Supplementary materials for this article are available at https://doi.org/10.1007/s13253-022-00487-1.

A Spatiotemporal Analytical Outlook of the Exposure to Air Pollution and COVID-19 Mortality in the USA

Sounak CHAKRABORTY, Tanujit DEY, Yoonbae JUN, Chae Young LIM, Anish MUKHERJEE, and Francesca DOMINICI

The world is experiencing a pandemic due to Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), also known as COVID-19. The USA is also suffering from a catastrophic death toll from COVID-19. Several studies are providing preliminary evidence that short- and long-term exposure to air pollution might increase the severity of COVID-19 outcomes, including a higher risk of death. In this study, we develop a spatiotemporal model to estimate the association between exposure to fine particulate matter $PM_{2.5}$ and mortality accounting for several social and environmental factors. More specifically, we implement a Bayesian zero-inflated negative binomial regression model with random effects that vary in time and space. Our goal is to estimate the association between air pollution and mortality accounting for the spatiotemporal variability that remained unexplained by the measured confounders. We applied our model to four regions of the USA with weekly data available for each county within each region. We analyze the data separately for each region because each region shows a different disease spread pattern. We found a positive association between long-term exposure to $PM_{2.5}$ and the mortality from the COVID-19 disease for all four regions with three of four being statistically significant. Data and code are available at our GitHub repository.

Supplementary materials accompanying this paper appear on-line.

Key Words: Air pollution; Bayesian inference; COVID-19; Markov Chain Monte Carlo; Negative binomial model; Spatial; Spatiotemporal; Zero inflation.

Chae Young Lim, Sounak Chakraborty, Tanujit Dey, Yoonbae Jun these authors have contributed equally and Joined First Author.

S. Chakraborty, Department of Statistics, University of Missouri, Columbia, MO, USA (E-mail: *chakrabortys@missouri.edu*).
T. Dey (✉), Center for Surgery and Public Health, Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA (E-mail: *tdey@bwh.harvard.edu*).
Y. Jun · C.Y. Lim, Department of Statistics, Seoul National University, Gwanak-gu, Korea (E-mail: *junpeea@snu.ac.kr*) (E-mail: *twinwood@snu.ac.kr*).
A. Mukherjee, Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, USA (E-mail: *anish.mukherjee@louisville.edu*).
F. Dominici, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA (E-mail: *fdominic@hsph.harvard.edu*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 27, Number 3, Pages 419–439
https://doi.org/10.1007/s13253-022-00487-1

1. INTRODUCTION

The world is experiencing an enormous death toll from COVID-19. The number of COVID-19 cases and deaths vary spatially and temporally and can be affected by many factors, some local some global. There is a large body of literature that investigates the key biological, socioeconomic, and environmental factors that might increase the degree of severity of the health outcomes after having contacted COVID-19 (Karmakar et al. 2021; Webb Hooper et al. 2020; Yancy 2020; Abdelzaher et al. 2020; Ali and Islam 2020; Fiasca et al. 2020; Giani et al. 2020). Concerning the environmental factors, it is well established that short- and long-term exposure to air pollution increases the risk of several chronic diseases, including cardiovascular and respiratory diseases, irrespective of COVID-19 (Jiang et al. 2016; Lelieveld and Münzel 2019). We and others (Chakrabarty et al. 2020; Liu et al. 2020; Ogen 2020; Jiang and Xu 2021; Yongjian et al. 2020; Conticini et al. 2020; Comunian et al. 2020) have hypothesized that exposure to air pollution increases the severity of COVID-19 outcomes, because air pollution can affect our immune, respiratory and cardiovascular system. This is a rapidly evolving area of research, see, for example (Bhaskar et al. 2020) for a review of the epidemiological studies on this topic.

