Spatiotemporal Dynamics, Nowcasting and Forecasting of COVID-19 in the United States

Li Wang[a], Guannan Wang[b], Lei Gao[a], Xinyi Li[c], Shan Yu[a], Myungjin Kim[a], Yueying Wang[a] and Zhiling Gu[a]

[a]*Iowa State University, USA,* [b]*College of William & Mary, USA and* [c]*SAMSI / University of North Carolina at Chapel Hill, USA*

Abstract: In response to the ongoing public health emergency of COVID-19, we investigate the disease dynamics to understand the spread of COVID-19 in the United States. In particular, we focus on the spatiotemporal dynamics of the disease, accounting for the control measures, environmental effects, socioeconomic factors, health service resources, and demographic conditions that vary from different counties. In the modeling of an epidemic, mathematical models are useful, however, pure mathematical modeling is deterministic, and only demonstrates the average behavior of the epidemic; thus, it is difficult to quantify the uncertainty. Instead, statistical models provide varieties of characterization of different types of errors. In this paper, we investigate the disease dynamics by working at the interface of theoretical models and empirical data by combining the advantages of mathematical and statistical models. We develop a novel nonparametric space-time disease transmission model for the epidemic data, and to study the spatial-temporal pattern in the spread of COVID-19 at the county level. The proposed methodology can be used to dissect the spatial structure and dynamics of spread, as well as to forecast how this outbreak may unfold through time and space in the future. To assess the uncertainty, projection bands are constructed from forecast paths obtained in bootstrap replications. A dashboard is established with multiple R shiny apps embedded to provide a 7-day forecast of the COVID-19 infection count and death count up to the county level, as well as a long-term projection of the next four months. The proposed method provides remarkably accurate short-term prediction results.

Key words and phrases: Coronavirus; Dynamic models in epidemics; Nonparametric modeling; Prediction; Spatial epidemiology; Varying coefficient models.

1 Introduction

Since the beginning of the reported cases in December 2019, the outbreak of COVID-19 has spread globally within weeks. On March 11, 2020, the World Health Organization (WHO) deemed COVID-19 to be a pandemic (WHO, 2020), and on March 24, they warned that the U.S. could be the next epicenter of the global coronavirus pandemic. The reported confirmed cases in the U.S. have soared in the following weeks. The coronavirus is spreading from the biggest cities in the U.S. to its suburbs, and it has begun encroaching on the nation's rural regions. According to the New York Times, as of April 30 8:01 A.M. EST, there are now at least 1,045,300 confirmed cases and 60,900 deaths from COVID-19 in the U.S.

Address for correspondence: Li Wang (lilywang@iastate.edu)

An essential question for developing a defense against COVID-19 is how far the virus will spread and how many lives it will claim. It is not clear to anyone where this crisis will lead us. Understanding the dynamics of the disease is therefore undoubtedly critical. One way to answer these questions is through scientific modeling. Several attempts have been made to model and forecast the spread and mortality of COVID-19 (Elmousalami and Hassanien, 2020; Fanelli and Piazza, 2020; Kucharski et al., 2020; Pan et al., 2020; Sun et al., 2020; Wang et al., 2020d; Zhang et al., 2020).

In epidemiology, the fundamental concept of infectious disease is the investigation of how infections spread. Mathematical methods, such as the class of susceptible-infectious-recovered (SIR) models (Allen et al., 2008; Chen et al., 2020; Lawson et al., 2016; Pfeiffer et al., 2008; Wakefield et al., 2019; Weiss, 2013), are widely used in epidemics to capture the dynamic process of the spread of the infectious disease. However, pure mathematical modeling is deterministic, and only demonstrates the average behavior of the epidemic. In addition, its focus is often on the form of models, not the parameter estimation for observed data; thus, it is difficult to quantify the uncertainty. Instead, statistical models provide a varying characterization of different types of errors. When it comes to analyzing the reported numbers of infectious diseases, other factors may also be responsible for temporal or spatial patterns. The spread of the disease varies a lot across different geographical regions. Local area features, like socioeconomic factors and demographic conditions, can dramatically influence the course of the epidemic. These data are usually supplemented with the population information at the county level. In addition, 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. SIR models with assumptions of random mixing can overestimate the health service needed by not taking into account the behavioral change and government-mandated action. In this paper, we propose a class of novel nonparametric dynamic epidemic models to analyze the infectious disease data by incorporating the spatiotemporal structure and the effect of explanatory variables.

In this paper, we borrow the mechanistic rules from the SIR model by including three compartments: infected, susceptible and removed states, and develop a class of data-driven statistical models to reconstruct the spatiotemporal dynamics of the disease transmission. We build a novel space-time epidemic modeling framework for the infected count data, to study the spatial-temporal pattern in the spread of COVID-19 at the county level. The proposed methodology can be used to dissect the spatial structure and dynamics of spread, as well as to assess how this outbreak may unfold through time and space.

Given an parametric epidemic model, the typical inference problem involves estimating the parameters associated with the parametric models from the data to hand. Such specifications are ad hoc, and if misspecified, can lead to substantial estimation bias problems. In practice, this question might be addressed by considering alternative parametric models, or sensitivity analyses if some of the underlying model parameters are assumed to be known. Nonparametric approaches to fitting epidemic models

to the data have received relatively little attention in the literature possibly due to the lack of data. By allowing the infection to depend on time and location, we consider a generalized additive varying coefficient model to estimate the unobserved process of the disease transmission. By adopting a non-parametric approach, we do not impose a particular parametric structure, which significantly enhances the flexibility of the epidemic models that practitioners use. For our model estimation, we propose a quasi-likelihood approach via the penalized spline approximation and the iteratively reweighted least squares technique.

Prediction models for COVID-19 at the county-level that combine local characteristics and actions are very beneficial for the community to understand the dynamics of the disease spread and support decision making at a time when they are urgently needed. Models can help predict rates of new infections, and estimate when the strain on the hospital system could peak. In this paper, we consider both the short-term and long-term impact of the virus. To assess the uncertainty associated with the prediction, we develop a projection band constructed based on the envelope of the bootstrap forecast paths, which are closest to the forecast path obtained on the basis of the original sample. Based on our research findings, we develop multiple R shiny apps embedded into a COVID-19 dashboard, which provides a 7-day forecast and a 4-month forecast of COVID-19 infected and death count at both the county level and state level.

The rest of the paper is organized as follows. Section 2 introduces our case study on COVID-19, including a detailed description of the data. Section 3 outlines the nonparametric spatiotemporal modeling framework and describes how to incorporate additional covariates. Section 4 introduces our estimation method, presents our algorithms, and discusses the details of the implementation. Section 5 starts with a description of the prediction of the infection count and provides the uncertainty with the band of the forecast path. Section 6 shows the results and findings of the case study. Section 7 concludes the paper with a discussion. The supplementary materials (Wang et al., 2020c) contain additional figures, and an animation of the estimation results.

2 COVID-19 Case Study and Data

2.1 Research Goal of the Study

The goal of this study is threefold. First, we develop a new dynamic epidemic modeling framework for public health surveillance data to study the spatial-temporal pattern in the spread of COVID-19. We aim to investigate whether the proposed model could be used to guide the modeling of the dynamic of the spread at the county level by moving beyond the typical theoretical conceptualization 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 in consideration of the demographic, environmental, behavioral, socioeconomic factors in the U.S. Third,

we project the spatial-temporal pattern of the spread of the virus in the U.S. For the short-term forecast, we provide the prediction of the daily infection count and death count up to the county level. As for the long-term forecast, we project the total infected and death cases in the next three months.

2.2 Epidemic Data from the COVID-19 Outbreak in the U.S.

This study analyzes data from the reported confirmed COVID-19 infections and deaths at the county level, which are reported by the 3,104 counties from the 48 mainland U.S. states and the District of Columbia. The aggregated COVID-19 cases are from January 20 until April 25, 2020. The data are collected, compiled and cleaned from a combination of public sources that aim to facilitate the research effort to confront COVID-19, including Health Department Website in each state or region, the New York Times (NYT, 2020), the COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University (CSSE, 2020), and the COVID Tracking Project (Atlantic, 2020). These data sources automatically updated every day or every other day. We have created a dashboard `https://covid19.stat.iastate.edu/` to visualize and track the infected and death cases, which was launched on March 27, 2020.

2.3 Information of the covariates

We consider a variety of county-level characteristics as covariate information in our study, which can be divided into six groups. The data sources and the operational definitions of these features are discussed as follows.

Policies. Government declarations are used to identify the dates that different jurisdictions implemented various social distancing policies (emergency declarations, school closures, bans on large gatherings, limits on bars, restaurants and other public places, the deployment of severe travel restrictions, and "stay-at-home" or "shelter-in-place" orders). President Trump declared a state of emergency on March 13, 2020, to enhance the federal government response to confront the COVID-19. By March 16, 2020, every state had made an emergency declaration. Since then, more severe social distancing actions have been taken by the majority of the states, especially those hardest hit by the pandemic. We compiled the dates of executive orders by checking national and state governmental websites, news articles and press releases.

