dynamic process of the spread of infectious diseases. These mechanistic models characterise disease transmission through a set of differential equations, and they demonstrate the average behaviour of the epidemic. However, these mathematical models typically focus more on the descriptions of physical phenomena rather than the parameter estimation for observed data, and thus it is challenging to use them to make statistical inferences. On the other hand, statistical models are a class of data-driven methods, and they usually focus on the inference about the relationships between variables. For example, when analysing the confirmed cases and deaths of COVID-19, other factors, such as demographics, socioeconomic status, mobility, and control policies, may also be responsible for temporal or spatial patterns. In this article, we create a state of the art interface between mathematical models and statistical models to understand the dynamic pattern of the spread of contagious diseases, such as COVID-19 and many others.

We borrow the mechanistic rules from the SIR model and form a data-driven model with three compartments: infectious, susceptible, and removed states. The capacity of the health care system, and control measures, such as government-mandated social distancing, also have a significant impact on the spread of the epidemic. Since it is challenging to incorporate those factors in mathematical models, we borrow the strength from statistical models and include various explanatory variables to study not only the spatiotemporal structure but also the effects of the explanatory variables. Notice that the spread of the disease varies a lot across different geographical regions; we incorporate discrete-time spatially varying coefficient models to different compartments to reconstruct the spatiotemporal dynamics of the disease transmission. In general, the spatiotemporal models are able to bring in more information to the epidemic study (Held, Hens, D O'Neill, and Wallinga 2019; Jia et al. 2020).

With an emerging disease such as COVID-19, it is hard to measure many features of the transmission process, which may take a long time to understand fully. Thus, it is desirable to make inferences from observed data as model-free as possible. For a parametric epidemic model, the typical inference problem involves estimating the parameters associated with the parametric models from the data at hand. Such specifications are ad hoc, and if misspecified, can lead to substantial estimation bias problems. This issue might be addressed in practice by considering alternative nonparametric models or sensitivity analyses if some of the underlying model parameters are assumed to be known. By adopting a nonparametric approach, we do not impose a particular parametric structure, which significantly enhances the flexibility of the parametric epidemic models. Nonparametric approaches to fitting epidemic models to the data have received relatively little attention in the literature, possibly due to the lack of data. With the rich COVID-19 epidemic data released every day, we can consider the nonparametric method to model the covariates and coefficient functions.

By allowing the response (such as infected and death counts) to depend on time and location, we consider a generalised additive varying coefficient model to estimate the unobserved process of the disease transmission. We propose a quasi-likelihood approach via the penalised spline approximation and an iteratively reweighted least-squares technique for our model estimation. Our proposed algorithm is sufficiently fast and efficient for the user to analyse large datasets within seconds. We derive the asymptotical normality of the constant coefficients in the linear components under some regularity conditions.

For the varying components, we show the consistency of the estimators and obtain their convergence rates.

Finally, as an empirical illustration, we apply the proposed model and estimation method to a study of COVID-19 at the county level in the US. We illustrate how the proposed method can be used to analyse the spatiotemporal dynamics of the disease spread and guide evidence-based decision-making. Modeling COVID-19 at the county level and combining local characteristics are beneficial for the community in understanding the dynamics of the disease spread and support decision-making when urgently needed.

The rest of the paper is organised as follows. Section 2 outlines the nonparametric spatiotemporal modelling framework and describes how to incorporate additional covariates. Section 3 introduces the estimation method, presents the asymptotic properties of the estimators, the computational algorithm, and details of the implementation. Section 4 evaluates the finite sample performance of the proposed method using a simulation study. Section 5 describes the epidemic and endemic data, the results and findings of the case study. Section 6 concludes the paper with a discussion. Supplementary Material A contains some animation videos of the dynamic estimation results in the COVID-19 study, and Supplementary Material B provides the technical assumptions and detailed proofs of the asymptotic results.

2. Space-time epidemic modelling

To study the spatiotemporal pattern of COVID-19, we develop a novel spatiotemporal epidemic model (STEM) to estimate and predict the infection and death cases at the area level based on the idea of the compartment models. For simplicity, we introduce the STEM based on the parsimonious SIR models, but it can be extended to the SEIR models with an extra 'exposed' compartment for infected but not infectious individuals.

For area i and day t, let Y_{it} be the number of new cases, and let I_{it}, D_{it}, R_{it}, and S_{it} be the number of accumulated active infectious cases, accumulated death cases, accumulated recovered cases, and susceptible population, respectively. Let N_i be the total population for the ith area, and denote $Z_{it} = \log(S_{it}/N_i)$. Let $\mathbf{U}_i = (U_{i1}, U_{i2})^\top$ be the GPS coordinates of the geographic centre of area i, which ranges over a bounded domain $\Omega \subseteq \mathbb{R}^2$ of the region under study. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iq})^\top$ be a q-dimensional vector of explanatory variables collected from the U.S. Census Bureau. For example, the socioeconomic factors, health resources, and demographic conditions. Let A_{ijt} be the jth dummy variable of actions or continuous measures taken for area i at time t, and let $\mathbf{A}_{it} = (A_{i1t}, \ldots, A_{ipt})^\top$, which varies with the time.

In this paper, we consider the exponential families of distributions. The conditional density of Y_{it} given $(I_{i-1}, Z_{i,t-1}, \mathbf{A}_{i,t-r}, \mathbf{X}_i, \mathbf{U}_i) = (w, z, \boldsymbol{a}, \boldsymbol{x}, \boldsymbol{u})$ can be represented as

$$f\left(y \mid w, z, \boldsymbol{a}, \boldsymbol{x}, \boldsymbol{u}\right) = \exp\left[\frac{1}{\sigma^2}\left\{y\zeta\left(w, z, \boldsymbol{a}, \boldsymbol{x}, \boldsymbol{u}\right) - \mathcal{B}\left\{\zeta\left(w, z, \boldsymbol{a}, \boldsymbol{x}, \boldsymbol{u}\right)\right\}\right\} + \mathcal{C}\left(y, \sigma^2\right)\right],$$

for some known functions $\mathcal{B}$ and $\mathcal{C}$, dispersion parameter σ^2 and the canonical parameter ζ.

As mentioned earlier, the idea of the proposed model is based on conventional compartmental models and time series regression models. The conventional compartmental models aim to predict the number of susceptible individuals, infected cases, and recoveries. Specifically, in a SIR model, the progression is that a susceptible individual becomes infected through contact with one or more infected individuals. Then after a period, the infected individual advances to the recovery state, i.e. a noncontagious state. Therefore, at any given time, the SIR model captures the dynamical mechanism of the disease spread, assuming that the rate at which susceptible individuals become infected depends on the number of susceptible and infected individuals. In general, the dynamics of the SIR system inform us of the deterministic skeleton on which the behaviour of the corresponding stochastic system is built. In contrast, time series regression predicts the future of the response variable based on its history. It produces a combination of the variables with weights that indicate the variable importance. It also provides a neat result of how the predictors affect the response. Furthermore, we assume that the determinants of the daily new cases of a particular area can be explained not only by the features of that area but also by the spread of the virus in the surrounding areas. Therefore, by combining the advantages of compartmental and statistical models, we develop a novel discrete-time spatial epidemic model comprising the susceptible state, infected state, removed state, and area-level characteristics. Below we use superscripts I, D, and R to denote infected, death, and recovered states. We assume that the conditional mean value of daily new positive cases (μ_{it}^{I}), fatal cases (μ_{it}^{D}) and recovery (μ_{it}^{R}) in area i and day t can be modelled via a link function g as follows:

$$g(\mu_{it}^{\mathrm{I}}) = \beta_{0t}^{\mathrm{I}}(\mathbf{U}_i) + \beta_{1t}^{\mathrm{I}}(\mathbf{U}_i)\log(I_{i,t-1}) + \alpha_{0t}^{\mathrm{I}} Z_{i,t-1} + \sum_{j=1}^{p} \alpha_{jt}^{\mathrm{I}} A_{ij,t-m} + \sum_{k=1}^{q} \gamma_{kt}^{\mathrm{I}}(X_{ik}), \quad (1)$$

$$g(\mu_{it}^{\mathrm{D}}) = \beta_{0t}^{\mathrm{D}}(\mathbf{U}_i) + \beta_{1t}^{\mathrm{D}}\log(I_{i,t-\delta}) + \sum_{j=1}^{p} \alpha_{jt}^{\mathrm{D}} A_{ij,t-m'} + \sum_{k=1}^{q} \gamma_{kt}^{\mathrm{D}}(X_{ik}), \quad (2)$$

$$\mu_{it}^{\mathrm{R}} = \nu_t^{\mathrm{R}} I_{i,t-\delta'}, \quad (3)$$

where α_{jt}^{I}'s, α_{jt}^{D}'s, β_{1t}^{D} and ν_t^{R} are unknown constant coefficients, $\beta_{0t}^{\mathrm{I}}(\cdot)$, $\beta_{1t}^{\mathrm{I}}(\cdot)$, and $\beta_{0t}^{\mathrm{D}}(\cdot)$ are unknown bivariate coefficient functions, $\gamma_{kt}^{\mathrm{I}}(\cdot), \gamma_{kt}^{\mathrm{D}}(\cdot), k = 1, \ldots, q$, are univariate functions to be estimated, δ and δ' are the time delay between illness and death or recovery, and the parameter m in $A_{ij,t-m}$'s denotes a small delay time allowing for the control measure to be effective (here we take $m = 7$, $m' = 7$). For the new infection, following the assumptions in SIR, we assume that the number of the newly infected cases at time t depends on the situation of the pandemic at time $t-1$. For the death, according to CDC (2021a), the median number of days from symptom onset to death is around an incubation period. For the recovery, according to CDC (2021b), people with mild to moderate COVID-19 remain infectious no longer than ten days after their symptoms began, and those with more severe illness or those who are severely immunocompromised remain infectious no longer than 20 days after their symptoms began. As a result, we choose δ to be 14 and $\delta' = 10$ in our real data analysis. For model identifiability, similar to the conventional generalised additive model literature (Hastie and Tibshirani 1990; Wood 2017), we assume $\mathrm{E}\{\gamma_{kt}^{\mathrm{I}}(X_k)\} = 0$, $\mathrm{E}\{\gamma_{kt}^{\mathrm{D}}(X_k)\} = 0$, $k = 1, \ldots, q$. The STEM encompasses many existing models as special cases. For example, the traditional generalised linear regression models, generalised additive models (Liu, Yang, and Härdle 2013), generalised partially linear additive models

(Wang, Liu, Liang, and Carroll 2011) and generalised additive coefficient model (Xue and Liang 2010).

Note that, for the log link, $\exp\{\beta_{0t}^{\mathrm{I}}(\boldsymbol{u})\}$ illustrates the transmission rate at location $\boldsymbol{u}$, $\exp\{\beta_{0t}^{\mathrm{D}}(\boldsymbol{u})\}$ represents the fatality rate at location $\boldsymbol{u}$, ν_t^{R} is the recovery rate, $\beta_{1t}^{\mathrm{I}}(\cdot)$, α_{0t}^{I} and β_{1t}^{D} are the mixing parameters of the contact process. The rationale for including $\beta_{1t}^{\mathrm{I}}(\cdot)$ and β_{1t}^{D} ($\beta_{1t}^{\mathrm{I}}(\cdot) > 0$, $\beta_{1t}^{\mathrm{D}} > 0$) is to allow for deviations from mass action and to account for the discrete-time approximation to the continuous time model (Finkenstädt and Grenfell 2000; Wakefield, Dong, and Minin 2019). In many cases, the standard bilinear form may not necessarily hold. The above proposed epidemic model incorporates the nonlinear incidence rates, which represents a much wider range of dynamical behaviour than those with bilinear incidence rates (Liu, Hethcote, and Levin 1987). These dynamical behaviours are determined mainly by $\beta_{0t}^{\mathrm{I}}(\cdot)$, $\beta_{1t}^{\mathrm{I}}(\cdot)$, $\beta_{0t}^{\mathrm{D}}(\cdot)$, and β_{1t}^{D}. For example, when $\beta_{1t}^{\mathrm{I}}(\cdot)$ and α_{0t}^{I} are both 1, it corresponds to the standard assumption of homogeneous mixing in Jong, Diekmann, and Heesterbeek (1995).