In this paper, we introduce a Bayesian spatiotemporal model to estimate the association between long-term exposure to $PM_{2.5}$ and COVID-19 health outcomes. To address this scientific question, we need to overcome several challenges. These include: 1) a large number of zero counts, especially at the beginning of the pandemic; 2) complex spatiotemporal variation that remained unexplained after having accounted for several measured confounders; and 3) computational feasibility. To overcome the challenges listed above and many others, we introduce a Bayesian model with multivariate spatiotemporal distributions of random effects while accounting for several measured socioeconomic and demographic factors. We modeled the COVID-19 death counts via a zero-inflated negative binomial (ZINB) distribution (Neelon et al. 2019). Since the frequentist approach to fitting the ZINB model is challenging for longitudinal, spatial, and spatiotemporal data, particularly when the model includes multivariate spatial random effects, we have chosen a more tractable Bayesian approach proposed by Neelon et al. (2019).

We apply our Bayesian model to a data set that includes weekly county-level death counts, air pollution levels and many other potential confounders from the USA. We incorporate the spatial and spatiotemporal information into the model by assigning a multivariate intrinsic conditionally autoregressive (ICAR) prior structure (Banerjee et al. 2014) to the random effects. Additional time-fixed effects are also considered. Then, we analyze the effectiveness of the zero-inflated model compared to an ordinary negative binomial model.

Wu et al. (2020) also considered a ZINB model at a county level to investigate the association between long-term exposure to $PM_{2.5}$ and COVID-19 deaths using an ecological and cross-sectional study design. These authors also considered state-specific random effects to capture variation between states. While Wu *et al.* (2020) investigated global association by using the data over most of US counties, no spatial dependence nor temporal dependence was assumed in the model, which is the main difference from our modeling. Instead, we decided to analyze spatiotemporal county-level data within regions consisting of multiple states that are geographically connected. This can account for heterogeneous dynamics of

the spread of disease across different regions of the USA. For our analysis, we consider four regions: Mid-Atlantic (New Jersey, New York, and Pennsylvania), Pacific (California, Oregon, and Washington), South Atlantic (Florida, Georgia, North Carolina, South Carolina), and Midwest (Iowa, Kansas, Missouri, Nebraska, North Dakota, and South Dakota). The results of our model show an overall positive association between long-term exposure to the ($PM_{2.5}$) and the mortality from the COVID-19 disease, which matches with other previous studies (Bhaskar et al. 2020).

The rest of the paper is organized as follows: In Sect. 2, we describe and visualize the data set for four regions in the USA. In Sect. 3, we present the methodology, including how we leverage the spatial and spatiotemporal information to estimate the COVID-19 spread. In this section, we also compare different statistical models. In Sect. 4, we present the results by applying the methods to the data set. A simulation study along with more details of the real data analysis is available in the Supplementary document. In Sect. 5, we discuss the results and comment on future directions.

2. DATA

2.1. COVID-19 DEATH COUNTS

We accessed the data from the repository maintained by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). We obtained daily number of deaths in each county from March 23, 2020, to August 31, 2020. Please note that the start date of the data source is March 22, 2020.

To investigate different scenarios of temporal and spatial dynamics of the COVID-19 deaths, we consider four regions: Mid-Atlantic (New Jersey, New York, and Pennsylvania), Pacific (California, Oregon, and Washington), South Atlantic (Florida, Georgia, North Carolina, South Carolina), and Midwest (Iowa, Kansas, Missouri, Nebraska, North Dakota, and South Dakota). Instead of considering the ratio of a COVID-19 count to a county-level population size, the county-level population size was considered as an offset variable in the model to remove the obvious effect of the county-level population size on the COVID-19 death counts. For spatiotemporal models, we computed 23 weekly COVID-19 death counts from March 23 to August 31, 2020 by aggregating daily counts from Monday to Sunday without overlapping. For spatial models, which we consider for comparison, we used cumulative counts since the beginning of the study period. For sensitivity analysis, we consider a different starting day of a week to calculate COVID-19 weekly death counts for spatiotemporal models. For spatial models, we consider a different length of aggregating COVID-19 death counts.

2.2. EXPOSURE TO PARTICULATE MATTER ($PM_{2.5}$)

We imported the data from the repository where code and data are publicly available for reproducing analyses in "Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study" (Wu et al. 2020). The county-level long-term averaged $PM_{2.5}$ ($\mu g/m^3$) is calculated from an established exposure prediction model ([21],

Van Donkelaar et al. 2019). Wu et al. (2020) considered the period of 2000–2016. We extended the period up to 2018 for our analysis. Thus, temporal variations of $PM_{2.5}$ during our study period for each county were not considered.