Demographic Characteristics. To capture the demographic characteristics of a county, five variables are considered in the analysis to describe the racial, ethnic, sexual and age structures: the percent of the population who identify as African American, the percent of the population who identify as Hispanic or Latino, the rate of aged people ($\geq$ 65 years) per capita, the ratio of male over female and population density over square mile of land area. The former two variables were obtained from the 2010 Census (U.S. Census Bureau, 2010), and the latter three variables are extracted from the 2010–2018 American Community Survey (ACS) Demographic and Housing Estimates (U.S. Census Bureau, 2010).

Healthcare Infrastructure. We incorporated three components in our analysis to describe the healthcare infrastructure in each county: percent of the population aged less than 65 years without health insurance, local government expenditures for health per capita, and total counts of hospital beds per 1,000 population. These components measure the access for residents to public health resources within and across counties. The first component is available in the USA Counties Database (U.S. Census Bureau, 2010), the second is from Economic Census 2012 (U.S. Census Bureau, 2012), and the last is compiled from Homeland Infrastructure Foundation-level Data (U.S. Department of Homeland Security).

Socioeconomic Status. A diverse of factors are considered to describe the socioeconomic status in each county. We first apply the factor analysis to seven factors collected from the 2005–2009 ACS 5-year estimates (U.S. Census Bureau, 2010), and generate two factors: social affluence and concentrated disadvantage. To be specific, the former is comprised of the percent of families with annual incomes higher than \$75,000 (factor loading = 0.86), percent of the population aged 25 years or older with a bachelor's degree or higher (factor loading = 0.92), percent of the people working in management, professional, and related occupations (factor loading = 0.73), and the median value of owner-occupied housing units (factor loading = 0.74); whereas the latter includes the percent of the households with public assistance income (factor loading = 0.34), the percent of households with female householders and no husband present (factor loading = 0.81), and civilian labor force unemployment rate (factor loading = 0.56). These two factors, affluence and disadvantage, explain more than 60% of the variation.

We also incorporate the Gini coefficient to measure income inequality. The Gini coefficient, also known as Gini index, is a well-known measure for income inequality and wealth distribution in economics, with value ranging from zero (complete equality, where everyone has exactly the same income) to one (total inequality, where one person occupies all of the income). The 2005–2009 ACS (U.S. Census Bureau, 2010) provided the household income data that allow us to calculate the Gini coefficient.

Rural/urban Factor. In the literature, rural/urban residence has been found to be associated with the spread of epidemics. Specifically, rural counties are often characterized by poor socio-economic profiles and limited access to healthcare services, indicating a potential higher risk. To capture rural/urban residence, we use the urban rate from the 2010 Census (U.S. Census Bureau, 2010).

Geographic Information. The longitude and latitude of the geographic center for each county in the U.S. are available in Gazetteer Files (U.S. Census Bureau, 2019).

3 Space-time Epidemic Modeling

In this section, we propose a class of nonparametric space-time models to estimate the infection count at the area level. In the following, let Y_{it} be the number of new cases at time t for area i, $i = 1, \ldots, n$. Also for area i, let I_{it}, D_{it} and R_{it} be the cumulative number of active infectious, death and recovered cases at time t, and let C_{it} be the number of cumulative confirmed cases up to time t. Then, it is clear

that $I_{it} = \sum_{j=1}^{t} Y_{ij} - D_{it} - R_{it}$. Further, denote N_i the population for area i, and the number of susceptible subjects at time t would be $S_{it} = N_i - C_{it}$. Define $Z_{it} = \log(S_{it}/N_i)$.

We denote $\mathbf{U}_i = (U_{i1}, U_{i2})^\top$ be the GPS coordinates of the geographic center 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 the covariates of area i that is not varying with time, see the description in Section 2.3. For example, the socioeconomic factors, health service resources, and demographic conditions. Let A_{ijt} denotes the jth dummy variable of actions or 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 given $(I, Z, \mathbf{A}, \mathbf{X}, \mathbf{U}) = (\mathrm{i}, z, \boldsymbol{a}, \boldsymbol{x}, \boldsymbol{u})$ can be represented as

$$f_{Y|I,Z,\mathbf{A},\mathbf{X},\mathbf{U}}\left(y|\,\mathrm{i}, z, \boldsymbol{a}, \boldsymbol{x}, \boldsymbol{u}\right) = \exp\left[\sigma^{-2}\left\{y\zeta\left(\mathrm{i}, z, \boldsymbol{a}, \boldsymbol{x}, \boldsymbol{u}\right) - \mathcal{B}\left\{\zeta\left(\mathrm{i}, z, \boldsymbol{a}, \boldsymbol{x}, \boldsymbol{u}\right)\right\}\right\} + \mathcal{C}\left(y, \sigma^2\right)\right],$$

for some known functions $\mathcal{B}$ and $\mathcal{C}$, dispersion parameter σ^2 and the canonical parameter ζ. Let $\mu\left(\mathrm{i}, z, \boldsymbol{a}, \boldsymbol{x}, \boldsymbol{u}\right)$ be the conditional expectation of Y given $(I, Z, \mathbf{A}, \mathbf{X}, \mathbf{U}) = (\mathrm{i}, z, \boldsymbol{a}, \boldsymbol{x}, \boldsymbol{u})$.

We assume that the determinants of the daily new cases of a certain area can be explained not only by the features of this area but also by the characteristics of the surrounding areas. Based on the idea of the SIR models, we propose a discrete-time spatial epidemic model comprising the susceptible, infected and removed states, and area-level characteristics. At time point t, we assume $\mu_{it} = \mu\left(I_{i,t-1}, Z_{i,t-1}, \mathbf{X}_i, \mathbf{A}_{i,t-r}, \mathbf{U}_i\right)$, which is modeled via a link function g as follows:

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

where α_{jt}'s are unknown time-varying coefficients, $\beta_{0t}(\cdot)$ and $\beta_{1t}(\cdot)$ are unknown bivariate coefficient functions, $\gamma_{kt}(\cdot)$, $k = 1, \ldots, q$, are univariate functions to be estimated. The parameter r in $A_{ij,t-r}$'s denotes a small delay time allowing for the control measure to be effective. For model identifiability, we assume $\mathrm{E}(\gamma_{kt}) = 0$, $k = 1, \ldots, q$. Note that $\exp\{\beta_{0t}(\boldsymbol{u})\}$ illustrates the transmission rate at location $\boldsymbol{u}$, β_{1t}, α_{0t} are the mixing parameters of the contact process. The rationale for including $\beta_{1t}(\cdot)$ ($0 < \beta_{1t} < 1$) is to allow for deviations from mass action and to account for the discrete-time approximation to the continuous time model; see Finkenstädt and Grenfell (2000); Wakefield et al. (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 behavior than those with bilinear incidence rates (Liu et al., 1987). These dynamical behaviors are determined mainly by β_{0t} and β_{1t}. When β_{1t} and α_{0t} are both 1, it is corresponding to the standard assumption of homogeneous mixing in De Jong et al. (1995).

Since Y_{it} is the number of new cases at time t for area i, $i = 1, \ldots, n$, Poisson or negative binomial (NB) might be an appropriate option for random component; see Yu et al. (2020), and Kim and Wang (2020). We assume that

- (Poisson) $\mathrm{E}(Y_{it}|\mathbf{Z}_{i,t-1}, \mathbf{A}_{i,t-r}, \mathbf{X}_i, \mathbf{U}_i) = \mu_{it}, \ \ \mathrm{Var}(Y_{it}|\mathbf{Z}_{i,t-1}, \mathbf{A}_{i,t-r}, \mathbf{X}_i, \mathbf{U}_i) = \mu_{it}$,
- (NB) $\mathrm{E}(Y_{it}|\mathbf{Z}_{i,t-1}, \mathbf{A}_{i,t-r}, \mathbf{X}_i, \mathbf{U}_i) = \mu_{it}, \ \ \mathrm{Var}(Y_{it}|\mathbf{Z}_{i,t-1}, \mathbf{A}_{i,t-r}, \mathbf{X}_i, \mathbf{U}_i) = \mu_{it}(1+\mu_{it}/I_{i,t-1})$,

where μ_{it} can be modeled via the same log link as follows:

$$\log(\mu_{it}) = \beta_{0t}(\mathbf{U}_i) + \beta_{1t}(\mathbf{U}_i)\log(I_{i,t-1}) + \alpha_{0t}Z_{i,t-1} + \sum_{j=1}^{p}\alpha_{jt}A_{ij,t-r} + \sum_{k=1}^{q}\gamma_{kt}(X_{ik}). \quad (2)$$

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 to account for the presence of over-dispersion.

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

4 Estimation of the STEM

4.1 Penalized Quasi-likelihood Method

In this section, we describe how to estimate the parameters and nonparameteric components in the proposed STEM model (2).