In our study, since we model the number of new cases at time t for area i, Poisson or Negative Binomial (NB) might be an appropriate option for random component (Kim and Wang in press; Yu, Wang, Wang, Liu, and Yang 2020). For example, for the infection model, we can assume that

- (Poisson) $\mathrm{E}(Y_{it} \mid I_{i-1}, Z_{i,t-1}, \mathbf{A}_{i,t-m}, \mathbf{X}_i, \mathbf{U}_i) = \mu_{it}^{\mathrm{I}}$, $\mathrm{Var}(Y_{it} \mid I_{i-1}, Z_{i,t-1}, \mathbf{A}_{i,t-m}, \mathbf{X}_i, \mathbf{U}_i) = \mu_{it}^{\mathrm{I}}$,
- (NB) $\mathrm{E}(Y_{it} \mid I_{i-1}, Z_{i,t-1}, \mathbf{A}_{i,t-m}, \mathbf{X}_i, \mathbf{U}_i) = \mu_{it}^{\mathrm{I}}$, $\mathrm{Var}(Y_{it} \mid I_{i-1}, Z_{i,t-1}, \mathbf{A}_{i,t-m}, \mathbf{X}_i, \mathbf{U}_i) = \mu_{it}^{\mathrm{I}}(1 + \frac{\mu_{it}^{\mathrm{I}}}{I_{i,t-1}})$,

where μ_{it}^{I} can be modelled via the same log link as follows:

$$\log(\mu_{it}^{\mathrm{I}}) = \beta_{0t}^{\mathrm{I}}(\mathbf{U}_i) + \beta_{1t}^{\mathrm{I}}(\mathbf{U}_i)\log(I_{i,t-1}) + \alpha_{0t}^{\mathrm{I}} Z_{i,t-1} + \sum_{j=1}^{p} \alpha_{jt}^{\mathrm{I}} A_{ij,t-m} + \sum_{k=1}^{q} \gamma_{kt}^{\mathrm{I}}(X_{ik}). \quad (4)$$

We can consider similar models for the death count. At the beginning of the outbreak, infected and death cases could be rare, so 'Poisson' might be a reasonable choice of the random component to describe the distribution of rare events in a large population. As the disease progresses, the variation of infected/death count increases across counties and states. So, at the acceleration phase of the disease, the negative binomial random component might be an appropriate option for the presence of over-dispersion.

The above spatiotemporal epidemic model (STEM) is developed based on the foundation of epidemic modelling. It can provide a rich characterisation of different types of errors for modelling uncertainty. Moreover, it accounts for both spatiotemporal nonstationarity and area-level local features simultaneously. It also offers flexibility in assessing the dynamics of the spread at different times and locations than various parametric models in the literature.

3. Estimation of the STEM

In this section, we describe the estimation method of the parameters and nonparametric components in the proposed STEM model. To capture the temporal dynamics, we consider the moving window approach. From Equations (1) and (2), we can see that the infectious model and the death model share many common features, so we adopt a similar approach when estimating them. Thus, in the following, we only use the infectious model (1) to illustrate the estimation method of the entire STEM system. Moreover, for notation simplicity, we drop the superscripts, such as I and D.

3.1. Penalized quasi-likelihood method

Suppose that $\mathrm{Var}(Y \mid I = w, Z = z, \mathbf{A} = \mathbf{a}, \mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u}) = \sigma^2 V(\mu(w, z, \mathbf{a}, \mathbf{x}, \mathbf{u}))$ for some positive function $V(\cdot)$ and dispersion parameter σ^2, and $L(\mu, y)$ is a quasi-likelihood function satisfying $\nabla_\mu L(\mu, y) = \frac{y-\mu}{\sigma^2 V(\mu)}$. We first describe the estimation of model (1). For the current time t, and roughness parameters λ_0 and λ_1, we consider the penalised quasi-likelihood problem defined as follows:

$$\sum_{i=1}^{n} \sum_{s=t-t_0}^{t} L\left[g^{-1}\left\{ \beta_{0t}(\mathbf{U}_i) + \beta_{1t}(\mathbf{U}_i)\log(I_{i,s-1}) + \alpha_{0t} Z_{i,s-1} + \sum_{j=1}^{p} \alpha_{jt} A_{ij,s-m} \right.\right.$$
$$\left.\left. + \sum_{k=1}^{q} \gamma_k(X_{ik}) \right\}, Y_{is} \right] - \frac{1}{2}\left\{ \lambda_0 \mathcal{E}(\beta_{0t}) + \lambda_1 \mathcal{E}(\beta_{1t}) \right\}, \qquad (5)$$

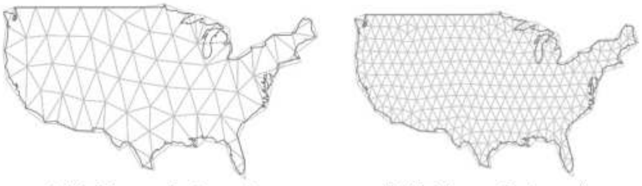
where $t_0 + 1$ is the window width for the model fitting, and it can can be selected by minimising the prediction errors or maximising the correlation between the predicted and observed values. The energy functional is defined as follows:

$$\mathcal{E}(\beta) = \int_\Omega \left\{ (\nabla_{u_1}^2 \beta)^2 + 2(\nabla_{u_1}\nabla_{u_2}\beta)^2 + (\nabla_{u_2}^2 \beta)^2 \right\} \mathrm{d}u_1\, \mathrm{d}u_2, \qquad (6)$$

where $\nabla_{u_j}^q \beta(\boldsymbol{u})$ is the qth order derivative in the direction $u_j, j = 1, 2$, at any location $\boldsymbol{u} = (u_1, u_2)^\top$.

Note that, except for parameters $\{\alpha_{jt}\}_{j=0}^p$, other functions are related to curse of dimensionality due to the nature of functions. To handle this difficulty, we employ the basis expansion approach to approximate the univariate and bivariate functions discussed below. The univariate additive components $\{\gamma_{kt}(\cdot)\}_{k=1}^q$ and the spatially varying coefficient components $\{\beta_{\ell t}(\cdot)\}_{\ell=0}^1$ in model (4) are approximated using univariate polynomial spline and bivariate penalised splines over triangulation (BPST), respectively. The BPST method is well known to be computationally efficient to deal with data distributed on complex domains with irregular shape or with holes inside; see the details in Lai and Wang (2013) and Sangalli, Ramsay, and Ramsay (2013).

Assume that X_k takes value on an interval $[a_k, b_k]$, $k = 1, \ldots, q$. To satisfy the model identification constraint $\mathrm{E}\{\gamma_{kt}(X_k)\} = 0$, in this paper, we consider the space of the centred spline functions $\mathcal{U}_k^0 = \{\phi \in \mathcal{U}_k : \mathrm{E}\phi(X_k) = 0\}$ for univariate additive components, where $\mathcal{U}_k = \mathcal{U}_k^{\varrho}([a_k, b_k])$ be the space of the polynomial splines of order $\varrho + 1$; see Xue and Liang (2010), Liu et al. (2013) and Wang, Xue, and Yang (2020). Let $\mathcal{J}$ be the index set of the basis functions, and then denote by $\{\Phi_{kJ}(x_k), J \in \mathcal{J}\}$ the B-spline basis functions of $\mathcal{U}_k^0$ for the kth covariate. For $J \in \mathcal{J}$ and $k = 1, \ldots, p$, $\Phi_{kJ}(x_k)$ satisfies $\mathrm{E}\Phi_{kJ}(X_k) = 0$ and $\mathrm{E}\Phi_{kJ}^2(X_k) = 1$; see Yu et al. (2020) for the details of basis construction. For all



(a) Triangulation $\triangle_1$ (b) Triangulation $\triangle_2$

Figure 1. Triangulations used in the bivariate spline estimation. (a) triangulation $\triangle_1$ and (b) triangulation $\triangle_2$.

$x_k \in [a_k, b_k]$, the estimator of $\gamma_k(x_k)$ is $\widehat{\gamma}_{kt}(x_k) = \sum_{J\in\mathcal{J}} \xi_{kJ}\Phi_{kJ}(x_k) = \boldsymbol{\Phi}_k^\top(x_k)\boldsymbol{\xi}_{kt}$, where $\boldsymbol{\Phi}_k(x_k) = \{\Phi_{kJ}(x_k), J \in \mathcal{J}\}^\top$, and $\boldsymbol{\xi}_{kt} = \{\xi_{kJt}, J \in \mathcal{J}\}^\top$ is a vector of coefficients.

For the bivariate coefficient functions $\beta_{0t}(\cdot)$ and $\beta_{1t}(\cdot)$ in the STEM model (4), we consider the bivariate spline over triangulation (Lai and Wang 2013). The spatial domain Ω with either an arbitrary shape or holes inside can be partitioned into a set of triangles. Denote $\triangle$ by a triangulation of the domain Ω (Lai and Schumaker 2007); see, for example, Figure 1. In practice, the triangulation can be obtained through varieties of software; see for example, the 'Delaunay' algorithm (*delaunay.m* in MATLAB or *DelaunayTriangulation* in MATHEMATICA), the R package 'Triangulation' (Wang and Lai 2019), and the 'DistMesh' Matlab code.

Let $\mathbb{C}^r(\Omega)$ be the space of rth continuously differentiable functions over the domain Ω. For $0 \le r < d$ and $\triangle$, we construct the spline space of degree d and smoothness r over $\triangle$ in the following:

$$\mathbb{S}_d^r(\triangle) = \{\mathcal{P} \in \mathbb{C}^r(\Omega) : \mathcal{P} \text{ is a polynomial function of degree up to } d \text{ on each } T \in \triangle\}. \tag{7}$$

For triangulation $\triangle$ with M triangles, denote a set of bivariate Bernstein basis polynomials for $\mathbb{S}_d^r(\triangle)$ as $\{B_M(\boldsymbol{u})\}_{M\in\mathcal{M}}$, where an index set $\mathcal{M}$ for basis functions. These bivariate spline basis can be generated via the R package 'BPST' (Wang et al. 2019). More discussions of the bivariate spline over triangulations can be found in Mu, Wang, and Wang (2018) and Yu et al. (2020). Then, we can approximate the bivariate functions $\beta_{\ell t}(\cdot) \in \mathbb{S}_d^r(\triangle)$ in the STEM model (4) by $\sum_{M\in\mathcal{M}} B_M(\boldsymbol{u})\theta_{\ell tM} = \mathbf{B}(\boldsymbol{u})^\top\boldsymbol{\theta}_{\ell t}$, where $\mathbf{B}(\boldsymbol{u}) = \{B_M(\boldsymbol{u}), M \in \mathcal{M}\}^\top$ and $\boldsymbol{\theta}_{\ell t} = \{\theta_{\ell tM}, M \in \mathcal{M}\}^\top$.