2.3. POTENTIAL CONFOUNDERS

The twelve potential risk factors or confounding variables were selected from the previous benchmark study (Wu et al. 2020). The list of variables includes percent of poverty, population density (4 levels by quartiles), median house value (thousand $), median household income (thousand $), percent of owner-occupied housing, percent of Hispanic population, percent of Black population, percent of the adult population with less than a high school education, and percent of the adult population older than age 65, percent of hospital beds per population and percent of smoking. These were collected from the 2000 and 2010 Census (https://www.census.gov) and the 2005–2016 American Community Surveys (https://www.census.gov/programs-surveys/acs/) according to Wu *et al.* (2020). Note that these variables do not vary temporally but only spatially.

3. METHODS

In this paper, we will explore spatiotemporal modeling of the count data for our main analysis. We account for the spatiotemporal nature of the county-level weekly COVID-19 death cases over the regions introduced in Sect. 2.

Depending on the over-dispersion and zero inflation characteristics in the outcome, we consider a negative binomial (NB) and a zero-inflated negative binomial (ZINB) distribution-based modeling approach. We will compare these results with the spatial model results using cumulative counts to provide general understanding of our implications of findings.

3.1. MODELS

3.1.1. Spatial Negative Binomial Model

Let y_i represent the death counts in county i for a certain week or the aggregated death count for a certain period. We model these cross-sectional spatial count data with a negative binomial distribution (Neelon et al. 2019) specified as,

$$y_i \sim \text{NB}\,(p_i, r) \tag{1}$$

where p_i represents the success probability in the negative binomial distribution for county i and r controls dispersion of the model since $var(y_i) = E(y_i)(1 + E(y_i)/r)$. We model y_i as a generalized linear mixed model with a logistic link function. We assume D fixed-effect covariates including exposure to $PM_{2.5}$, population density, age distribution and several socioeconomic variables. County-specific spatial random intercepts, b_i, are introduced to

allow spatial dependence among the counties. The model is then defined as:

$$\text{logit}\,(p_i) = \theta_i = \log(x_{oi}) + \beta_0 + \sum_{d=1}^{D} \beta_d X_d + b_i. \tag{2}$$

$x_{o\,i}$ is a population size of the i^{th} county so that $\log(x_{o\,i})$ indicates an offset variable. Note that a positive value of β_d associated with a unit change in X_d accounts for an increase in the expected number of counts. We assign $\{b_i\}$ an intrinsic conditional autoregressive (ICAR) prior (Banerjee et al. 2014), which is specified by the following conditional structure:

$$b_i \mid b_{(-i)}, \sigma_b^2 \sim \text{N}\left(\frac{1}{m_i}\sum_{l\in\partial_i} b_l, \frac{\sigma_b^2}{m_i}\right), \tag{3}$$

where m_i is the number of neighbors of the i^{th} county, ∂_i is the set of indices for the neighbors of the i^{th} county and $b_{(-i)}$ is the set of random intercepts except the one for the i^{th} county. σ_b^2/m_i represents the conditional variance given the random intercepts corresponding to the rest of the counties. Note that we assume the first-order neighbor structure. This model is similar to the model considered in Wu et al. (2020), but we consider spatial random intercepts. This model is used to compare with a spatiotemporal model. In the rest of the paper, we will refer to this negative binomial spatial model by SNB.