To capture the temporal dynamics, we consider the moving window approach. For the current time t, and nonnegative smoothness parameters λ_ℓ for $\ell = 0, 1$, we consider the following penalized quasi-likelihood problem:

$$\sum_{i=1}^{n}\sum_{s=t-t_0}^{t} L\left[g^{-1}\left\{\beta_{0s}(\mathbf{U}_i) + \beta_{1s}(\mathbf{U}_i)\log(I_{i,s-1}) + \alpha_{0s}Z_{i,s-1} + \sum_{j=1}^{p}\alpha_{js}A_{ij,s-r} + \sum_{k=1}^{q}\gamma_{ks}(X_{ik})\right\}, Y_{is}\right] - \frac{1}{2}\left\{\lambda_0\mathcal{E}(\beta_0) + \lambda_1\mathcal{E}(\beta_1)\right\}, \quad (3)$$

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

$$\mathcal{E}(\beta) = \int_\Omega \left\{(\nabla^2_{u_1}\beta)^2 + 2(\nabla_{u_1}\nabla_{u_2}\beta)^2 + (\nabla^2_{u_2}\beta)^2\right\}\mathrm{d}u_1\mathrm{d}u_2, \quad (4)$$

where $\nabla^q_{u_j}\beta(\boldsymbol{u})$ 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 overcome this difficulty, we introduce the basis expansion for 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 (2) are approximated using univariate polynomial spline and bivariate penalized 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 Schumaker (2007), Lai and Wang (2013) and Sangalli et al. (2013). We introduce a brief review of univariate splines and bivariate splines in the following.

Suppose that the covariate X_k is distributed on an interval $[a_k, b_k]$, $k = 1, \ldots, q$. For $k = 1, \ldots, q$, denote $\delta k = \{a_k = \delta k, 0 < \delta k, 1 < \cdots < \delta_{k,J_n} < \delta k, J_n + 1 = b_k\}$ a partition of $[a_k, b_k]$ with J_n interior knots. Let $\mathcal{U}k = \mathcal{U}_k^\varrho([a_k, b_k], \delta k)$ be the space of the polynomial splines of order $\varrho + 1$, which are polynomial functions with ϱ-degree (or less) on intervals $[\delta k, j, \delta_{k,j+1})$, $j = 0, \ldots, J_n - 1$, and $[\delta k, J_n, \delta_{k,J_n+1}]$, and have $\varrho - 1$ continuous derivatives globally. Next, let $\mathcal{U}_k^0 = \{\phi \in \mathcal{U}_k : \mathrm{E}\phi(X_k) = 0\}$, which ensures that the spline functions are centered; see Yu et al. (2020).

Let $\{\varphi_{kj}(x_k), j \in \mathcal{J}\}$ be the original B-spline basis functions for the kth covariate, where $\mathcal{J}$ is the index set of the basis functions. Let $\varphi^0_{kj}(x_k) = \varphi_{kj}(x_k) - \varphi_{k1}(x_k)\mathrm{E}\varphi_{kj}(X_k)/\mathrm{E}\varphi_{k1}(X_k)$, $\phi_{kj}(x_k) = \varphi^0_{kj}(x_k)/\mathrm{SD}\{\varphi^0_{kj}(X_k)\}$, $j \in \mathcal{J}$, then $\mathrm{E}\phi_{kj}(X_k) = 0$ and $\mathrm{E}\phi^2_{kj}(X_k) = 1$. Suppose the nonlinear component can be well approximated by a spline function so that, for all $x_k \in [a_k, b_k]$, $\gamma_{kt}(x_k) \approx \sum_{j=1}^{J_n+\varrho+1} \xi_{ktj}\phi_{kj}(x_k) = \boldsymbol{\Phi}_k^\top(x_k)\boldsymbol{\xi}_{kt}$, where $\boldsymbol{\Phi}_k(x_k) = (\phi_{k1}(x_k), \ldots, \phi_{k,J_n+\varrho+1}(x_k))^\top$ and $\boldsymbol{\xi}_{kt} = (\xi_{kt1}, \ldots, \xi_{kt,J_n+\varrho+1})^\top$ is a vector of coefficients.

For the bivariate coefficient functions $\beta_{0t}(\cdot)$ and $\beta_{1t}(\cdot)$ in the STEM model (2), we introduce bivariate spline over triangulation. The spatial domain Ω with either an arbitrary shape or holes inside can be partitioned into finitely many M triangles, $T_1, \ldots, T_M$, that is, $\Omega = \cup_{m=1}^M T_m$, and any nonempty intersection between a pair of triangles in $\triangle$ is either a shared vertex or a shared edge. A collection of these triangles, $\triangle := \{T_1, \ldots, T_M\}$, is called a triangulation of the domain Ω Lai and Schumaker (2007); Lai and Wang (2013). For a triangle $T \in \triangle$ in $\mathbb{R}^2$ with vertices $\boldsymbol{v}_i$, for $i = 1, 2, 3$, numbered in counter-clockwise order, we can write $T :=< \boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3 >$. Then, any point $\boldsymbol{v} \in \mathbb{R}^2$ can be uniquely represented as $\boldsymbol{v} = b_1\boldsymbol{v}_1 + b_2\boldsymbol{v}_2 + b_3\boldsymbol{v}_3$ such that $b_1 + b_2 + b_3 = 1$, where the coefficients (b_1, b_2, b_3) are called the "barycentric coordinates" of point $\boldsymbol{v} \in T$. The Bernstein basis polynomials of degree $d \geq 1$ relative to T are defined as $B^d_{T;ijk}(\boldsymbol{v}) = d!/(i!j!k!)b_1^i b_2^j b_3^k$, for $i + j + k = d$.

Given an integer $d \geq 0$, let $\mathbb{P}_d(T)$ be the space of all polynomials of degree $\leq d$ on T. Note that the barycentric coordinates b_1, b_2, b_3 of $\boldsymbol{v} \in T$ are all linear functions of the Cartesian coordinates, therefore, the set of Bernstein basis polynomials forms a basis for $\mathbb{P}_d(T)$. For a triangle T and coefficients $\{\theta_{T;ijk}\}$, any polynomial $\mathcal{P} \in \mathbb{P}_d(T)$ can be uniquely written as $\mathcal{P}(\boldsymbol{v})|_T = \sum_{i+j+k=d} \theta_{T;ijk}B^d_{T;ijk}(\boldsymbol{v})$

called the B-form of $\mathcal{P}$ relative to T. Let $\mathbb{C}^r(\Omega)$ be the space of rth continuously differentiable functions over the domain Ω. Given $0 \leq r < d$ and a triangulation $\triangle$, the spline space of degree d and smoothness r over $\triangle$ is defined as

$$\mathbb{S}_d^r(\triangle) = \{\mathcal{P} \in \mathbb{C}^r(\Omega) : \mathcal{P}|_{T_i} \in \mathbb{P}_d(T_m), T_m \in \triangle, m = 1, \ldots, M\}. \tag{5}$$

For triangulation $\triangle$ with M triangles, denote a set of bivariate Bernstein basis polynomials for $\mathbb{S}_d^r(\triangle)$ as $\{B_m\}_{m\in\mathcal{M}}$, where $\mathcal{M}$ is an index set for basis functions on triangulation $\triangle$ with cardinality $|\mathcal{M}| = M(d+1)(d+2)/2$. Then, we can approximate the bivariate functions $\beta_{\ell t} \in \mathbb{S}_d^r(\triangle)$ in the STEM model (2) by $\sum_{m\in\mathcal{M}} B_m(\boldsymbol{u})\theta_{\ell tm} = \mathbf{B}(\boldsymbol{u})^\top \boldsymbol{\theta}_{\ell t}$, where at a location point $\boldsymbol{u}$, $\mathbf{B}(\boldsymbol{u}) = \{B_m(\boldsymbol{u}), m \in \mathcal{M}\}^\top$ and $\boldsymbol{\theta}_{\ell t} = \{\theta_{\ell tm}, m \in \mathcal{M}\}^\top$ are the vector of bivariate basis functions and the corresponding spline coefficient vector at a time point t, respectively.

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. The bivariate spline basis are generated via the R package "BPST" (Wang et al., 2019).

Considering the basis expansion, for the current time t, the maximization problem (3) is changed to minimize

$$-\sum_{i=1}^{n}\sum_{s=t-t_0}^{t} L\left(g^{-1}\left[\mathbf{B}(\mathbf{U}_i)^\top\{\boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 \log(I_{i,s-1})\} + \alpha_0 Z_{i,s-1} + \sum_{j=1}^{p}\alpha_j A_{ij,s-r} + \sum_{k=1}^{q}\boldsymbol{\Phi}_k^\top(X_{ik})\boldsymbol{\xi}_k\right], Y_{is}\right) + \frac{1}{2}(\lambda_0\boldsymbol{\theta}_0^\top\mathbf{P}\boldsymbol{\theta}_0 + \lambda_1\boldsymbol{\theta}_1^\top\mathbf{P}\boldsymbol{\theta}_1) \text{ subject to } \mathbf{H}\boldsymbol{\theta}_\ell = \mathbf{0}, \ell = 0, 1. \tag{6}$$

In addition, we consider the energy functional $\mathcal{E}(\beta_\ell)$ in (4) can be approximated by $\mathcal{E}(\mathbf{B}^\top\boldsymbol{\theta}_\ell) = \boldsymbol{\theta}_\ell^\top\mathbf{P}\boldsymbol{\theta}_\ell$, for $\ell = 0, 1$, where $\mathbf{P}$ is the block diagonal penalty matrix. Introducing the constraint matrix $\mathbf{H}$ which satisfies $\mathbf{H}\boldsymbol{\theta}_\ell = \mathbf{0}$, $\ell = 0, 1$, is a common strategy to reflect global smoothness in $\mathbb{S}_d^r(\triangle)$ in (5).