Considering the basis expansion, the energy functional $\mathcal{E}(\beta_{\ell t})$ in (6) can be approximated by $\mathcal{E}(\mathbf{B}(\cdot)^\top\boldsymbol{\theta}_\ell) = \boldsymbol{\theta}_\ell^\top\mathbf{P}\boldsymbol{\theta}_\ell$, for $\ell = 0, 1$, where $\mathbf{P}$ is a block diagonal penalty matrix. By introducing the constraint matrix $\mathbf{H}$ which satisfies $\mathbf{H}\boldsymbol{\theta}_\ell = \mathbf{0}$, $\ell = 0, 1$, we can reflect global smoothness in $\mathbb{S}_d^r(\triangle)$ in (7). For the current time t, the maximisation problem (5) is changed to minimise

$$-\sum_{i=1}^{n}\sum_{s=t-t_0}^{t} L\left(g^{-1}\left[\mathbf{B}(\mathbf{U}_i)^\top\{\boldsymbol{\theta}_{0t}+\boldsymbol{\theta}_{1t}\log(I_{i,s-1})\}+\alpha_0 Z_{i,s-1}+\sum_{j=1}^{p}\alpha_{jt}A_{ij,s-m} + \sum_{k=1}^{q}\boldsymbol{\Phi}_k^\top(X_{ik})\boldsymbol{\xi}_{kt}\right], Y_{is}\right)+\frac{1}{2}(\lambda_0\boldsymbol{\theta}_{0t}^\top\mathbf{P}\boldsymbol{\theta}_{0t}+\lambda_1\boldsymbol{\theta}_{1t}^\top\mathbf{P}\boldsymbol{\theta}_{1t}) \quad \text{subject to } \mathbf{H}\boldsymbol{\theta}_{\ell t}=\mathbf{0}, \ell=0,1. \tag{8}$$

Directly solving the optimisation problem in (8) is not straightforward due to the smoothness constraints involved. Instead, suppose that the rank r_H matrix $\mathbf{H}^\top$ is decomposed into $\mathbf{QR} = (\mathbf{Q}_1\ \mathbf{Q}_2)\binom{\mathbf{R}_1}{\mathbf{R}_2}$, where $\mathbf{Q}_1$ is the first r_H columns of an orthogonal matrix $\mathbf{Q}$, and $\mathbf{R}_2$ is a matrix of zeros, which is a submatrix of an upper triangle matrix $\mathbf{R}$. Then, reparametrization of $\boldsymbol{\theta}_{\ell t} = \mathbf{Q}_2\boldsymbol{\theta}^*_{\ell t}$ for some $\boldsymbol{\theta}^*_{\ell t}, \ell = 0, 1$, enforces $\mathbf{H}\boldsymbol{\theta}_{\ell t} = \mathbf{0}$. Thus, the constraint problem in (8) can be changed to an unconstrained optimisation problem as follows:

$$-\sum_{i=1}^{n}\sum_{s=t-t_0}^{t} L\left(g^{-1}\left[\mathbf{B}(\mathbf{U}_i)^\top\mathbf{Q}_2\{\boldsymbol{\theta}^*_{0t} + \boldsymbol{\theta}^*_{1t}\log(I_{i,s-1})\} + \alpha_{0t}Z_{i,s-1} + \sum_{j=1}^{p}\alpha_{jt}A_{ij,s-m}\right.\right.$$
$$\left.\left. + \sum_{k=1}^{q}\boldsymbol{\Phi}_k^\top(X_{ik})\boldsymbol{\xi}_{kt}\right], Y_{is}\right) + \frac{1}{2}\left(\lambda_0\boldsymbol{\theta}^{*\top}_{0t}\mathbf{Q}_2^\top\mathbf{P}\mathbf{Q}_2\boldsymbol{\theta}^*_{0t} + \lambda_1\boldsymbol{\theta}^{*\top}_{1t}\mathbf{Q}_2^\top\mathbf{P}\mathbf{Q}_2\boldsymbol{\theta}^*_{1t}\right). \qquad (9)$$

Let $(\widehat{\boldsymbol{\theta}}^*_{0t}, \widehat{\boldsymbol{\theta}}^*_{1t})^\top$, $(\widehat{\alpha}_{0t}, \widehat{\alpha}_{1t}, \ldots, \widehat{\alpha}_{pt})^\top$, and $(\widehat{\boldsymbol{\xi}}_{1t}, \ldots, \widehat{\boldsymbol{\xi}}_{qt})^\top$ be the minimisers of (9) at time point t. Then, the estimators of $\boldsymbol{\beta}_{\ell t}(\cdot)$ are $\widehat{\boldsymbol{\beta}}_{\ell t}(u) = \mathbf{B}(u)^\top\mathbf{Q}_2\widehat{\boldsymbol{\theta}}^*_{\ell t}$, $\ell = 0, 1$, the estimators of α_{jt} are $\widehat{\alpha}_{jt}$, $j = 1, \ldots, p$, and the spline estimators of $\gamma_{kt}(\cdot)$ are $\widehat{\gamma}_{kt}(x_k) = \boldsymbol{\Phi}_k(x_k)^\top\widehat{\boldsymbol{\xi}}_{kt}$, $k = 1, \ldots, q$.

3.2. Penalized iteratively reweighted least squares algorithm

In this subsection, we are going to describe the estimating algorithm in detail. For the current time t, let $\mathbf{Y} = (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_t^\top)^\top$ be the vector of the response variable where $\mathbf{Y}_s = (Y_{1s}, \ldots, Y_{ns})^\top$. Denote $\boldsymbol{\Phi}_i^\top = \{\boldsymbol{\Phi}_1(X_{i1})^\top, \ldots, \boldsymbol{\Phi}_q(X_{iq})^\top\}$, $\mathbf{A}_{is}^\top = (A_{i1,s-m}, \ldots, A_{ip,s-m})$, and $\mathbf{F} = (\mathbf{F}_1, \ldots, \mathbf{F}_t)^\top$, where $\mathbf{F}_s = (\mathbf{F}_{1s}, \ldots, \mathbf{F}_{ns})$, and $\mathbf{F}_{is}^\top = (\mathbf{A}_{is}^\top,\ \boldsymbol{\Phi}_i^\top,\ [\{1, \log(I_{i,s-1})\}^\top \otimes \{\mathbf{Q}_2^\top\mathbf{B}(\mathbf{U}_i)\}]^\top)$. Denote $\boldsymbol{\vartheta}_t = (\boldsymbol{\alpha}_t^\top, \boldsymbol{\xi}_t^\top, \boldsymbol{\theta}_t^{*\top})^\top$, and let $\eta_{is}(\boldsymbol{\vartheta}_t) = \mathbf{B}(\mathbf{U}_i)^\top\mathbf{Q}_2\{\boldsymbol{\theta}^*_{0t} + \boldsymbol{\theta}^*_{1t}\log(I_{i,s-1})\} + \alpha_{0_t}Z_{i,s-1} + \sum_{j=1}^{p}\alpha_{jt}A_{ij,s-m} + \sum_{k=1}^{q}\boldsymbol{\Phi}_{kt}^\top (X_{ik})\boldsymbol{\xi}_k$, and $\boldsymbol{\eta}(\boldsymbol{\vartheta}_t) = \{\eta_{is}(\boldsymbol{\vartheta}_t)\}_{i=1,s=1}^{n,t}$. In addition, let the mean vector $\boldsymbol{\mu}(\boldsymbol{\vartheta}_t) = \{\mu_{is}(\boldsymbol{\vartheta}_t)\}_{i,s=1}^{n,t} = \{g^{-1}(\eta_{is}(\boldsymbol{\vartheta}_t))\}_{i,s=1}^{n,t}$, the variance function matrix $\mathbf{V} = \text{diag}\{V(\mu_{is})\}_{i,s=1}^{n,t}$, the diagonal matrix $\mathbf{G} = \text{diag}\{g'(\mu_{is})\}_{i,s=1}^{n,t}$ with the derivative of link function as element, and the weight matrix $\widetilde{\mathbf{V}} = \text{diag}[\{V(\mu_{is})g'(\mu_{is})^2\}^{-1}w_{st}, i = 1, \ldots, n, s = 1, \ldots, t]$, where $w_{st} = I(t - s \geq t_0)$.

Iteratively reweighted least squares algorithm (IRLS) is commonly used to find the maximum likelihood estimates of a generalised linear model. Therefore, in this work, we design a penalised iteratively reweighted least squares (PIRLS) algorithm as described below. Suppose at the jth iteration, we have $\boldsymbol{\mu}^{(j)} = \boldsymbol{\mu}(\boldsymbol{\vartheta}^{(j)})$, $\boldsymbol{\eta}_t^{(j)} = \boldsymbol{\eta}(\boldsymbol{\vartheta}_t^{(j)})$ and $\mathbf{V}^{(j)}$. Then at $(j+1)$th iteration, we consider the following objective function:

$$L_P^{(j+1)} = \left\|\left\{\mathbf{V}^{(j)}\right\}^{-1/2}\left\{\mathbf{Y} - \boldsymbol{\mu}\left(\boldsymbol{\vartheta}_t^{(j)}\right)\right\}\right\|^2 + \frac{1}{2}\sum_{\ell=0}^{1}\lambda_\ell\boldsymbol{\theta}^{*\top}_{\ell t}\mathbf{Q}_2^\top\mathbf{P}\mathbf{Q}_2\boldsymbol{\theta}^*_{\ell t}.$$

Algorithm 1 The Penalized Iteratively Reweighted Least Squares (PIRLS) Algorithm.

```
Step 1. Initialize $\eta^{(0)}$ and $\mu^{(0)}$ and calculate $\widetilde{\mathbf{V}}^{(0)}$ and $\widetilde{\mathbf{Y}}^{(0)}$ from $g'(\mu_{is}^{(0)})$ and $V(\mu_{is}^{(0)})$, $i = 1, \ldots, n$, and $s = 1, \ldots, t$.
Step 2. Set step $j = 0$.
    while $\{\alpha, \xi, \theta^*\}$ not converge do
        (i) Obtain $\alpha^{(j+1)}, \xi^{(j+1)}, \theta^{*(j+1)}$ by minimising the (10) with respect to $\vartheta$, and $\eta^{(j+1)} = \eta(\vartheta^{(j+1)})$ and $\mu^{(j+1)} = \mu(\vartheta^{(j+1)})$.
        (ii) Update $\widetilde{\mathbf{V}}^{(j+1)}$ and $\widetilde{\mathbf{Y}}^{(j+1)}$ with $g'(\mu_{is}^{(j+1)})$ and $V(\mu_{is}^{(j+1)})$, $i = 1, \ldots, n$, $s = 1, \ldots, t$, using $\eta^{(j+1)}$ and $\mu^{(j+1)}$.
        (iii) Set $j = j + 1$.
```

Take the first order Taylor expansion of $\boldsymbol{\mu}(\boldsymbol{\vartheta})$ around $(\boldsymbol{\vartheta}^{(j)})$, then

$$L_P^{(j+1)} \approx \left\| \left\{ \mathbf{V}^{(j)} \right\}^{-1/2} \left[\mathbf{Y} - \boldsymbol{\mu}^{(j)} - \{\mathbf{G}^{(j)}\}^{-1} \mathbf{F}(\boldsymbol{\vartheta}_t - \boldsymbol{\vartheta}_t^{(j)}) \right] \right\|^2 + \frac{1}{2} \sum_{\ell=0}^{1} \lambda_\ell \boldsymbol{\theta}_{\ell t}^{*\top} \mathbf{Q}_2^\top \mathbf{P} \mathbf{Q}_2 \boldsymbol{\theta}_{\ell t}^*$$

$$= \left\| \left\{ \widetilde{\mathbf{V}}^{(j)} \right\}^{1/2} \left[\widetilde{\mathbf{Y}}^{(j)} - \mathbf{F}(\boldsymbol{\vartheta}_t) \right] \right\|^2 + \frac{1}{2} \sum_{\ell=0}^{1} \lambda_\ell \boldsymbol{\theta}_{\ell t}^{*\top} \mathbf{Q}_2^\top \mathbf{P} \mathbf{Q}_2 \boldsymbol{\theta}_{\ell t}^*, \tag{10}$$

where $\widetilde{\mathbf{Y}}^{(j)} = (\widetilde{\mathbf{Y}}_1^{(j)\top}, \ldots, \widetilde{\mathbf{Y}}_t^{(j)\top})^\top$ with $\widetilde{Y}_{is}^{(j)} = g'(\mu_{is}^{(j)})(Y_{is} - \mu_{is}^{(j)}) + \eta_{is}^{(j)}$ for $s = 1, \ldots, t$. The detailed procedure for the PIRLS is illustrated in Algorithm 1. In the numerical studies, we consider the following initial values $\mu_{is}^{(0)} = Y_{is} + 0.1$ and $\eta_{is}^{(0)} = g(\mu_{is}^{(0)})$ for $i = 1, \ldots, n$ and $s = 1, \ldots, t$.

Compared with the traditional nonparametric techniques, such as kernel smoothing, the proposed algorithm is much more computationally efficient. Therefore, we can easily apply our method to analyse massive spatiotemporal data sets.