3.1.2. Spatiotemporal Negative Binomial Model

The NB distribution-based modeling of spatiotemporal count data can be expressed as:

$$y_{ij} \sim \text{NB}\left(p_{ij}, r\right), \tag{4}$$

where y_{ij} and p_{ij} are the count and the success probability corresponding to the negative binomial distribution for the i^{th} county at time t_{ij}, $j = 1, \ldots, n_i$, respectively. We model p_{ij} as a generalized linear mixed model with the logistic link function in a following way:

$$\text{logit}\left(p_{ij}\right) = \theta_{ij} = \log(x_{o\,ij}) + \beta_0 + \sum_{m=1}^{M} \beta_{1m} T_m + \sum_{d=1}^{D} \beta_{d+M} X_d + b_{i1} + b_{i2} t_{ij}. \tag{5}$$

$x_{o\,ij}$ is a population size of the i^{th} county at the time t_{ij} so that $\log(x_{o\,ij})$ indicates an offset variable. $\sum_{m=1}^{M} \beta_{1m} T_m$ is a flexible nonlinear time-fixed effect using M cubic B-splines to capture the time trend fully. $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots \beta_{M+D})^T$ represents the coefficient vector for the fixed effect covariates. $\boldsymbol{b}_i = (b_{i1}, b_{i2})^T$ represents the spatial bivariate random effects for the i^{th} county. b_{i1} is a random intercept and b_{i2} is a random slope for a linear time trend. In this model, we account for time trend both as a fixed effects and also as a random effects because there might be heterogeneity across counties in the temporal dynamic of COVID-19 weekly deaths that remained unexplained by the time invariant covariates.

We modeled the county-specific random effects vector $\boldsymbol{b}_i$, as a bivariate ICAR prior to incorporate spatial dependence in the intercept and slope of the linear county-specific time trends:

$$\boldsymbol{b}_i \mid \boldsymbol{b}_{(-i)}, \sigma_b^2 \sim \mathrm{N}_2\left(\frac{1}{m_i}\sum_{l\in\partial_i}\boldsymbol{b}_l, \frac{1}{m_i}\boldsymbol{\Gamma}\right), \tag{6}$$

where m_i is the number of neighbors for the i^{th} county and ∂_i is the set of neighbors for the i^{th} county. $\boldsymbol{\Gamma}/m_i$ is a 2×2 conditional covariance matrix of $\boldsymbol{b}_i$ given $\boldsymbol{b}_{(-i)}$, the random effects for the rest of the counties. We shall refer to this model as STNB in the rest of this paper.

3.1.3. Spatiotemporal Zero-Inflated Negative Binomial Model

In order to explain the zero inflation in the count data across different counties over time, we consider a spatiotemporal ZINB model (Neelon et al. 2019) for y_{ij} as,

$$y_{ij} \sim \left(1-q_{ij}\right)\mathbb{1}_{(\mathrm{w}_{ij}=0\wedge y_{ij}=0)} + q_{ij}\mathrm{NB}\left(p_{ij}, r\right)\mathbb{1}_{(\mathrm{w}_{ij}=1)}, \tag{7}$$

where q_{ij} represents the probability that the i^{th} county at time t_{ij} belongs to the negative binomial component and w_{ij} represents the corresponding indicator variable. $\wedge$ is a symbol of logical conjunction. We can interpret q_{ij} as the probability that the i^{th} county at time t_{ij} potentially can have death counts. The rest of the parameters are same as defined in (4). To consider both a population size offset and nonlinear time-fixed effect similar to the model in (5), we model q_{ij} and p_{ij} as,

$$\begin{aligned}
\mathrm{logit}\left(q_{ij}\right) &= \mathrm{logit}\left[\Pr\left(\mathrm{w}_{ij}=1 \mid \boldsymbol{\beta}_1, \boldsymbol{b}_{1i}\right)\right] \\
&= \theta_{1ij} = \log(x_{oij}) + \beta_{10} + \sum_{m=1}^{M}\beta_{1m}T_m + \sum_{d=1}^{D_1}\beta_{1(d+M)}Z_d + b_{1i1} + b_{1i2}t_{ij}, \\
\mathrm{logit}\left(p_{ij}\right) &= \theta_{2ij} = \log(x_{oij}) + \beta_{20} + \sum_{m=1}^{M}\beta_{2m}T_m + \sum_{d=1}^{D_2}\beta_{2(d+M)}X_d + b_{2i1} + b_{2i2}t_{ij}
\end{aligned} \tag{8}$$