Directly solving the optimization problem in (6) is not straightforward due to the smoothness constraints inside. Instead, suppose that the rank r 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 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 $\boldsymbol{\theta}_\ell = \mathbf{Q}_2\boldsymbol{\theta}_\ell^*$ for some $\boldsymbol{\theta}_\ell^*$, $\ell = 0, 1$, enforces $\mathbf{H}\boldsymbol{\theta}_\ell = \mathbf{0}$. Thus, the constraint problem in (6) can be changed to an unconstrained optimization

problem as follows:

$$-\sum_{i=1}^{n}\sum_{s=t-t_0}^{t} L\left(g^{-1}\left[\mathbf{B}(\mathbf{U}_i)^\top\mathbf{Q}_2\{\boldsymbol{\theta}_0^*+\boldsymbol{\theta}_1^*\log(I_{i,s-1})\}+\alpha_0 Z_{i,s-1}+\sum_{j=1}^{p}\alpha_j A_{ij,s-r}\right.\right.$$
$$\left.\left.+\sum_{k=1}^{q}\boldsymbol{\Phi}_k^\top(X_{ik})\boldsymbol{\xi}_k\right], Y_{is}\right)+\frac{1}{2}\left(\lambda_0\boldsymbol{\theta}_0^{*\top}\mathbf{Q}_2^\top\mathbf{P}\mathbf{Q}_2\boldsymbol{\theta}_0^*+\lambda_1\boldsymbol{\theta}_1^{*\top}\mathbf{Q}_2^\top\mathbf{P}\mathbf{Q}_2\boldsymbol{\theta}_1^*\right). \quad (7)$$

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 maximizers of (7) at time point t. We obtain the estimators of $\boldsymbol{\beta}_{\ell t}(\cdot)$:

$$\widehat{\boldsymbol{\beta}}_{\ell t}(\boldsymbol{u})=\mathbf{B}(\boldsymbol{u})^\top\mathbf{Q}_2\widehat{\boldsymbol{\theta}}_{\ell t}^*,\ \ell=0,1,$$

the estimator of α_{jt} is $\widehat{\alpha}_{jt}$, $j=1,\ldots,p$, and the spline estimator $\gamma_{kt}(\cdot)$ is $\widehat{\gamma}_{kt}(x_k)=\boldsymbol{\Phi}_k(x_k)^\top\widehat{\boldsymbol{\xi}}_{kt}$, $k=1,\ldots,q$.

4.2 A Penalized Iteratively Reweighted Least Squares Algorithm

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 $\boldsymbol{\Phi}_i^\top=\{\boldsymbol{\Phi}_1(X_{i1})^\top,\cdots,\boldsymbol{\Phi}_q(X_{iq})^\top\}$, $\mathbf{A}_{is}^\top=(A_{i1,s-r},\cdots,A_{ip,s-r})$, and $\mathbf{F}=(\mathbf{F}_1,\ldots,\mathbf{F}_t)^\top$, where $\mathbf{F}_s=(\mathbf{F}_{1s},\cdots,\mathbf{F}_{ns})$, and $\mathbf{F}_{is}^\top=(\mathbf{A}_{is}^\top,\ \boldsymbol{\Phi}_i^\top,\ [\{1,\log(I_{i,s-1})\}^\top\otimes\mathbf{B}^*(\mathbf{U}_i)]^\top)$ and $\mathbf{B}^*(\mathbf{U}_i)=\mathbf{Q}_2^\top\mathbf{B}(\mathbf{U}_i)$. Let $\eta_{is}(\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\theta}^*)=\mathbf{B}^*(\mathbf{U}_i)^\top\{\boldsymbol{\theta}_0^*+\boldsymbol{\theta}_1^*\log(I_{i,s-1})\}+\alpha_0 Z_{i,s-1}+\sum_{j=1}^{p}\alpha_j A_{ij,s-r}+\sum_{k=1}^{q}\boldsymbol{\Phi}_k^\top(X_{ik})\boldsymbol{\xi}_k$, and $\boldsymbol{\eta}(\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\theta}^*)=\{\eta_{is}\}_{i=1,s=1}^{n,t}$. In addition, let the mean vector $\boldsymbol{\mu}(\boldsymbol{\beta}^*)=\{\mu_{is}\}_{i,s=1}^{n,t}=\{g^{-1}(\eta_{is})\}_{i,s=1}^{n,t}$, the variance function matrix $\mathbf{V}=\mathrm{diag}\{V(\mu_{is})\}_{i,s=1}^{n,t}$, the diagonal matrix $\mathbf{G}=\mathrm{diag}\{g'(\mu_{is})\}_{i,s=1}^{n,t}$ with the derivative of link function as element, and the weight matrix $\mathbf{W}=\mathrm{diag}[\{V(\mu_{is})g'(\mu_{is})^2\}^{-1}w_{st},i=1,\ldots,n,s=1,\ldots,t]$, where $w_{st}=I(t-s\geq t_0)$.

In order to numerically solve the minimization in (7), we design the penalized iteratively reweighted least squares (PIRLS) algorithm as described below. Suppose at the jth iteration, we have $\boldsymbol{\mu}^{(j)}=\boldsymbol{\mu}(\boldsymbol{\alpha}^{(j)},\boldsymbol{\xi}^{(j)},\boldsymbol{\theta}^{*(j)})$, $\boldsymbol{\eta}^{(j)}=\boldsymbol{\eta}(\boldsymbol{\alpha}^{(j)},\boldsymbol{\xi}^{(j)},\boldsymbol{\theta}^{*(j)})$ and $\mathbf{V}^{(j)}$. Then at $(j+1)$th iteration, we consider the following objective function:

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

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

$$L_P^{(j+1)}\approx\left\|\left\{\mathbf{V}^{(j)}\right\}^{-1/2}\left[\mathbf{Y}-\boldsymbol{\mu}^{(j)}-\{\mathbf{G}^{(j)}\}^{-1}\mathbf{F}\left\{\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\xi}\\ \boldsymbol{\theta}^*\end{pmatrix}-\begin{pmatrix}\boldsymbol{\alpha}^{(j)}\\ \boldsymbol{\xi}^{(j)}\\ \boldsymbol{\theta}^{*(j)}\end{pmatrix}\right\}\right]\right\|^2+\frac{1}{2}\sum_{\ell=0}^{1}\lambda_\ell\boldsymbol{\theta}_\ell^{*\top}\mathbf{Q}_2^\top\mathbf{P}\mathbf{Q}_2\boldsymbol{\theta}_\ell^*$$
$$=\left\|\left\{\mathbf{W}^{(j)}\right\}^{1/2}\left[\widetilde{\mathbf{Y}}^{(j)}-\mathbf{F}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\xi}\\ \boldsymbol{\theta}^*\end{pmatrix}\right]\right\|^2+\frac{1}{2}\sum_{\ell=0}^{1}\lambda_\ell\boldsymbol{\theta}_\ell^{*\top}\mathbf{Q}_2^\top\mathbf{P}\mathbf{Q}_2\boldsymbol{\theta}_\ell^*, \quad (8)$$

Algorithm 1 The PIRLS Algorithm.

Step 1. Start with the initial values $\boldsymbol{\eta}^{(0)}$ and $\boldsymbol{\mu}^{(0)}$. Calculate weight matrix $\mathbf{W}^{(0)}$ and working variable $\widetilde{\mathbf{Y}}^{(0)}$ 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 $\{\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\theta}^*\}$ *not converge* **do**

(i) Obtain $\boldsymbol{\alpha}^{(j+1)}, \boldsymbol{\xi}^{(j+1)}, \boldsymbol{\theta}^{*(j+1)}$ by minimizing the (8) with respect to $\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\theta}^*$, and update $\boldsymbol{\eta}^{(j+1)} = \boldsymbol{\eta}(\boldsymbol{\alpha}^{(j+1)}, \boldsymbol{\xi}^{(j+1)}, \boldsymbol{\theta}^{*(j+1)})$ and $\boldsymbol{\mu}^{(j+1)} = \boldsymbol{\mu}(\boldsymbol{\alpha}^{(j+1)}, \boldsymbol{\xi}^{(j+1)}, \boldsymbol{\theta}^{*(j+1)})$.

(ii) Update $\mathbf{W}^{(j+1)}$ 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 $\boldsymbol{\eta}^{(j+1)}$ and $\boldsymbol{\mu}^{(j+1)}$.

(iii) Set $j = j + 1$.

end

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 PIRLS procedure is represented in Algorithm 1. In the numerical analysis, we set $\mu_{is}^{(0)} = Y_{is} + 0.1$ and $\eta_{is}^{(0)} = g(\mu_{is}^{(0)})$ as the initial values to start the iteration.

4.3 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}, while such kind of data tends to suffer from missingness, large error and inconsistency due to its difficulty in data collection; see the discussions in KCRA (2020).