3.3. Asymptotic results

Let $\eta_{it}^0 = g(\mu_{it}^0)$, where μ_{it}^0 is the conditional mean based on the true parameter α_{jt}^0's and functions $\beta_{\ell t}^0$'s and γ_{kt}^0's. Let $\varepsilon_{it} = Y_{it} - g^{-1}(\eta_{it}^0)$ be the error term. For the quasi-likelihood function, $L\{g^{-1}(\eta), y\}$, denote $q_1(\eta, y) = \frac{\partial}{\partial \eta} L\{g^{-1}(\eta), y\} = \{y - g^{-1}(\eta)\}\rho_1(\eta)$, and $q_2(\eta, y) = \frac{\partial^2}{\partial \eta^2} L\{g^{-1}(\eta), y\} = \{y - g^{-1}(\eta)\}\rho_1'(\eta) - \rho_2(\eta)$, where $\rho_j(\eta) = \{\frac{\partial}{\partial \eta} g^{-1}(\eta)\}^j / [\sigma^2 V\{g^{-1}(\eta)\}] = [\{g'(g^{-1}(\eta)\}^j \sigma^2 V\{g^{-1}(\eta)\}]^{-1}, j = 1, 2$. For a vector valued function $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_p)^\top$, let $\|\boldsymbol{\phi}\|_{L_2} = \{\sum_{k=0}^p \|\phi_k\|_{L_2}^2\}^{1/2}$ and $\|\boldsymbol{\phi}\|_\infty = \max_{0 \le k \le p} \|\phi_k\|_\infty$. Define $|\phi|_{\upsilon,\infty} = \max_{i+j=\upsilon} \|\nabla_{u_1}^i \nabla_{u_2}^j \phi\|_\infty$ for a nonnegative integer ν.

The following theorem provides the L_2 convergence rate of the spline estimators, $\widehat{\beta}_{\ell t}(\boldsymbol{u})$, for $\ell = 0, 1$. The detailed proof is illustrated in the Supplemental Material B.

Theorem 3.1: *Under Assumptions* (A1)–(A7) *in the Supplemental Material B, the univariate spline estimators* $\widehat{\gamma}_{kt}$, $k = 1, \ldots, q$, *and the bivariate spline estimators* $\widehat{\beta}_{\ell t}(\cdot)$, $\ell = 0, 1$, *satisfy that* $\sum_{\ell=0}^1 \|\widehat{\beta}_{\ell t} - \beta_{\ell t}^0\|_{L_2} + \sum_{k=1}^q \|\widehat{\gamma}_{kt} - \gamma_{kt}^0\|_{L_2} = O_{\text{a.s.}}\{(h^{-1/2} + |\Delta|^{-1}) n^{-1/2} (\log n)^{1/2} + h^{\varrho+1} + |\Delta|^{d+1} + \lambda_{\max}(n|\Delta|^4)^{-1}\}$ *as* $n \to \infty$, *where* $\lambda_{\max} = \max(\lambda_0, \lambda_1)$.

Let $\mathbf{W}^\top = \{1, \log(I_{s-1})\}$, $\mathbf{Z}^\top = \{Z_{s-1}, A_{1,s-m}, \ldots, A_{p,s-m}\}$, and $\eta^0(\mathbf{X}, \mathbf{Z}, \mathbf{U}, \mathbf{W}) = \mathbf{Z}^\top \boldsymbol{\alpha}^0 + \sum_{k=1}^{q} \gamma_k^0(X_k) + \sum_{\ell=0}^{1} \beta_\ell^0(\mathbf{U}) W_{\ell s}$. In the following, define

$$\Gamma(\mathbf{x}, \mathbf{u}, \mathbf{w}) = \frac{\mathrm{E}\left[\rho_2\{\eta^0(\mathbf{Z}, \mathbf{X}, \mathbf{U}, \mathbf{W})\}\mathbf{Z} \mid \mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u}, \mathbf{W} = \mathbf{w}\right]}{\mathrm{E}[\rho_2\{\eta^0(\mathbf{Z}, \mathbf{X}, \mathbf{U}, \mathbf{W})\} \mid \mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u}, \mathbf{W} = \mathbf{w}]}, \quad \widetilde{\mathbf{Z}} = \mathbf{Z} - \boldsymbol{\Gamma}(\mathbf{X}, \mathbf{U}, \mathbf{W}). \tag{11}$$

The next theorem shows that the maximum quasi-likelihood estimator of $\boldsymbol{\alpha}^0$ is root-n consistent and asymptotically normal.

Theorem 3.2: *Under Assumptions* (A1)–(A8) *in the Supplemental Material B, the estimator $\widehat{\boldsymbol{\alpha}}_t$ is asymptotically normally distributed, i.e.* $\sqrt{n}(\widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^0) \to N(\mathbf{0}, \boldsymbol{\Sigma}^{-1})$, *where* $\boldsymbol{\Sigma} = \mathrm{E}\{\rho_2(\eta^0)\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top\}$ *with $\widetilde{\mathbf{Z}}$ given in* (11).

3.4. Modeling the number of fatal and recovered cases

To fit the proposed STEM and make predictions for cumulative positive cases, one obstacle is the lack of direct observations for the number of active cases, I_{it}. Instead, the most commonly reported number is the count of total confirmed cases, C_{it}. Some departments of public health also release information about fatal cases D_{it} and recovered cases R_{it}. Based on the fact that $I_{it} = C_{it} - R_{it} - D_{it}$, we attempt to model D_{it} and R_{it} to facilitate the estimation and prediction of newly confirmed cases Y_{it} based on the proposed STEM model. For the death model (2), we can use the same maximum quasi-likelihood approach to fit the model.

Ideally, if sufficient data for recovered cases can be collected from each area, a similar model can be fitted to explain the growth of the recovered cases. However, most people who become sick with COVID-19 only experience mild illness and can recover at home without medical care within a week. Meanwhile, according to the U.S. Centers for Disease Control and Prevention, severe cases are often hospitalised to receive supportive care, which may take several weeks. Although there have been regional, national, and global data on confirmed cases and deaths, not much has been reported on recovery. Currently, there is a lack of a uniform method for reporting recoveries across the U.S. (Howard and Yu 2020).

Only a few states regularly update the number of recovered patients, but the counts can seldom be mapped to counties. Due to a lack of data, we are no longer able to use all the explanatory variables discussed above to model daily new recovered cases. Instead, we mimic the relationship between the number of recovered and active cases from some compartmental models in epidemiology (Siettos and Russo 2013; Anastassopoulou, Russo, Tsakris, and Siettos 2020). At current time point t, we assume that $\Delta R_{is} = R_{is} - R_{i,s-1} = \nu_t^{\mathrm{R}} I_{i,s-\delta'} + \varepsilon_{is}$, $s = t - t_0, \ldots, t$, in which δ' represents the time delay from infection to recovery ($\delta' = 10$ in our analysis), and ε_{is} is the random noise. The recovery rate ν_t^{R} enables us to make reasonable predictions for future recovered patients counts and provide researchers with the foresight of when the epidemic will end. The rate ν_t^{R} can be either estimated from available state-level data, or obtained from prior medical studies to alleviate the under-reporting issue in actual data.

3.5. Zero-inflated models at the early stage of the outbreak

It is well known that in the early stage of an epidemic, the quality of any model output can be affected by the restricted quality of data that pertain to under-detection or inconsistent detection of cases, reporting delays, and poor documentation, regarding infections, deaths, tests, and other factors. There are many counties with zero daily counts at the early stage of disease spread. Therefore, we consider zero-inflated models based on a zero-inflated probability distribution, i.e. a distribution that allows for frequent zero-valued observations. Following the previous works (Arab, Holan, Wikle, and Wildhaber 2012; Wood, Pya, and Säfken 2016), we assume the observed counts Y_{it} contributes to a zero-inflated Poisson (ZIP) distribution, $\mathrm{ZIP}(\mu_{it}^{\mathrm{I}}, p_{it}^{\mathrm{I}})$, specifically, we assume that

$$P(Y_{it} = y \mid I_{i,t-1}, \mathbf{Z}_{i,t-1}, \mathbf{A}_{i,t-m}, \mathbf{X}_i, \mathbf{U}_i) = \begin{cases} 1 - p_{it}^{\mathrm{I}}, & y = 0, \\ p_{it}^{\mathrm{I}} \dfrac{(\mu_{it}^{\mathrm{I}})^y}{\{\exp(\mu_{it}^{\mathrm{I}}) - 1\} y!}, & y > 0, \end{cases}$$

where $p_{it}^{\mathrm{I}} = \frac{\exp(\eta_{it}^{\mathrm{I}})}{1+\exp(\eta_{it}^{\mathrm{I}})}$ with $\eta_{it}^{\mathrm{I}} = a_1 + \exp(a_2)\log(\mu_{it}^{\mathrm{I}})$, μ_{it}^{I} is generated from (4), and a_1, a_2 are unknown parameters estimated along with the roughness parameters. See Wood et al. (2016) for the estimation of a_1 and a_2.

Let $\Delta D_{it} = D_{it} - D_{i,t-1}$ be the number of new fatal cases on day t. Similarly, we can consider zero-inflated models for fatal cases, in which we assume the observed count ΔD_{it} contributes to a ZIP distribution $\mathrm{ZIP}(\mu_{it}^{\mathrm{D}}, p_{it}^{\mathrm{D}})$:

$$P(\Delta D_{it} = d \mid I_{i,t-1}, \mathbf{A}_{i,t-m}, \mathbf{X}_i, \mathbf{U}_i) = \begin{cases} 1 - p_{it}^{\mathrm{D}}, & d = 0, \\ p_{it}^{\mathrm{D}} \dfrac{(\mu_{it}^{\mathrm{D}})^d}{\{\exp(\mu_{it}^{\mathrm{D}}) - 1\} d!}, & d > 0, \end{cases}$$

where $p_{it}^{\mathrm{D}} = \frac{\exp(\eta_{it}^{\mathrm{D}})}{1+\exp(\eta_{it}^{\mathrm{D}})}$ with $\eta_{it}^{\mathrm{D}} = \nu_1 + \exp(\nu_2)\log(\mu_{it}^{\mathrm{D}})$, μ_{it}^{D} is generated from (2), and ν_1, ν_2 are unknown parameters that can be similarly estimated as in the above.

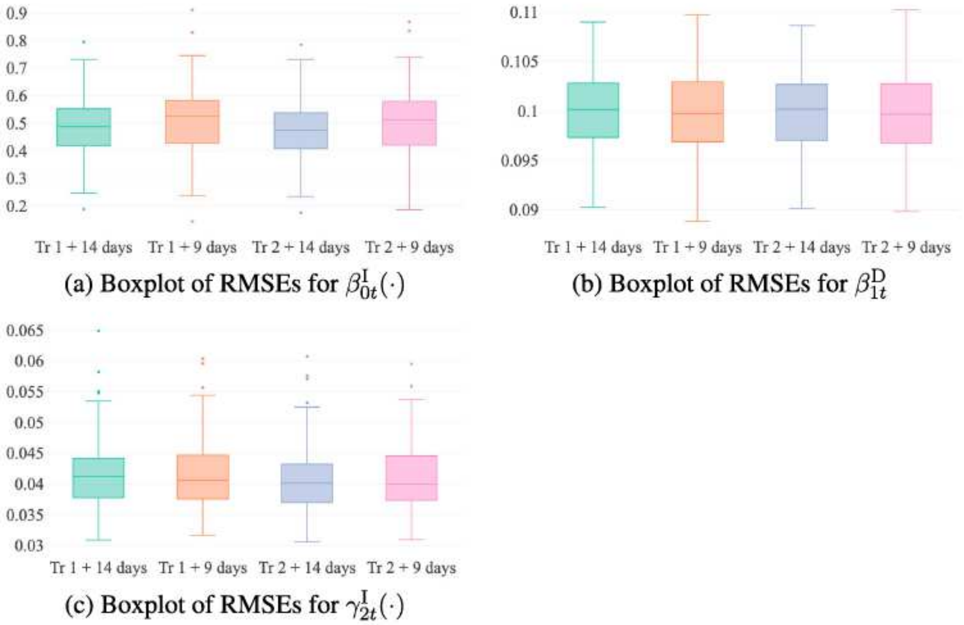
4. A simulation study

In this section, we conduct a simulation study to evaluate the finite sample performance of the proposed method. In the simulation, we use a subset of covariates of the county-level characteristics analysed in Section 5. The response variable Y_{it} and ΔD_{it} are generated from a ZIP distribution with the logarithm of Poisson parameters generated as following:

$$\log(\mu_{it}^{\mathrm{I}}) = \beta_{0t}^{\mathrm{I}}(\mathbf{U}_i) + \beta_{1t}^{\mathrm{I}}(\mathbf{U}_i)\log(I_{i,t-1}) + \sum_{j=1}^{p} \alpha_{jt}^{\mathrm{I}} A_{ij,t-m} + \sum_{k=1}^{q} \gamma_{kt}^{\mathrm{I}}(X_{ik}), \tag{12}$$

$$\log(\mu_{it}^{\mathrm{D}}) = \beta_{0t}^{\mathrm{D}}(\mathbf{U}_i) + \beta_{1t}^{\mathrm{D}}\log(I_{i,t-\delta}), \tag{13}$$

where $\delta = 14$, $p = 2$, $q = 5$, $m = 7$, and A_{ijt}, X_{ik}, $j = 1, 2$, $k = 1, \ldots, 5$, come from the covariates in the COVID-19 dataset described in Section 5. The true univariate functions $\gamma_{1t}(x)$, $\gamma_{2t}(x)$, …, $\gamma_{5t}(x)$, together with their estimate and confidence band in one typical iteration, are displayed in Figure 3(a–e). Figure 4(a–c) depict the bivariate coefficient



(a) Boxplot of RMSEs for $\beta^{\mathrm{I}}_{0t}(\cdot)$

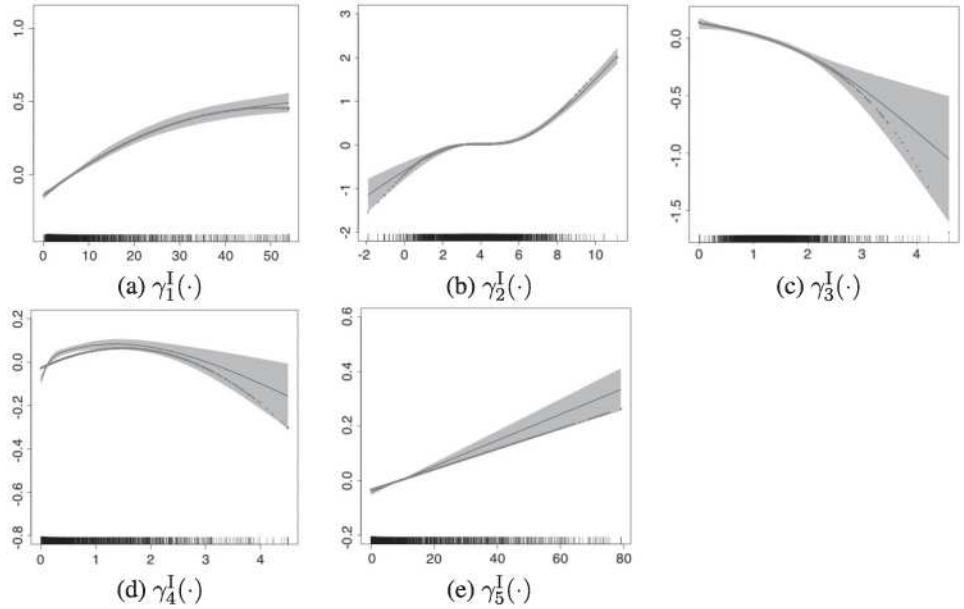
(b) Boxplot of RMSEs for β^{D}_{1t}

(c) Boxplot of RMSEs for $\gamma^{\mathrm{I}}_{2t}(\cdot)$

Figure 2. Boxplot of RMSEs. (a) Boxplot of RMSEs for $\beta^{\mathrm{I}}_{0t}(\cdot)$. (b) Boxplot of RMSEs for β^{D}_{1t} and (c) Boxplot of RMSEs for $\gamma^{\mathrm{I}}_{2t}(\cdot)$.

functions $\beta^{\mathrm{I}}_{0t}(\cdot)$, $\beta^{\mathrm{I}}_{1t}(\cdot)$ and $\beta^{\mathrm{D}}_{0t}(\cdot)$, which are generated to mimic the spatial pattern of infection/mortality rate in the pandemic. We also include the corresponding estimated functions from one experiment in Figure 4(d–f) to show that the spatial pattern can be very well captured using the proposed method. For recovery data, the daily recovered cases are simulated by $\Delta R_{it} = \nu^{\mathrm{R}} I_{i,s-1}$, where $\nu^{\mathrm{R}} = 0.07$. We simulate data by assuming that a pandemic emerged on March 15 with 1 case showed up in each of the 420 selected counties. These counties are selected if COVID-19 cases had been found by March 15 in real data. Then, daily confirmed, fatal and recovered cases are generated based on model (12) and (13) from a ZIP distribution with the complimentary log of the zero probability being linearly dependent on the log of the Poisson parameter μ^{I}_{it} and μ^{D}_{it}.

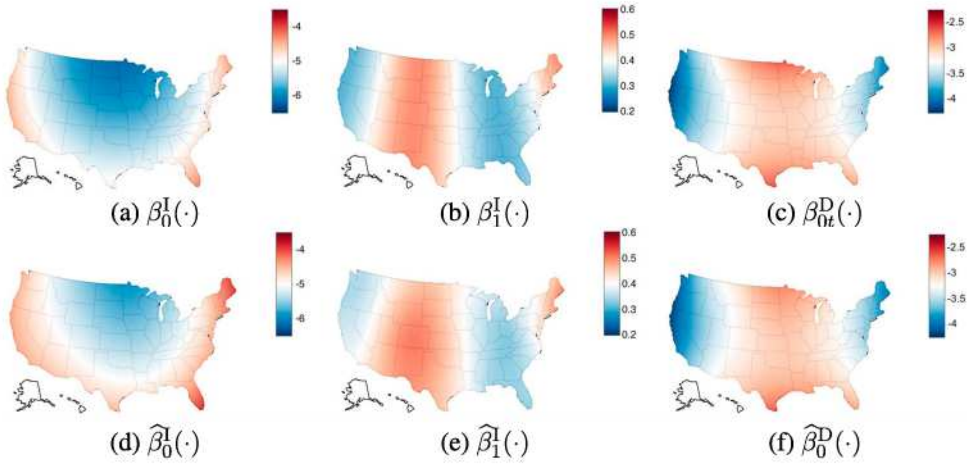
To evaluate the performance numerically, we conduct 100 Monte Carlo experiments with 9 or 14 days as the training window sizes. For the univariate spline smoothing, we use cubic splines with two interior knots; and for the bivariate spline smoothing, we consider degree $d = 2$, smoothness $r = 1$, and two different triangulations in Figure 1: $\triangle_1$ (119 triangles with 87 vertices) and $\triangle_2$ (522 triangles with 306 vertices). The root mean squared errors (RMSEs) for some of the parametric and nonparametric components in models (12) and (13) are illustrated in Figure 2, and the boxplots of the RMSEs for all the parametric and nonparametric components are shown in Supplemental Material A. Moreover, the average RMSEs over 100 experiments are reported in Table 1. According to the numeric results, the proposed model is not sensitive to the choice of triangulation. Based on Figure 2 and Table 1, one can see that increasing the window size of training data can help improve the accuracy in estimating most of the coefficient functions while increasing the computational burden at the same time. Thus, in practice, users can balance the choice of the window size with the power of the computation resource.



(a) $\gamma_1^{\mathrm{I}}(\cdot)$ (b) $\gamma_2^{\mathrm{I}}(\cdot)$ (c) $\gamma_3^{\mathrm{I}}(\cdot)$

(d) $\gamma_4^{\mathrm{I}}(\cdot)$ (e) $\gamma_5^{\mathrm{I}}(\cdot)$

Figure 3. The true and estimated univariate component functions in model (12) (estimation window: 04/17/20–04/30/20). The red curves represent true functions, while black curves and dark area indicate estimated coefficients and their 95% confidence bands. (a) $\gamma_1^{\mathrm{I}}(\cdot)$. (b) $\gamma_2^{\mathrm{I}}(\cdot)$. (c) $\gamma_3^{\mathrm{I}}(\cdot)$. (d) $\gamma_4^{\mathrm{I}}(\cdot)$ and (e) $\gamma_5^{\mathrm{I}}(\cdot)$.



(a) $\beta_0^{\mathrm{I}}(\cdot)$ (b) $\beta_1^{\mathrm{I}}(\cdot)$ (c) $\beta_{0t}^{\mathrm{D}}(\cdot)$

(d) $\widehat{\beta}_0^{\mathrm{I}}(\cdot)$ (e) $\widehat{\beta}_1^{\mathrm{I}}(\cdot)$ (f) $\widehat{\beta}_0^{\mathrm{D}}(\cdot)$

Figure 4. True and estimated bivariate varying coefficients (estimation window: 04/17/20–04/30/20) in the simulation. (a) $\beta_0^{\mathrm{I}}(\cdot)$. (b) $\beta_1^{\mathrm{I}}(\cdot)$. (c) $\beta_{0t}^{\mathrm{D}}(\cdot)$. (d) $\widehat{\beta}_0^{\mathrm{I}}(\cdot)$. (e) $\widehat{\beta}_1^{\mathrm{I}}(\cdot)$ and (f) $\widehat{\beta}_0^{\mathrm{D}}(\cdot)$.

5. COVID-19 case study

The goals of the following study are two-fold. First, we develop a new dynamic epidemic modelling framework for public health surveillance data to study the spatiotemporal pattern in the spread of COVID-19. We aim to investigate whether the proposed model

Table 1. The average of root mean squared errors (RMSEs) of the estimated components in infection model and death model in the simulation ('–' indicates not applicable).

	Window			Component							
Model	Δ	Size	β_0	β_1	α_1	α_2	γ_1	γ_2	γ_3	γ_4	γ_5
Infection	Δ_1	9 days	0.5094	0.0453	0.0534	0.0258	0.0105	0.0415	0.0214	0.0262	0.0114
	Δ_1	14 days	0.4862	0.0395	0.0499	0.0193	0.0103	0.0416	0.0186	0.0256	0.0119
	Δ_2	9 days	0.5038	0.0437	0.0521	0.0253	0.0110	0.0409	0.0216	0.0261	0.0112
	Δ_2	14 days	0.4729	0.0374	0.0479	0.0190	0.0109	0.0406	0.0195	0.0255	0.0120
Death	Δ_1	9 days	0.0456	0.0999	–	–	–	–	–	–	–
	Δ_1	14 days	0.0373	0.1000	–	–	–	–	–	–	–
	Δ_2	9 days	0.0459	0.0999	–	–	–	–	–	–	–
	Δ_2	14 days	0.0375	0.1000	–	–	–	–	–	–	–

could guide the modelling of the dynamics of the spread at the county level by moving beyond the typical theoretical conceptualisation of context where a county's infection is only associated with its own features. Second, to understand the factors that contribute to the spread of COVID-19, we model the daily infected cases at the county level, considering the demographic, environmental, behavioural, and socioeconomic factors in the U.S.

5.1. Data description

The data for the COVID-19 outbreak in the U.S. is collected and cleaned from a combination of public data repositories, including official state Health Department Websites, the New York Times (NYT 2020), the COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University (JHU CSSE 2020), the COVID Tracking Project (Atlantic 2020) and the USAFacts (USAFacts 2020). Wang et al. (2020) provides a thorough comparison of the COVID-19 data collected from the above four sources, details on anomaly detection and repair, as well as how to integrate the data with other local characteristics.

The USA Counties Database compiled by the U.S. Census Bureau and the Homeland Infrastructure Foundation-level Data prepared by the U.S. Department of Homeland Security (see Table 2) were used as the source of local information. The county-level features can be categorised into the following groups.

Control measures. We mainly consider two control measures in our work, emergency declarations and 'stay-at-home' or 'shelter-in-place' orders, among a variety of social distancing policies (e.g. school closures, closures of non-essential services focussed on bars and restaurants, bans on large gatherings and the deployment of severe travel restrictions). Dates of interventions were compiled by checking national and state government websites, executive orders, and newly-initiated COVID-19 laws. Starting in Washington on February 29, 2020, the declarations of state emergency soon swept the nation. By March 16, 2020, every state had made an emergency declaration, with most taking the form of a State of Emergency or a Public Health Emergency. The executive orders of 'stay-at-home' or 'shelter-in-place' started in California in the middle of March, and within one to two weeks, the majority of the states had taken similar actions. Due to the immense pressures of the crippled economy and anxious public, states in the U.S. started to reopen successively in late April. A state is treated as 'reopening' once its stay-at-home order lifts,

Table 2. County-level predictors used in the STEM modeling.