Based on the fact that $I_{it} = C_{it} - R_{it} - D_{it}$, we attempt to modeling D_{it} and R_{it} in order to facilitate the estimation and prediction of newly confirmed cases Y_{it} based on the proposed STEM model. Let $\Delta D_{it} = D_{it} - D_{i,t-1}$ be the new fatal cases on day t, and following similar notations in the STEM model (2), we assume that

$$\Delta D_{it} | \mathbf{X}_i, \mathbf{U}_i, I_{i,t-1}, \mathbf{A}_{i,t-r} \ \sim \ \text{Poisson}(\mu_{it}^{\mathrm{D}}), \tag{9}$$

where

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

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, there are no uniform criteria to collect recovery reports across the U.S. (CNN, 2020). According to the U.S. Centers for Disease Control and Prevention, severe cases with COVID-19 often require medical care and receive supportive care in the hospital. At the same time, in general, most people with the mild illness are not hospitalized and suggested to recover at home. Currently, only a few states regularly update the number of recovered patients, but seldom can the counts be mapped to counties.

Due to the 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 (Anastassopoulou et al., 2020; Siettos and Russo, 2013). At current time point t, we assume that $\Delta R_{is} = \nu_t I_{i,s-1} + \varepsilon_{is},\ s = t-t_0, \ldots, t$, in which the recovery rate ν_t 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 can be either estimated from available state-level data, or obtained from prior medical studies due to the under-reporting issue in actual data.

4.4 Zero-inflated Models at the Early Stage of the Outbreak

Early in an epidemic, the quality of data on infections, deaths, tests, and other factors often are limited by underdetection or inconsistent detection of cases, reporting delays, and poor documentation, all of which affect the quality of any model output. 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, which allows for frequent zero-valued observations. Following the works by Arab et al. (2012), Beckett et al. (2014) and Wood et al. (2016), we assume the observed counts Y_{it} contributes to a zero-inflated Poisson distribution

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

where μ_{it} follows (2), and $p_{it} = \text{logit}(\eta_{it})$ with $\eta_{it} = a_1 + \{b + \exp(a_2)\} \log(\mu_{it})$. Here we take $b = 0$ and a_1, a_2 are estimated with the roughness parameters. See Wood et al. (2016) for more details in the estimation of a_1 and a_2.

Similarly, we also consider zero-inflated models, in which we assume the observed count ΔD_{it} contributes to a zero-inflated Poisson distribution

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

where μ_{it}^{D} follows (9), $p_{it}^{\mathrm{D}} = \text{logit}(\eta_{it})$, and $\eta_{it}^{\mathrm{D}} = v_1 + \{b + \exp(v_2)\} \log(\mu_{it}^{\mathrm{D}})$ with $b = 0$ and (v_1, v_2) estimated in a parallel fashion to (a_1, a_2).

5 Forecast and Band of the Forecast Path

To understand the impact of COVID-19, it requires accurate forecast for the spread of infectious cases along with analysis of the number of death and recovery cases. In this section, we describe our prediction procedure of these counts, specifically, we are interested in predicting Y_{it}, I_{it} and D_{it}. We also provide the prediction intervals to quantify the uncertainty of the prediction.

We consider an h-step ahead prediction. As described in Section 3, if we observe $C_{is}, I_{is}, R_{is}, D_{is}$ for $s = 1, \ldots, t$, then the infection model and fatal cases model can be fitted by regressing $\{Y_{is}\}_{i=1,s=t-t_0}^{n,t}$, $\{\Delta D_{is}\}_{i=1,s=t-t_0}^{n,t}$ on $\{I_{i,s-1}, Z_{i,s-1}, \mathbf{A}_{i,s-r}, \mathbf{X}_i\}_{i=1,s=1}^{n,t}$, respectively. The predictions of infectious count at time $t+1$ and iteratively at $t+h$ are

$$\widehat{Y}_{i,t+1} = \exp\left\{\widehat{\beta}_{0t}(\mathbf{U}_i) + \widehat{\beta}_{1t}(\mathbf{U}_i)\log(I_{it}) + \widehat{\alpha}_{0t}Z_{it} + \sum_{j=1}^{p}\widehat{\alpha}_{jt}A_{ij,t+1-r} + \sum_{k=1}^{q}\widehat{\gamma}_{kt}(X_{ik})\right\}, \tag{10}$$

$$\widehat{Y}_{i,t+h} = \exp\left\{\widehat{\beta}_{0t}(\mathbf{U}_i) + \widehat{\beta}_{1t}(\mathbf{U}_i)\log(\widehat{I}_{i,t+h-1}) + \widehat{\alpha}_{0t}\widehat{Z}_{i,t+h-1} + \sum_{j=1}^{p}\widehat{\alpha}_{jt}A_{ij,t+h-r} + \sum_{k=1}^{q}\widehat{\gamma}_{kt}(X_{ik})\right\},$$

respectively, where $\widehat{I}_{i,t+h-1} = I_{it} + \sum_{s=t+1}^{t+h-1}\widehat{Y}_{is} - \widehat{R}_{i,t+h-1} - \widehat{D}_{i,t+h-1}$ and $\widehat{Z}_{i,t+h-1} = \log(N_i - C_{i,t+h-1}) - \log(N_i)$. Meanwhile, let

$$\widehat{\Delta D}_{i,t+h} = \exp\left\{\widehat{\beta}^{\mathrm{D}}_{0t}(\mathbf{U}_i) + \widehat{\beta}^{\mathrm{D}}_{1t}(\mathbf{U}_i)\log(I_{i,t+h-1}) + \sum_{j=1}^{p}\widehat{\alpha}^{\mathrm{D}}_{jt}A_{ij,t+h-r} + \sum_{k=1}^{q}\widehat{\gamma}^{\mathrm{D}}_{kt}(X_{ik})\right\},$$

and $\widehat{\Delta R}_{i,t+h} = \widehat{\nu}\widehat{I}_{i,t+h-1}$, where we predict $R_{i,t+h}$ by $\widehat{R}_{i,t+h} = R_{it} + \sum_{s=t+1}^{t+h}\widehat{\Delta R}_{i,s}$, and $D_{i,t+h}$ by $\widehat{D}_{i,t+h} = D_{it} + \sum_{s=t+1}^{t+h}\widehat{\Delta D}_{i,s}$. Then, the predicted number of active cases and susceptible cases are $\widehat{I}_{i,t+h} = C_{i,t+h-1} + \widehat{Y}_{i,t+h} - \widehat{R}_{i,t+h} - \widehat{D}_{i,t+h}$, and $\widehat{S}_{i,t+h} = N_i - (C_{i,t+h-1} + \widehat{Y}_{i,t+h})$. The above one-step predicted values can be thus plugged back into equation (10) to obtain the predictions for the following days by repeating the same procedure.

There is substantial interest in the problem of how to quantify the uncertainty for the forecasts with a succession of periods. To construct the band for forecast path $\{Y_{i,t+h}, h = 1, \ldots, H\}$, we consider the bootstrap method (Staszewska-Bystrova, 2009), in which the bootstrap samples are generated using the bias-corrected bootstrap procedure; see Algorithms 2 and 3 for the details.

6 Analysis and Findings

In this section, we present our analysis results and findings for the COVID-19 study.

6.1 Estimation and Inference Results

For the model estimation, we consider the data collected from March 23 to April 25. Based on the data described in Section 2, we consider the following model for the infection count:

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

Algorithm 2 A bootstrap procedure to correct the bias.

Step 1. Fit models (2) and (9) using $(Y_{is}, I_{i,s-1}, Z_{i,s-1}, \mathbf{A}_{i,s-r}, \mathbf{X}_i, \mathbf{U}_i)_{i=1,s=1}^{n,t}$ and $(\Delta D_{is}, I_{i,s-1}, \mathbf{A}_{i,s-r}, \mathbf{X}_i, \mathbf{U}_i)_{i=1,s=1}^{n}$, obtain $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\gamma}}$, $\widehat{\boldsymbol{\beta}}^{\mathrm{D}}$ $\widehat{\boldsymbol{\alpha}}^{\mathrm{D}}$, $\widehat{\boldsymbol{\gamma}}^{\mathrm{D}}$.