Predictors	Description
Control	Dummy variable for declaration of 'shelter-in-place' or 'stay-at-home' order (Control $= 1$, for 'shelter-in-place', and Control $= 0$, for no restriction or restriction lifted)
Socioeconomic Status	
Affluence	Social affluence
Disadvantage	Concentrated disadvantage
Gini	Gini coefficient
Healthcare Infrastructure	
NHIC	Percent of persons under 65 years without health insurance
EHPC	Local government expenditures for health per capita
TBed*	Total bed counts per 1000 population
Demographic Characteristics	
AA	Percent of African American population
HL	Percent of Hispanic or Latino population
PD*	Population density per square mile of land area
Old	Aged people (age $\geq$ 65 years) rate per capita
Sex	Ratio of male over female
Environment Characteristics	
Mobility	Daily number of trips within each county
Urban	Urban rate

Note: The covariates with * represent that they are transformed from the original value by $f(x) = \log(x + \delta)$. For example, $\mathrm{PD}^* = \log(\mathrm{PD} + \delta)$, where δ is a small number.

or once reopening is permitted in at least one primary sector (restaurants, retail stores, personal care businesses), or once reopening is permitted in a combination of smaller sectors.

Socioeconomic status contains (a) social affluence (b) concentrated disadvantage and (c) Gini coefficient. Social affluence is a measure of more economically privileged areas, including factors: (i) percent of households with income over $75,000; (ii) percent of adults obtaining bachelor's degree or higher; (iii) percent of employed persons in management, professional and related occupations; (iv) median value of owner-occupied housing units. The concentrated disadvantage is a measure for conditions of economic disadvantage, including factors: (i) percent of households with public assistance income; (ii) percent of households with a female householder and no husband present; (iii) civilian labour force unemployment rate. Gini coefficient, known as the Gini index, is a measure of economic inequality and wealth distribution among a population.

Healthcare infrastructure contains (d) local government expenditures for health per capita, (e) percent of persons under 65 years without health insurance, and (f) logarithm of total bed counts per 1000 population.

Demographic characteristics contain (g) percent of African American population, (h) percent of Hispanic or Latino population, (i) logarithm of population density per square mile of land area, (j) aged people (age $\geq$ 65 years) rate per capita, (k) ratio of male over female, and (l) urban rate.

Mobility data are collected and cleaned from the U.S. Department of Transportation, Bureau of Transportation Statistics, and Descartes Labs. It describes the daily number of trips within each county produced from an anonymized national panel of mobile device data from multiple sources. Trips are defined as movements that include a stay of longer than 10 min at an anonymized location away from home.

5.2. Analysis and findings in COVID-19

For the model estimation, we consider the following model for the infected count:

$$
\begin{aligned}
&\log(\mu_{it}^{\mathrm{I}}) \\
&\quad = \beta_{0t}^{\mathrm{I}}(\mathbf{U}_i) + \beta_{1t}^{\mathrm{I}}(\mathbf{U}_i)\log(I_{i,t-1}) + \alpha_{0t}^{\mathrm{I}} Z_{i,t-1} + \alpha_{1t}^{\mathrm{I}}\mathrm{Control}_{i,t-7} + \alpha_{2t}^{\mathrm{I}}\mathrm{Mobility}_{i,t-7} \\
&\qquad + \gamma_{1t}^{\mathrm{I}}(\mathrm{Gini}_i) + \gamma_{2t}^{\mathrm{I}}(\mathrm{Affluence}_i) + \gamma_{3t}^{\mathrm{I}}(\mathrm{Disadvantage}_i) + \gamma_{4t}^{\mathrm{I}}(\mathrm{Urban}_i) + \gamma_{5t}^{\mathrm{I}}(\mathrm{PD}_i) \\
&\qquad + \gamma_{6t}^{\mathrm{I}}(\mathrm{Tbed}_i) + \gamma_{7t}^{\mathrm{I}}(\mathrm{NHIC}_i) + \gamma_{8t}^{\mathrm{I}}(\mathrm{EHPC}_i) \\
&\qquad + \gamma_{9t}^{\mathrm{I}}(\mathrm{AA}_i) + \gamma_{10t}^{\mathrm{I}}(\mathrm{HL}_i) + \gamma_{11t}^{\mathrm{I}}(\mathrm{Sex}_i) + \gamma_{12t}^{\mathrm{I}}(\mathrm{Old}_i), \qquad (14)
\end{aligned}
$$

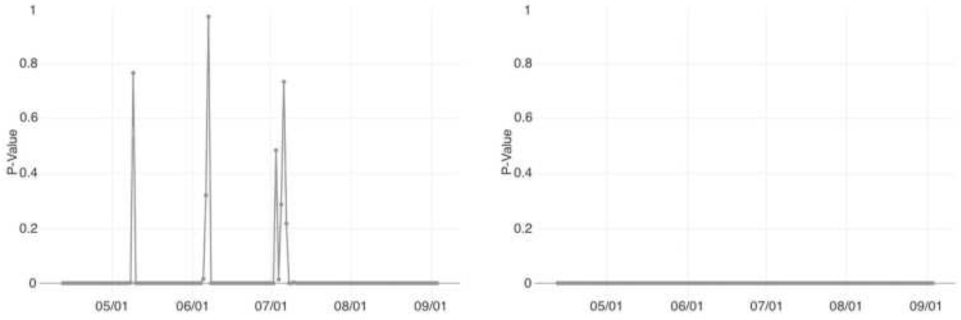
where $i = 1, \ldots, 3104$. For the death count, we consider the following semiparametric model:

$$
\begin{aligned}
&\log(\mu_{it}^{\mathrm{D}}) \\
&\quad = \beta_{0t}^{\mathrm{D}}(\mathbf{U}_i) + \beta_{1t}^{\mathrm{D}}\log(I_{i,t-\delta}) + \alpha_{1t}^{\mathrm{D}}\mathrm{Control}_{i,t-7} + \alpha_{2t}^{\mathrm{D}}\mathrm{Mobility}_{i,t-7} \\
&\quad + \gamma_{1t}^{\mathrm{D}}\mathrm{Gini}_i + \gamma_{2t}^{\mathrm{D}}\mathrm{Affluence}_i + \gamma_{3t}^{\mathrm{D}}\mathrm{Disadvantage}_i + \gamma_{4t}^{\mathrm{D}}\mathrm{Urban}_i + \gamma_{5t}^{\mathrm{D}}\mathrm{PD}_i \\
&\quad + \gamma_{6t}^{\mathrm{D}}\mathrm{Tbed}_i + \gamma_{7t}^{\mathrm{D}}\mathrm{NHIC}_i + \gamma_{8t}^{\mathrm{D}}\mathrm{EHPC}_i + \gamma_{9t}^{\mathrm{D}}\mathrm{AA}_i + \gamma_{10t}^{\mathrm{D}}\mathrm{HL}_i + \gamma_{11t}^{\mathrm{D}}\mathrm{Sex}_i + \gamma_{12t}^{\mathrm{D}}\mathrm{Old}_i. \\
&\qquad\qquad (15)
\end{aligned}
$$

We consider the data collected from March 16 to September 3, 2020; see the data description in Section 5.1. Note that in Models (14)–(15), the covariate $\mathrm{Control}_{it}$ is a dummy variable for the executive order 'shelter-in-place' or 'stay-at-home', namely $\mathrm{Control}_{it} = 1$ suggesting 'shelter-in-place' taken place for county i at time t, while $\mathrm{Control}_{it} = 0$ suggesting no restriction or restriction lifted. See Table 2 for details of other county-level predictors.

We use 28 days, two incubation periods, as an estimation window to examine how the covariates affect the newly infected and fatal cases, and we choose $\delta = 14$. The roughness parameters are selected by generalised cross-validation (GCV). The performance of the univariate and bivariate splines depends on the choice of the knots and triangulations, respectively. We use cubic splines with two interior knots for the univariate spline smoothing. We generate the triangulations according to the 'max-min' criterion, which maximises the minimum angle of all the angles of the triangles in the triangulation. We consider two triangulation choices, $\triangle_1$ and $\triangle_2$, as shown in Figure 1. By the 'max-min' criterion, $\triangle_2$ is better than $\triangle_1$, but it also significantly increases the number of parameters to estimate. As a trade-off, for the estimation of $\beta_{0t}^{\mathrm{I}}(\cdot)$ and $\beta_{1t}^{\mathrm{I}}(\cdot)$, we adopt the finer triangulation $\triangle_2$, and use the rough triangulation $\triangle_1$ to estimate $\beta_{0t}^{\mathrm{D}}(\cdot)$ due to the sparsity problem in the death count and many zeros observed.

First of all, we describe our findings from modelling the COVID-19 related infection counts in 3104 counties from the 48 mainland US states and the District of Columbia. To examine the effect of the control measures ('shelter-in-place' or 'stay-at-home' orders) and mobility level after 7 days, we test the hypothesis: $H_0 : \alpha_{jt}^{\mathrm{I}} = 0,\ j = 1, 2$ in model (14). Figure 5(a) shows that the control measure is significant for the infected count most of the time. Figure 5(b) shows that the p-value of the mobility is always very close to zero, and thus the mobility is significant for the entire study period.



(a) Control (α^{I}_{1t})

(b) Mobility (α^{I}_{2t})

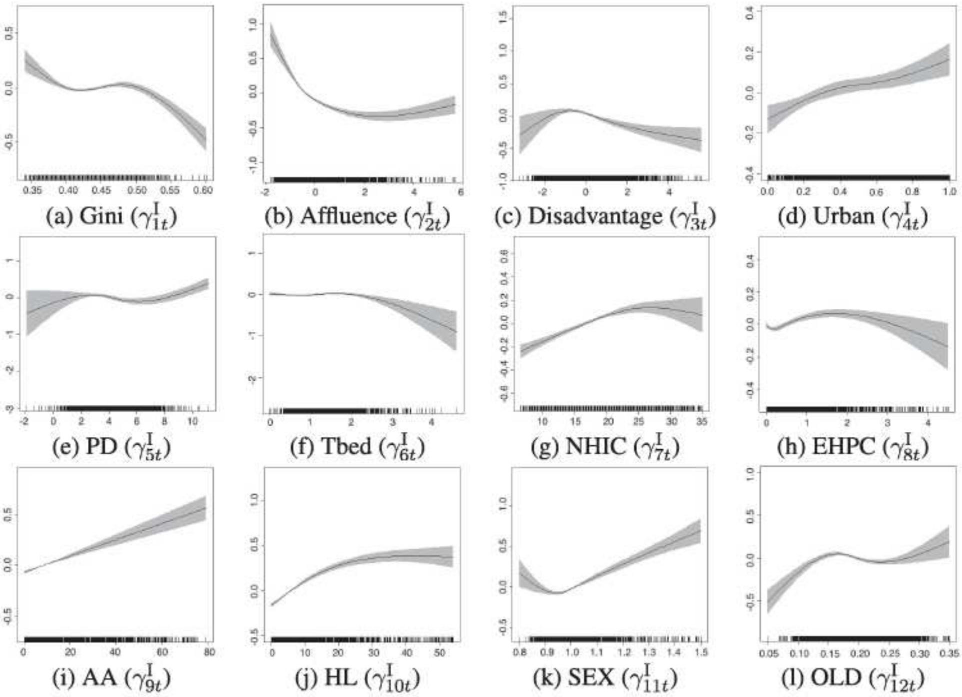
Figure 5. *P*-values of hypothesis tests of constant coefficients in model (14). (a) Control (α^{I}_{1t}) and (b) Mobility (α^{I}_{2t}).