Step 2. Generate bootstrap samples to correct the bias in the estimator of the coefficients.

foreach $1 \le b \le B$ **do**

(i) Generate the bootstrap sample as follows.

foreach $1 \le s \le t$ **do**

Generate $Y_{is}^b \sim \text{Poisson}(\widehat{\mu}_{is})$, $\Delta D_{is}^b \sim \text{Poisson}(\widehat{\mu}_{is}^{\mathrm{D}})$, and $\Delta R_{is}^b \sim \text{Poisson}(\widehat{\mu}_{is}^{R})$, where

$$\widehat{\mu}_{is} = \exp\{\widehat{\beta}_0(\mathbf{U}_i) + \widehat{\beta}_1(\mathbf{U}_i)\log(I_{i,s-1}) + \widehat{\alpha}_0 Z_{i,s-1} + \sum_{j=1}^{p}\widehat{\alpha}_j A_{ij,s-r} + \sum_{k=1}^{q}\widehat{\gamma}_k(X_{ik})\},$$

$$\widehat{\mu}_{is}^{\mathrm{D}} = \exp\{\widehat{\beta}_0^{\mathrm{D}}(\mathbf{U}_i) + \widehat{\beta}_1^{\mathrm{D}}(\mathbf{U}_i)\log(I_{i,s-1}) + \sum_{j=1}^{p}\widehat{\alpha}_j^{\mathrm{D}} A_{ij,s-r} + \sum_{k=1}^{q}\widehat{\gamma}_k^{\mathrm{D}}(X_{ik})\},$$

$$\widehat{\mu}_{is}^{R} = \widehat{\nu} I_{i,s-1}.$$

Update $Z_{is}^b = \log(S_{is}^b/N_i)$, where $S_{is}^b = S_{i,s-1}^b - Y_{is}^b$ and $I_{is}^b = I_{i,s-1} + Y_{is}^b - \Delta D_{is}^b - \Delta R_{is}^b$.

end

(ii) Fit the models (2) and (9) based on $(Y_{is}^b, I_{i,s-1}^b, Z_{i,s-1}, \mathbf{A}_{i,s-1}, \mathbf{X}_i, \mathbf{U}_i)_{i=1,s=1}^{n,t}$ and $(\Delta D_{is}^b, I_{i,s-1}^b, \mathbf{A}_{i,s-1}, \mathbf{X}_i, \mathbf{U}_i)_{i=1,s=1}^{n,t}$, respectively, and obtain $(\widehat{\boldsymbol{\beta}}^b, \widehat{\boldsymbol{\alpha}}^b, \widehat{\boldsymbol{\gamma}}^b)$ and $(\widehat{\boldsymbol{\beta}}^{\mathrm{D},b}, \widehat{\boldsymbol{\alpha}}^{\mathrm{D},b}, \widehat{\boldsymbol{\gamma}}^{\mathrm{D},b})$.

end

Step 3. Calculate the bias of the coefficients based on the above bootstrap samples. For example, for $\ell = 0, 1$, let $\text{bias}(\widehat{\beta}_\ell) = B^{-1}\sum_{b=1}^{B}\widehat{\beta}_\ell^b - \widehat{\beta}_\ell$, and let $\widehat{\beta}_\ell^{\mathrm{c}} = \widehat{\beta}_\ell - \text{bias}(\widehat{\beta}_\ell)$ be the corrected coefficient function. Similarly, we obtain the bias-corrected coefficients of $\widehat{\boldsymbol{\alpha}}_t$ and $\widehat{\boldsymbol{\gamma}}_t$, denoted by $\widehat{\boldsymbol{\alpha}}_t^{\mathrm{c}}$, $\widehat{\boldsymbol{\gamma}}_t^{\mathrm{c}}$, respectively.

Algorithm 3 A bootstrap procedure to calculate the prediction band.

Step 1. Generate bootstrap samples to construct prediction band.

foreach $1 \le b \le B$ **do**

  foreach $1 \le h \le H$ **do**

    Generate $Y^b_{i,t+h} \sim$ Poisson$(\widehat{\mu}^{\mathrm{c},b}_{i,t+h})$, $\Delta D^b_{i,t+h} \sim$ Poisson$(\widehat{\mu}^{\mathrm{D},\mathrm{c},b}_{i,t+h})$, and $\Delta R^b_{i,t+h} \sim$ Poisson$(\widehat{\mu}^{R,\mathrm{c},b}_{i,t+h})$ based on bootstrap estimators, where

    $$\widehat{\boldsymbol{\beta}}^{\mathrm{c},b} = 2\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^b,\ \widehat{\boldsymbol{\alpha}}^{\mathrm{c},b} = 2\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^b,\ \widehat{\boldsymbol{\gamma}}^{\mathrm{c},b} = 2\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}^b,$$

    $$\widehat{\boldsymbol{\beta}}^{\mathrm{D},\mathrm{c},b} = 2\widehat{\boldsymbol{\beta}}^{\mathrm{D}} - \widehat{\boldsymbol{\beta}}^{\mathrm{D},b},\ \widehat{\boldsymbol{\alpha}}^{\mathrm{D},\mathrm{c},b} = 2\widehat{\boldsymbol{\alpha}}^{\mathrm{D}} - \widehat{\boldsymbol{\alpha}}^{\mathrm{D},b},\ \widehat{\boldsymbol{\gamma}}^{\mathrm{D},\mathrm{c},b} = 2\widehat{\boldsymbol{\gamma}}^{\mathrm{D}} - \widehat{\boldsymbol{\gamma}}^{\mathrm{D},b},$$

    $$\widehat{\mu}^{\mathrm{c},b}_{i,t+h} = \exp\{\widehat{\beta}^{\mathrm{c},b}_0(\mathbf{U}_i) + \widehat{\beta}^{\mathrm{c},b}_1(\mathbf{U}_i)\log(I_{i,t+h-1}) + \widehat{\alpha}^{\mathrm{c},b}_0 Z_{i,t+h-1} + \sum_{j=1}^{p}\widehat{\alpha}^{\mathrm{c},b}_j A_{ij,t+h-r} + \sum_{k=1}^{q}\widehat{\gamma}^{\mathrm{c},b}_k(X_{ik})\},$$

    $$\widehat{\mu}^{\mathrm{D},\mathrm{c},b}_{i,t+h} = \exp\{\widehat{\beta}^{\mathrm{D},\mathrm{c},b}_0(\mathbf{U}_i) + \widehat{\beta}^{\mathrm{D},\mathrm{c},b}_1\log(I_{i,t+h-1}) + \sum_{j=1}^{p}\widehat{\alpha}^{\mathrm{D},\mathrm{c},b}_j A_{ij,t+h-r} + \sum_{k=1}^{q}\widehat{\gamma}^{\mathrm{D},\mathrm{c},b}_k(X_{ik})\},$$

    $$\widehat{\mu}^{R,\mathrm{c},b}_{i,t+h} = \widehat{\nu} I_{i,t+h-1}.$$

    Update $Z^b_{i,t+h} = \log(S^b_{i,t+h}/N_i)$, where $S^b_{i,t+h} = S^b_{i,t+h-1} - Y^b_{i,t+h}$ and $I^b_{i,t+h} = I^b_{i,t+h-1} + Y^b_{i,t+h} - \Delta D^b_{i,t+h} - \Delta R^b_{i,t+h}$.

  end

end

Step 2. Construct the $100(1-\alpha)\%$ prediction band by the above B bootstrap paths with the most extreme αB paths discarded. Start with setting $\kappa = 0$.

while $\kappa < \alpha B$ **do**

  (i) For each forecast time point $h = 1, \ldots, H$ (there are in total $B - \kappa$ constructed paths available), identify the largest and the smallest bootstrap forecast values, and the associated paths. Notice there are $2H$ extreme values and at most corresponding $2H$ paths.

  (ii) Compute the distances from each of the bootstrap path (at most $2H$) to the bootstrap sample, based on: $\sum_{h=1}^{H}(\widehat{\mu}^{\mathrm{c}}_{i,t+h} - Y^b_{i,t+h})^2$ or $\sum_{h=1}^{H}|\widehat{\mu}^{\mathrm{c}}_{i,t+h} - Y^b_{i,t+h}|$.

  (iii) Remove the path with the largest distance, and set $\kappa = \kappa + 1$.

end

Step 3. Obtain the $100(1-\alpha)\%$ prediction band from the envelope of the remaining $(1-\alpha)B$ bootstrap paths.

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

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

We use 14 days as an estimation window to examine how the covariates affect the new infected cases and fatal cases. The roughness parameters are selected by the generalized cross-validation (GCV). The performance of the univariate/bivariate splines is dependent upon the choice of the knots/triangulation. Knots selection and triangulation selection are one of the key ingredients for obtaining satisfactory results. We use cubic splines with 2 interior knots for the univariate spline smoothing. We generate the triangulations according to "max-min" criterion, which maximizes the minimum angle of all the angles of the triangles in the triangulation. Figure 2 shows the triangulations adopted by our method: $\triangle_1$ (119 triangles with 87 vertices) and $\triangle_2$ (522 triangles with 306 vertices). 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 $\beta_0(\cdot)$ and $\beta_1(\cdot)$, we adopt the finer triangulation $\triangle_2$, and use the rough triangulation $\triangle_1$ to estimate $\beta_0^{\mathrm{D}}(\cdot)$.

6.1.1 Estimation and inference for the infection model

First, we report our findings from modeling the infection count using model (11). To examine the effect of the control measures ("shelter-in-place" or "stay-at-home" order) after 7 days, we test the hypothesis: $H_0 : \alpha_{2t} = 0$ in model (11). We found that the p-values are smaller than 0.0001 at almost all the time points.

The estimated coefficient functions of $\beta_{0t}(\cdot)$ and $\beta_{1t}(\cdot)$ in model (11) using the data from March 23 to April 25, 2020, are shown in the supplementary materials (Wang et al., 2020c). We can see that the transmission rate varies at different locations and in different phases of the outbreak, and $\beta_{1t}(\cdot)$ is also varying, which indicates that the homogeneous mixing assumption of the simple SIR models does not hold. Transmission rate is high in the majority of states at the end of March, however, in many states, it becomes much lower in the middle of April or late of April.

Next, we examine the effect of the predictors and test the following hypothesis of the individual functions $H_0 : \gamma_{kt}(\cdot) = 0$, $k = 1, \ldots, 12$. The figures on Pages 2–14 in the supplementary materials (Wang et al., 2020c) show the estimate and the SCB of the nonparametric functions γ_{kt}, $k = 1, \ldots, 12$, at different time point in the STEM model (11).

From these figures, we can find the effect of the county-level predictors on the spread. Healthcare coverage is essential for a person's health status, and sometimes, a self-selection process. After controlling to social-economic factors, the percent of persons under 65 years without health insurance has a significant impact on the COVID-19 breakout in the community. We can observe a sharp increasing

pattern between the non-healthy-coverage rate and the COVID-19 infection rate. An under-covered population is much easier to be infected with the virus. Because there are more uninsured people in the urban area, an increasing pattern is observed in the urban rate impact analysis. "PD" is often considered to have a linear relationship with COVID-19 infection cases in most studies and news reports. Our results are consistent with the intuition. The higher the "PD" is, the higher the logarithm of new COVID-19 cases is. The local healthcare expenditure, "EHPC", has a similar impact on COVID-19 infections. The elderly population's impact pattern is an inverse U-shape. This pattern is because they are easier to be infected. However, when the older population dominates the community, people are less active and more risk-averse, thus stay home more often, so that it hinders the spread of the virus.

6.1.2 Estimation and inference for the death model

We report our findings from estimating the death count using model (12).

To examine the effect of the infection count, control measures and the county-level predictors, we test the following hypothesis: $H_0 : \beta_{1t}^{\mathrm{D}} = 0$, $H_0 : \alpha_{2t}^{\mathrm{D}} = 0$ and $H_0 : \gamma_{kt}^{\mathrm{D}} = 0$, $k = 1, \ldots, 12$, in model (12). Figure 1 plots the p-values of the above tests. From Figure 1, we find that the "Infection", "AA", "HL", "Disadvantage", and "Old" are very significant with p-values are smaller than 0.05 all the time. The rest of the predictors are significant on some days, but insignificant on other days.

Page 17 on the supplementary materials (Wang et al., 2020c) shows the pattern of $\widehat{\beta}_{0t}^{\mathrm{D}}(\cdot)$ in model (12). From this animation, we observe a general decrease pattern in the entire U.S. from March 23 to April 25, 2020.

6.2 Forecasting Performance and Results

In this section, we investigate the short-term prediction performance of the proposed method. In the following, we consider h-day ahead prediction based on the forecasting method described in Section 5. An R shiny app (Wang et al., 2020a) is developed to provide a 7-day forecast of COVID-19 infection and death count at both the county level and state level, in which the state level forecast is obtained by aggregating forecasts across counties in each state. This app was launched on 03/27/2020 for displaying results of our forecasting.

We demonstrate the accuracy of the STEM for h-day ahead predictions, $h = 1, \ldots, 7$. For comparison, we also consider the two naive models that assume a linear or exponential growth pattern for total confirmed cases for each county:

- (Linear) $\mathrm{E}(C_{it}|t) = \beta_{i0} + \beta_{i1}t$, $\mathrm{Var}(C_{it}|t) = \sigma_i^2$, $i = 1, \ldots, n$;
- (Exponential, Poisson) $\log\{\mathrm{E}(C_{it}|t)\} = \beta_{i0} + \beta_{i1}t$, $\mathrm{Var}(C_{it}|t) = \exp(\beta_{i0} + \beta_{i1}t)$, $i = 1, \ldots, n$;

and the following simple epidemic method (EM):

- (EM) $\log(\mu_{it}) = \beta_0 + \beta_1 \log(I_{i,t-1})$, $\log(\mu_{it}^{\mathrm{D}}) = \beta_0^{\mathrm{D}} + \beta_1^{\mathrm{D}} \log(I_{i,t-1})$, $i = 1, \ldots, n$.

We consider the data collected from March 23 to April 18. To predict the counts in the next 7 days, we use the previous 9 days as a training set for model fit. To show the accuracy of different methods, we compute the following root mean-squared prediction errors (RMSPEs):

$$R_h = T^{-1}\sum_{t=1}^{T}\left\{n^{-1}\sum_{i=1}^{n}(\widehat{Y}_{i,t+h} - Y_{i,t+h})^2\right\}^{1/2}, \ h = 1, \ldots, 7,$$

where $T = 18$.

Table 2 shows the average of the RMSPEs for h-day. From this table, we can see that our proposed method is much more accurate compared to all the other methods.

6.3 Findings from the Long-term Forecast

There has been an increasing public health concern regarding the adequacy of resources to treat infected cases. It is well known that hospital beds, intensive care units (ICU), and ventilators are critical for the treatment of patients with severe illness. To project the timing of the outbreak peak and the number of health resources required at a peak, in this section, we also provide the long-term forecast of the infection count and death count.

In Figure 3, we show the reported COVID-19 confirmed infectious cases and deaths, and the corresponding predicted counts for the next four months in the State of New York based on the observed data from April 16-22, 2020. Given the lack of reliable recovered data, we consider two different daily recovery rates: 0.10 and 0.15.

Based on our research results, we develop an R shiny app Wang et al. (2020b) to provide a forecast of COVID-19 infection count and death count for the next four months. The forecast for other states can be found from Wang et al. (2020b), which is updated every week.

7 Discussion

This work has aimed to bridge the gap between mathematical models and statistical analysis in the infectious disease study. In this paper, we created a state-of-art interface between mathematical models and statistical models for understanding and forecasting the dynamic pattern of the spread of infectious diseases. Our proposed model enhances the dynamics of the SIR mechanism by means of spatiotemporal analysis.

When it comes to analyzing the reported numbers of COVID-19 cases, other factors may also be responsible for temporal or spatial patterns. We investigated the spatial associations between the infection count, death count, and factors or characteristics of the counties across the U.S. by modeling the daily infected/fatal cases at the county level in consideration of the county-level factors. To examine spatial nonstationarity in transmission rate of the disease, we proposed a spatially varying coefficient model, which allows the transmission to vary from one area to another area. The proposed method can

be used as an important tool for understanding the dynamic of the disease spread, as well as to assess how this outbreak may unfold through time and space.

From our empirical studies, we found that our method provides a very accurate short-term forecast in the COVID-19 study. Since our model incorporates the epidemiological mechanism, it can also be used for long-term prediction. We also provided a projection band to quantify the uncertainty of the long-term forecast path.

Based on our results, a disease mapping can easily be implemented to illustrate high-risk areas, and thus help policy making 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 more realistic population structure.

Our paper did not take the under-reported issue into account. Assuming that the data used is reliable and that the future will continue to follow the past pattern of the disease, our forecasts suggest a continuing increase in the confirmed COVID-19 cases with sizable associated uncertainty. Our prediction method helps to understand where the state stands in combatting COVID-19 and give a sense of what to expect going forward. In predicting the future of the COVID-19 pandemic, many key assumptions have been based on limited data. Models may capture aspects of epidemics effectively while neglecting to account for other factors, such as the accuracy of diagnostic tests; whether immunity will wane quickly; and if reinfection could occur.

Data Availability Statement

- A full list of data citations are available by contacting the corresponding author.
- The R package "`STEM`" of the proposed method can be downloaded from the Github Repository: `https://github.com/covid19-dashboard-us/covid19`.
- The R shiny apps demonstrating the proposed methods can be found from `https://covid19.stat.iastate.edu/`.

Bibliography

Allen, L. J., Brauer, F., Van den Driessche, P., and Wu, J. (2008), *Mathematical epidemiology*, vol. 1945, Springer.

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," Available at `https://covidtracking.com/api`.

Beckett, S., Jee, J., Ncube, T., Pompilus, S., Washington, Q., Singh, A., and Pal, N. (2014), "Zero-inflated Poisson (ZIP) distribution: parameter estimation and applications to model data from natural calamities," *Involve, a Journal of Mathematics*, 7, 751–767.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., et al. (2020), "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study," *The Lancet*, 395, 507–513.

CNN (2020), "Most people recover from Covid-19. Here's why it's hard to pinpoint exactly how many," Available at `https://www.cnn.com/2020/04/04/health/recovery-coronavirus-tracking-data-explainer/index.html`.

CSSE, J. H. U. (2020), "2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository," Available at `https://github.com/CSSEGISandData/COVID-19`.

De Jong, M., Diekmann, O., and Heesterbeek, J. (1995), "How does the transmission depend on population size?" *Epidemic models: their structure and relation to data*, 5, 84.

Elmousalami, H. H. and Hassanien, A. E. (2020), "Day level forecasting for Coronavirus disease (COVID-19) spread: Analysis, modeling and recommendations," .

Fanelli, D. and Piazza, F. (2020), "Analysis and forecast of COVID-19 spreading in China, Italy and France," *Chaos, Solitons & Fractals*, 134.

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 (Applied Statistics)*, 49, 187–205.

KCRA (2020), "COVID-19: Why patient recovery data is scarce," Available at `https://www.kcra.com/article/covid-19-questions-recovery-numbers/32093456`.

Kim, M. and Wang, L. (2020), "Generalized spatially varying coefficient models," *Journal of Computational and Graphical Statistics*, accepted.

Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., and Eggo, R. M. (2020), "Early dynamics of transmission and control of COVID-19: A mathematical modelling study," *medRxiv*.

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

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*, 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.

NYT (2020), "Coronavirus (Covid-19) Data in the United States," Available at `https://github.com/nytimes/covid-19-data`.