Next, we examine the effect of the other predictors in Model 14. To test hypotheses $H_0 : \gamma^{\mathrm{I}}_{kt}(\cdot) = 0, k = 1, \ldots, 12$, we construct the 95% simultaneous confidence band (SCB) for $\gamma^{\mathrm{I}}_{kt}(\cdot)$'s. In function estimation problems, SCBs are an important tool to quantify and visualise the variability of the functional components; see Wang and Yang (2009), Cao, Yang, and Todem (2012), and Zheng, Liu, Yang, and Härdle (2016) for some related theory and applications. Figure 6 illustrates the estimated curves for different explanatory variables together with the corresponding SCBs based on the data period 03/22/2020–04/18/2020. Based on Figure 6, we can observe that at the beginning of the pandemic, the infected cases increase with the population density (PD), which is consistent with our intuition. We also find that the infections increase with African American Ratio and Hispanic Latino Ratio at the beginning of the outbreak.

We also study the effect of the covariates over time. Figure 7 shows the effect of the aged people rate at different time points over the outbreak. In the early stage of the COVID-19 pandemic, from March to April, COVID-19 struck the elderly more severely than the younger people. By mid-April and May, we saw that those communities with fewer aged people suffered more from COVID-19. Counties with a very high rate of aged people still experience high infection rates. However, when people understood the virus more and took action to protect the senior people, from mid-June to September, those counties with a higher rate of aged people became those least infected. However, from mid-June to September, older people tended to stay home and were more cautious about the virus. Also, as many states reopened bars, restaurants, and offices, people in their 20s and 30s were more likely to go out socialising, and the coronavirus spread more widely to young people (Bosman and Mervosh 2020).

Movies 1–12 in the Supplementary Material A show the estimates and SCBs of the nonparametric functions $\gamma^{\mathrm{I}}_{kt}(\cdot), k = 1, \ldots, 12$, over the entire study period in the STEM model (14).

After the discussion of our finding in the infection model, let us focus on the death model. For Model (15), we focus on the following hypothesis tests: $H_0 : \alpha^{\mathrm{D}}_{jt} = 0, j = 1, 2$, $H_0 : \beta^{\mathrm{D}}_{1t} = 0$ and $H_0 : \gamma^{\mathrm{D}}_{kt} = 0, k = 1, \ldots, 12$. Figure 8(b) shows that 'Mobility' is significant over the entire study period. "OLD" is significant in the beginning of the pandemic.



(a) Gini (γ^{I}_{1t}) (b) Affluence (γ^{I}_{2t}) (c) Disadvantage (γ^{I}_{3t}) (d) Urban (γ^{I}_{4t})

(e) PD (γ^{I}_{5t}) (f) Tbed (γ^{I}_{6t}) (g) NHIC (γ^{I}_{7t}) (h) EHPC (γ^{I}_{8t})

(i) AA (γ^{I}_{9t}) (j) HL ($\gamma^{\mathrm{I}}_{10t}$) (k) SEX ($\gamma^{\mathrm{I}}_{11t}$) (l) OLD ($\gamma^{\mathrm{I}}_{12t}$)

Figure 6. The estimated univariate functional components and the corresponding simultaneous confidence bands in the infection model. (a) Gini (γ^{I}_{1t}). (b) Affluence (γ^{I}_{2t}). (c) Disadvantage (γ^{I}_{3t}). (d) Urban (γ^{I}_{4t}). (e) PD (γ^{I}_{5t}). (f) Tbed (γ^{I}_{6t}). (g) NHIC (γ^{I}_{7t}). (h) EHPC (γ^{I}_{8t}). (i) AA (γ^{I}_{9t}). (j) HL ($\gamma^{\mathrm{I}}_{10t}$). (k) SEX ($\gamma^{\mathrm{I}}_{11t}$) and (l) OLD ($\gamma^{\mathrm{I}}_{12t}$).

For other county-level covariates, 'Affluence', 'Disadvantage', 'EHPC', 'AA' and 'SEX' are significant with p-values smaller than 0.05 most of time, while the rest of the predictors are significant on some days, but insignificant on other days.

In addition, movies 13 and 14 in the Supplementary Material A illustrate the estimated coefficient functions of $\beta^{\mathrm{I}}_{0t}(\cdot)$ and $\beta^{\mathrm{I}}_{1t}(\cdot)$ in model (14). From Movie 13, we can see that the transmission rate, $\beta^{\mathrm{I}}_{0t}(\cdot)$, varies at different locations and in different phases of the outbreak, especially the high rate in late March and April. Movie 14 shows that $\beta^{\mathrm{I}}_{1t}(\cdot)$ also varies from one location to another location, which indicates that the homogeneous mixing assumption of the simple SIR models does not hold. The transmission rate is high in most states at the end of April; however, it has become much lower since June. Movie 15 on the Supplementary Material A shows the pattern of $\widehat{\beta}^{\mathrm{D}}_{0t}$ in model (15). From this animation, we observe a severe fatality condition in the southern states in July and a pattern of a general decrease in the entire U.S. since August 2020.

6. Conclusion and discussion

This work has aimed to bridge the gap between mathematical models and statistical analysis in infectious disease studies. We created a state-of-art interface between mathematical and statistical models to understand the dynamic pattern of the spread of contagious

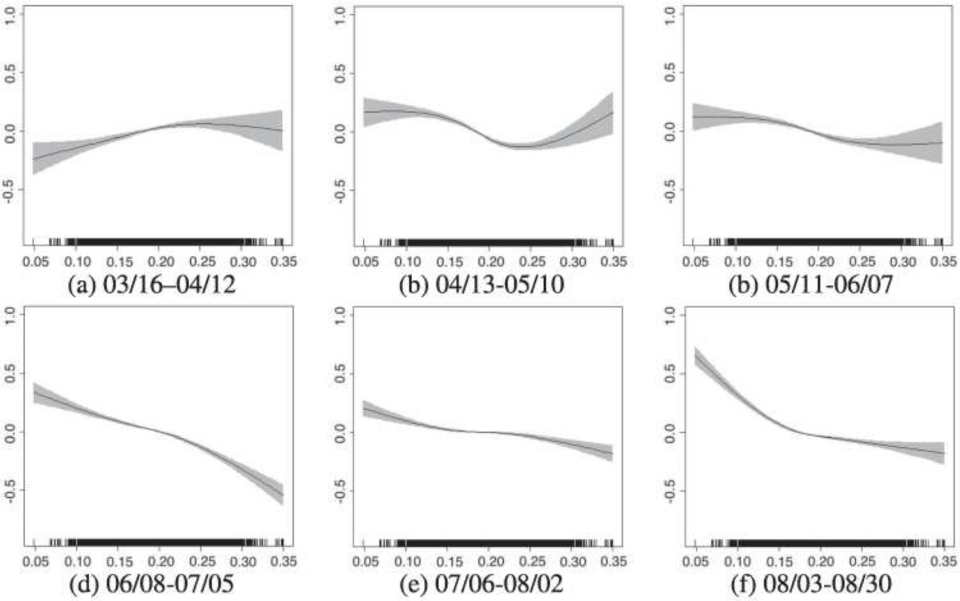


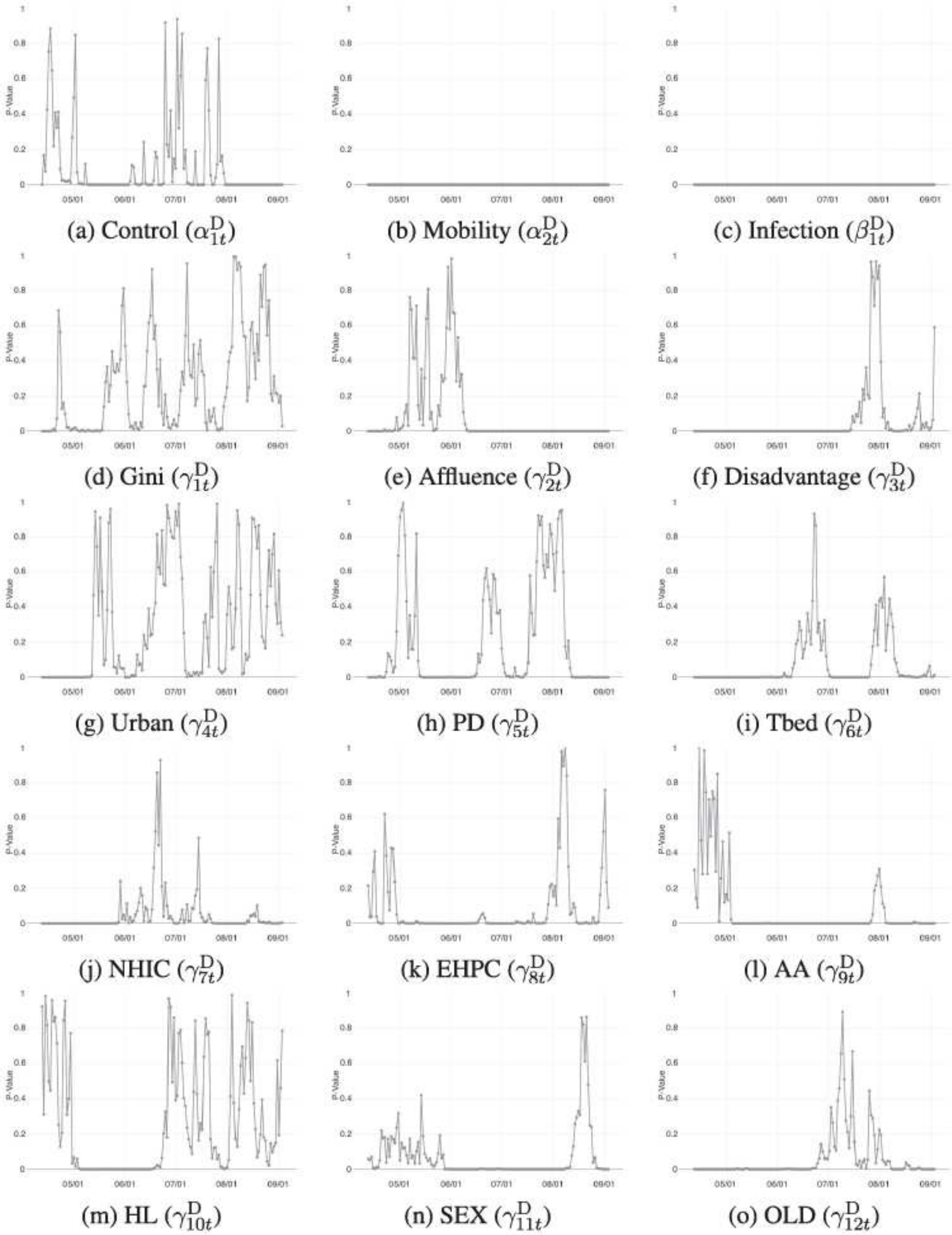
Figure 7. The estimated univariate functional components corresponding to the proportion of the elderly during different periods. (a) 03/16–04/12. (b) 04/13–05/10. (c) 05/11–06/07. (d) 06/08–07/05. (e) 07/06–08/02 and (f) 08/03–08/30 (γ_{6t}^{I}).

diseases. Our proposed model enhances the dynamics of the SIR mechanism through spatiotemporal analysis.

For analysing the confirmed and death cases of COVID-19, other factors may also be responsible for temporal or spatial patterns. We investigated the spatial associations between the infected count, death count, and factors or characteristics of the counties across the U.S. by modelling the daily infected/fatal cases at the county level considering the county-level factors. Modeling COVID-19 at the county-level and combining local characteristics are very beneficial for the community to understand the dynamics of the disease spread and support decision-making when urgently needed. To examine spatial nonstationarity in the transmission rate of the disease, we proposed a nonparametric spatially varying coefficient model, which allows the transmission to vary from one area to another area. The proposed method can be used as an essential tool for understanding the dynamics of the disease spread, as well as for assessing how this outbreak may unfold through time and space.

Based on our results, disease mapping can easily be implemented to illustrate high-risk areas and thus help policymaking and resource allocation. Our method can also be extended to other situations, including epidemic models in which there are several types of individuals with potentially different area characteristics or more complex models that include features such as latent periods or a more realistic population structure.

Our paper did not take the under-reported issue (for example, the asymptomatic coronavirus infectious cases) into account. Although our model may have partially corrected the problem with the spatiotemporal information, some better ways are proposed in several recent pieces of research, such as Pullano et al. (2021), Shaman (2021), Giordano



(a) Control (α^{D}_{1t}) (b) Mobility (α^{D}_{2t}) (c) Infection (β^{D}_{1t})

(d) Gini (γ^{D}_{1t}) (e) Affluence (γ^{D}_{2t}) (f) Disadvantage (γ^{D}_{3t})

(g) Urban (γ^{D}_{4t}) (h) PD (γ^{D}_{5t}) (i) Tbed (γ^{D}_{6t})

(j) NHIC (γ^{D}_{7t}) (k) EHPC (γ^{D}_{8t}) (l) AA (γ^{D}_{9t})

(m) HL (γ^{D}_{10t}) (n) SEX (γ^{D}_{11t}) (o) OLD (γ^{D}_{12t})

Figure 8. *P*-values of the hypothesis test of the constant coefficient in model (15). (a) Control (α^{D}_{1t}). (b) Mobility (α^{D}_{2t}). (c) Infection (β^{D}_{1t}). (d) Gini (γ^{D}_{1t}). (e) Affluence (γ^{D}_{2t}). (f) Disadvantage (γ^{D}_{3t}). (g) Urban (γ^{D}_{4t}). (h) PD (γ^{D}_{5t}). (i) Tbed (γ^{D}_{6t}). (j) NHIC (γ^{D}_{7t}). (k) EHPC (γ^{D}_{8t}). (l) AA (γ^{D}_{9t}). (m) HL (γ^{D}_{10t}). (n) SEX (γ^{D}_{11t}) and (o) OLD (γ^{D}_{12t}).

et al. (2021), and Moore, Hill, Dyson, Tildesley, and Keeling (2021). Furthermore, some of the newly developed methods also investigate the effect of vaccinations and different COVID-19 variants. For example, Giordano et al. (2021) and Moore et al. (2021) proposed an extended SEIR-type framework. In these models, individuals start from the susceptible-unvaccinated or susceptible-vaccinated states. Then those in the asymptomatic state will recover, and those in the symptomatic state may become either recover or die. Moreover, the infected individuals could also be divided into several groups based on COVID-19 variants. As discussed in Section 2, although we introduce our discrete-time spatial epidemic model based on the SIR model, we can extend it to such a kind of SEIR-type framework as well.

Acknowledgments

The authors would like to thank the Editor-in-Chief, an associate editor, and two referees for their constructive comments and suggestions that significantly improved an earlier version of this article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Myungjin Kim's research was partially supported by National Science Foundation (Division of Computing and Communication Foundations) award CCF-1934884, DMS-2135493, DMS-1916204 and Laurence H. Baker Center for Bioinformatics & Biological Statistics. Shan Yu's research was partially supported by the Iowa State University Plant Sciences Institute Scholars Program. Li Wang's research was partially supported by National Science Foundation (Division of Mathematical Sciences) award DMS-1916204 and DMS-2135493.

ORCID

Yueying Wang http://orcid.org/0000-0003-4861-2658
Myungjin Kim http://orcid.org/0000-0001-7784-0516
Shan Yu http://orcid.org/0000-0002-0271-5726
Xinyi Li http://orcid.org/0000-0003-0080-7034
Guannan Wang http://orcid.org/0000-0001-6551-4465
Li Wang http://orcid.org/0000-0001-8432-9986

References

Anastassopoulou, C., Russo, L., Tsakris, A., and Siettos, C. (2020), 'Data-Based Analysis, Modelling and Forecasting of the COVID-19 Outbreak', *PLOS ONE*, 15, 1–21.

Arab, A., Holan, S.H., Wikle, C.K., and Wildhaber, M.L. (2012), 'Semiparametric Bivariate Zero-Inflated Poisson Models with Application to Studies of Abundance for Multiple Species', *Environmetrics*, 23, 183–196.

Atlantic (2020), 'The COVID Tracking Project Data', Dataset. https://covidtracking.com/api.

Bosman, J., and Mervosh, S. (2020), 'As Virus Surges, Younger People Account for "Disturbing" Number of Cases'. https://www.nytimes.com/2020/06/25/us/coronavirus-cases-young-people.html.

Brauer, F., Van den Driessche, P., and Wu, J. (2008), *Mathematical Epidemiology*, (Vol. 1945), Berlin: Springer.

Cao, G., Yang, L., and Todem, D. (2012), 'Simultaneous Inference for the Mean Function Based on Dense Functional Data', *Journal of Nonparametric Statistics*, 24, 359–377.

CDC (2021a), 'COVID-19 Pandemic Planning Scenarios'. https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html.

CDC (2021b), 'Ending Home Isolation for Persons with COVID-19 Not in Healthcare Settings'. https://www.cdc.gov/coronavirus/2019-ncov/hcp/disposition-in-home-patients.html.

Finkenstädt, B.F., and Grenfell, B.T. (2000), 'Time Series Modelling of Childhood Diseases: A Dynamical Systems Approach', *Journal of the Royal Statistical Society, Series C*, 49, 187–205.

Giordano, G., Colaneri, M., Di Filippo, A., Blanchini, F., Bolzern, P., De Nicolao, G., Sacchi, P., Colaneri, P., and Bruno, R. (2021), 'Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in Italy', *Nature Medicine*, 27, 993–998.

Gog, J.R. (2020), 'How You Can Help with COVID-19 Modelling', *Nature Reviews Physics*, 2, 274–275.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, Vol. 43, 1st ed., London: Chapman and Hall/CRC Press.

Held, L., Hens, N., D O'Neill, P., and Wallinga, J. (2019), *Handbook of Infectious Disease Data Analysis*, New York: CRC Press.

Howard, J., and Yu, G. (2020), 'Most People Recover from Covid-19. Here's Why It's Hard to Pinpoint Exactly How Many', CNN News. https://www.cnn.com/2020/04/04/health/recovery-coronavirus-tracking-data-explainer/index.html.

JHU CSSE (2020), '2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository', Dataset. https://github.com/CSSEGISandData/COVID-19.

Jia, J.S., Lu, X., Yuan, Y., Xu, G., Jia, J., and Christakis, N.A. (2020), 'Population Flow Drives Spatio-Temporal Distribution of COVID-19 in China', *Nature*, 582, 389–394.

Jong, M., Diekmann, O., and Heesterbeek, J. (1995), 'How Does Transmission Depend on Population Size?', *Publication of the Newton Institute*, 5, 84.

Kim, M., and Wang, L. (in press), 'Generalized Spatially Varying Coefficient Models', *Journal of Computational and Graphical Statistics*, 30, 1–10.

Kucharski, A.J., Russell, T.W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., Eggo, R.M., Sun, F., Jit, M., Munday, J.D., Davies, N. (2020), 'Early Dynamics of Transmission and Control of COVID-19: A Mathematical Modelling Study', *The Lancet Infectious Diseases*, 20, 553–558.

Lai, M.J., and Schumaker, L.L. (2007), *Spline Functions on Triangulations* (1st ed.), Cambridge: Cambridge University Press.

Lai, M.J., and Wang, L. (2013), 'Bivariate Penalized Splines for Regression', *Statistica Sinica*, 23, 1399–1417.

Lawson, A.B., Banerjee, S., Haining, R.P., and Ugarte, M.D. (2016), *Handbook of Spatial Epidemiology*, New York: CRC Press.

Liu, W.M., Hethcote, H.W., and Levin, S.A. (1987), 'Dynamical Behavior of Epidemiological Models with Nonlinear Incidence Rates', *Journal of Mathematical Biology*, 25, 359–380.

Liu, R., Yang, L., and Härdle, W.K. (2013), 'Oracally Efficient Two-Step Estimation of Generalized Additive Model', *Journal of the American Statistical Association*, 108, 619–631.

Moore, S., Hill, E.M., Dyson, L., Tildesley, M.J., and Keeling, M.J. (2021), 'Modelling optimal vaccination strategy for SARS-CoV-2 in the UK', *PLoS computational biology*, 17, e1008849.

Mu, J., Wang, G., and Wang, L. (2018), 'Estimation and Inference in Spatially Varying Coefficient Models', *Environmetrics*, 29, e2485.

NYT (2020), 'Coronavirus (Covid-19) Data in the United States', Dataset. https://github.com/nytimes/covid-19-data.

Pfeiffer, D.U., Robinson, T.P., Stevenson, M., Stevens, K.B., Rogers, D.J., and Clements, A.C. (2008), *Spatial Analysis in Epidemiology*, New York: Oxford University Press.

Pullano, G., Di Domenico, L., Sabbatini, C.E., Valdano, E., Turbelin, C., Debin, M., Guerrisi, C, Kengne-Kuetche, C., Souty, C., Hanslik, T., Blanchon, T., Boëlle, P.Y., Figoni, J, Vaux, S., Campèse, C., Bernard-Stoecklin, S, and Colizza, V. (2021), 'Underdetection of cases of COVID-19 in France threatens epidemic control', *Nature*, 590, 134–139.

Sangalli, L., Ramsay, J., and Ramsay, T. (2013), 'Spatial Spline Regression Models', *Journal of the Royal Statistical Society, Series B*, 75, 681–703.

Shaman, J. (2021), 'An estimation of undetected COVID cases', *Nature*, 590, 38–39.

Siettos, C.I., and Russo, L. (2013), 'Mathematical Modeling of Infectious Disease Dynamics', *Virulence*, 4, 295–306.

Sun, H., Qiu, Y., Yan, H., Huang, Y., Zhu, Y., Gu, J., and Chen, S.X. (2020), 'Tracking Reproductivity of COVID-19 Epidemic in China with Varying Coefficient SIR Model', *Journal of Data Science*, 18, 455–472.

USAFacts (2020), 'Coronavirus Locations: COVID-19 Map by County and State', Dataset. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map.

Vespignani, A., Tian, H., Dye, C., Lloyd-Smith, J.O., Eggo, R.M., Shrestha, M., Scarpino, S.V., Gutierrez, B., Kraemer, M.U.G., Wu, J., Leung, K., and Leung, G.M. (2020), 'Modelling COVID-19', *Nature Reviews Physics*, 2, 279–281.

Wakefield, J., Dong, T.Q., and Minin, V.N. (2019), 'Spatio-Temporal Analysis of Surveillance Data', in *Handbook of Infectious Disease Data Analysis*, New York: Chapman and Hall/CRC Press, pp. 455–476.

Wang, G., Gu, Z., Li, X., Yu, S., Kim, M., Wang, Y., Gao, L., and Wang, L. (2020), 'Comparing and Integrating US COVID-19 Data from Multiple Sources with Anomaly Detection and Repairing', Preprint. arXiv:2006.01333v3.

Wang, L., and Lai, M.J. (2019), 'Triangulation', R package version 1.0. https://github.com/funstatpackages/Triangulation.

Wang, L., Liu, X., Liang, H., and Carroll, R. (2011), 'Estimation and Variable Selection for Generalized Additive Partial Linear Models', *The Annals of Statistics*, 39, 1827–1851.

Wang, G., Wang, L., Lai, M.J., Kim, M., Li, X., Mu, J., Wang, Y., and Yu, S. (2019), 'BPST: Bivariate Spline over Triangulation', R package version 1.0. https://github.com/funstatpackages/BPST.

Wang, L., Xue, L., and Yang, L. (2020), 'Estimation of Additive Frontier Functions with Shape Constraints', *Journal of Nonparametric Statistics*, 32, 262–293.

Wang, J., and Yang, L. (2009), 'Polynomial Spline Confidence Bands for Regression Curves', *Statistica Sinica*, 19, 325–342.

Wood, S.N. (2017), *Generalized additive models: an introduction with R*, New York: Chapman and Hall/CRC Press.

Wood, S.N., Pya, N., and Säfken, B. (2016), 'Smoothing Parameter and Model Selection for General Smooth Models', *Journal of the American Statistical Association*, 111, 1548–1563.

Xue, L., and Liang, H. (2010), 'Polynomial Spline Estimation for A Generalized Additive Coefficient Model', *Scandinavian Journal of Statistics*, 37, 26–46.

Yu, S., Wang, G., Wang, L., Liu, C., and Yang, L. (2020), 'Estimation and Inference for Generalized Geoadditive Models', *Journal of the American Statistical Association*, 115, 761–774.

Zheng, S., Liu, R., Yang, L., and Härdle, W.K. (2016), 'Statistical Inference for Generalized Additive Models: Simultaneous Confidence Corridors and Variable Selection', *Test*, 25, 607–626.