Pan, A., Liu, L., Wang, C., Guo, H., Hao, X., Wang, Q., Huang, J., He, N., Yu, H., Lin, X., Wei, S., and Wu, T. (2020), "Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China," *JAMA*.

Pfeiffer, D., Robinson, T. P., Stevenson, M., Stevens, K. B., Rogers, D. J., Clements, A. C., et al. (2008), *Spatial analysis in epidemiology*, vol. 142, Oxford University Press Oxford.

Sangalli, L., Ramsay, J., and Ramsay, T. (2013), "Spatial Spline Regression Models," *Journal of the Royal Statistical Society B*, 75, 681–703.

Siettos, C. I. and Russo, L. (2013), "Mathematical modeling of infectious disease dynamics," *Virulence*, 4, 295–306.

Staszewska-Bystrova, A. (2009), "Bootstrap Confidence Bands for Forecast Paths," *Available at SSRN 1507451*.

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*, accepted.

Wakefield, J., Dong, T. Q., and Minin, V. N. (2019), "Spatio-temporal analysis of surveillance data," *Handbook of Infectious Disease Data Analysis*, 455–476.

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. Available at `https://github.com/funstatpackages/BPST`.

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

Wang, L., Wang, G., Gao, L., Li, X., Yu, S., Kim, M., and Wang, Y. (2020a), "An R shiny app to visualize, track, and predict real-time infected cases of COVID-19 in the United States," Available at `https://covid19.stat.iastate.edu/`.

Wang, L., Wang, G., Gao, L., Li, X., Yu, S., Kim, M., Wang, Y., and Gu, Z. (2020b), "An R Shiny App to predict the infected and death cases of COVID-19 in the U.S. in the next three months." Available at `https://covid19.stat.iastate.edu/longtermproj.html`.

— (2020c), "Supplementary materials for 'Spatiotemporal Dynamics, Nowcasting and Forecasting of COVID-19 in the United States'," Available at `https://faculty.sites.iastate.edu/lilywang/page/arxiv`.

Wang, L., Zhou, Y., He, J., Zhu, B., Wang, F., Lu, T., Eisenberg, M. C., and Song, P. X.-K. (2020d), "An Epidemiological Forecast Model and Software Assessing Interventions on COVID-19 Epidemic in China." *Journal of Data Science*, accepted.

Weiss, H. H. (2013), "The SIR model and the foundations of public health," *Materials matematics*, 01–17.

WHO (2020), "WHO Director-General's opening remarks at the media briefing on COVID-19 – 11 March 2020," Available at "https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-m

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.

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*, 1–27.

Zhang, Y., You, C., Cai, Z., Sun, J., Hu, W., and Zhou, X.-H. (2020), "Prediction of the COVID-19 outbreak based on a realistic stochastic model," *medRxiv*.

Table 1: County-level predictors used in the modeling.

Covariates	Description
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 ≥ 65 years) rate per capita
Sex	Ratio of male over female
Socioeconomic Status	
Affluence	Social affluence, a measure of more economically privileged areas, including: Percent of households with income over $75,000 Percent of adults obtaining bachelor's degree or higher Percent of employed persons in management, professional and related occupations Median value of owner-occupied housing units
Disadvantage	Concentrated disadvantage, a measure for conditions of economic disadvantage, including: Percent of households with public assistance income Percent of households with female householder and no husband present Civilian labor force unemployment rate
Gini	Gini coefficient, a measure of economic inequality and wealth distribution
Rural/urban Factor	
Urban	Urban rate
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
Policies	
Control_1	dummy variable for emergency declaration of state
Control_2	dummy variable for declaration of "shelter-in-place" or "stay-at-home" order
Geographic Information	
Lat, Lon	Latitude and longitude of the approximate geographic center of the county

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

Table 2: The average of root mean squared prediction errors (RMSPE_h) of the infection or death count, for the h-day ahead prediction, $h = 1, \ldots, 7$.

	Method	RMSPE_1	RMSPE_2	RMSPE_3	RMSPE_4	RMSPE_5	RMSPE_6	RMSPE_7
Infection	Linear	40.332	56.581	74.074	94.038	117.661	143.440	167.763
	Exponential	>1000	>1000	>1000	>1000	>1000	>1000	>1000
	EM	41.323	69.217	97.766	130.247	166.116	199.284	236.642
	STEM	35.632	56.097	74.460	94.121	118.569	141.809	168.008
Death	Linear	6.899	9.917	13.297	16.944	21.272	25.393	29.586
	Exponential	>1000	>1000	>1000	>1000	>1000	>1000	>1000
	EM	3.799	7.322	10.405	13.617	17.221	20.304	23.282
	STEM	3.755	7.200	10.287	13.535	17.208	20.529	23.868

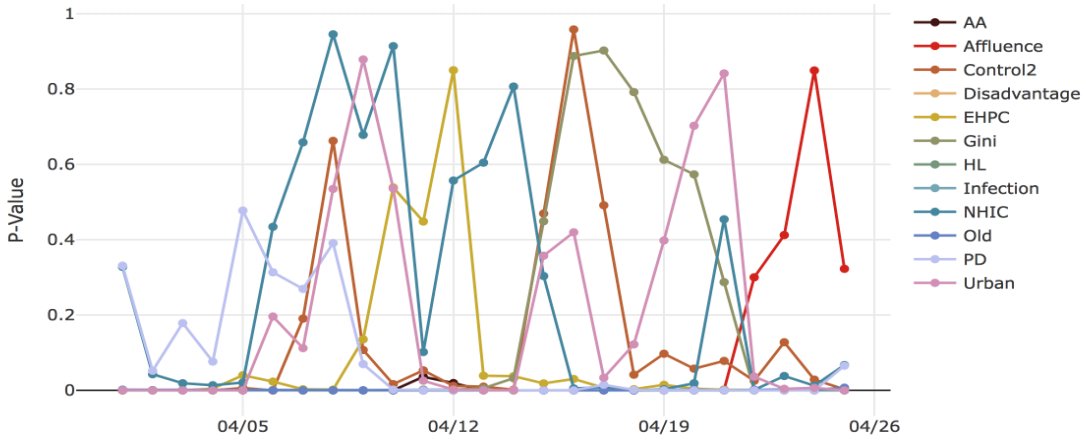
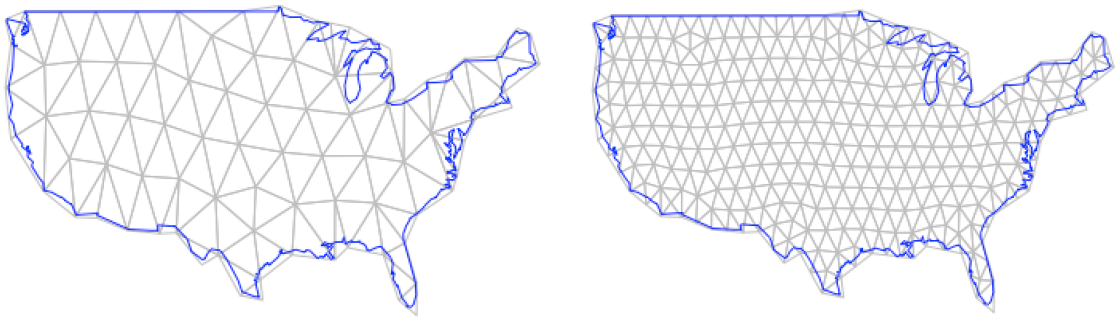
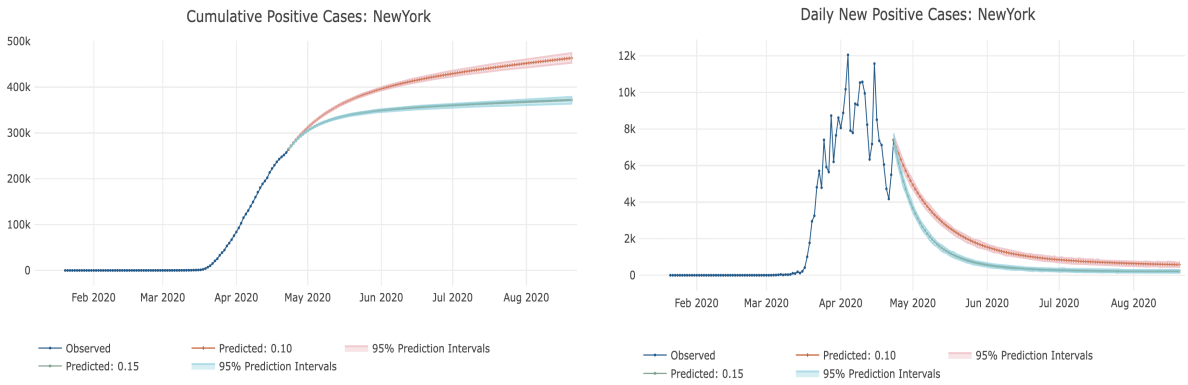


Figure 1: P-values of the hypothesis of the coefficient in the model (12).



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

Figure 2: Triangulations used in the bivariate spline estimation.



(a) Cumulative infection count (b) New infection count

Figure 3: Project of the cumulative and new infection count for State of New York in the next four months based on the observed data on April 16-22, 2020.