where $\boldsymbol{b}_{1i} = (b_{1i1}, b_{1i2})^T$ and $\boldsymbol{b}_{2i} = (b_{2i1}, b_{2i2})^T$ represent the random effects corresponding to q_{ij} and p_{ij}, respectively. We impose a multivariate ICAR prior structure (Neelon et al. 2019) on $\boldsymbol{\phi}_i = \left(\boldsymbol{b}_{1i}^T, \boldsymbol{b}_{2i}^T\right)^T$ as,

$$\boldsymbol{\phi}_i \mid \boldsymbol{\phi}_{(-i)}, \boldsymbol{\Gamma} \sim \mathrm{N}_4\left(\frac{1}{m_i}\sum_{l\in\partial_i}\boldsymbol{\phi}_l, \frac{1}{m_i}\boldsymbol{\Gamma}\right), \tag{9}$$

where $\boldsymbol{\Gamma}/m_i$ is a 4×4 conditional covariance matrix. This multivariate ICAR allows spatial dependence and county-specific random time trend within count components and excess-zero components as well as dependence between them. This model will be referred to as

STZINB-NLT in the rest of this paper, where NLT refers to nonlinear fixed-time trend for q_{ij}.

We hypothesize that the assumption of nonlinear fixed time trend for q_{ij} might not be necessary in the sense of model parsimony. Thus, we consider a simpler version by assuming linear time fixed effect in a binary component. That is,

$$\text{logit}\left(q_{ij}\right) = \theta_{1ij} = \log(x_{oij}) + \beta_{10} + \beta_{11}t_{ij} + \sum_{d=1}^{D_1} \beta_{1(d+1)} Z_d + b_{1i1} + b_{1i2}t_{ij}. \quad (10)$$

This model will be referred to as STZINB-LT in the rest of this paper, where LT refers to linear fixed time trend for q_{ij}.

3.2. Bayesian Inference

3.2.1. Prior Specification and MCMC Settings

We illustrate prior and hyper-parameter specifications and Markov Chain Monte Carlo (MCMC) settings under the STZINB model framework. For the latent at-risk indicators w_{ij}, probability was given as $\exp(\theta_{1ij})/[1 + \exp(\theta_{1ij})]$, where θ_{1ij} is defined as either equation (8) or (10). Prior distributions for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ were assumed to be $N_p\left(\beta_0 = \mathbf{0}, \Sigma_0 = 100\mathcal{I}_p\right)$, respectively, where $\mathcal{I}_p$ is a $p \times p$ identity matrix. For r in negative binomial component, a uniform prior was considered. To construct time-basis functions $T_m, m = 1, \cdots, M$, we standardized time points t_{ij} to be ranged from 0 to 1, i.e., $t_{ij} = \frac{j}{J}$, where $J = 23$ is the number of weeks during the study period, and $j = 1, \cdots, J = 23$ is a week indicator. We set three internal knot points 0.25, 0.50, 0.75 considering 0 and 1 as boundaries, so that $M = 7$ throughout our analysis. DIC is calculated as described in Gelman et al. (2013) for model comparison.

For each model, we ran three MCMC chains with 11,000 iterations, and 1,000 burn-in. For each model, convergence of each model was determined by conventional MCMC diagnostics such as trace plots and Geweke z-statistics.

3.2.2. Conditional Posterior Distribution and Model Fitting

A posterior sampling algorithm of STZINB is adopted from Neelon et al. (2019), and it is straightforward to implement the algorithms for the other models as they are simpler models. As outlined in Neelon et al. (2019), we need to update at-risk indicators $\boldsymbol{w}$, coefficients for the binary model component $\boldsymbol{\beta}_1$, coefficients for the count model component $\boldsymbol{\beta}_2$, a dispersion parameter r for the negative binomial distribution, the set of spatial random effects $\boldsymbol{\phi}$ with Γ. The following illustrate the steps of MCMC to update the parameters.

STEP1 Update at-risk indicators $\boldsymbol{w}$

Given current parameter values, we draw w_{ij} from a Bernoulli distribution with probability η_{ij} such that

$$\eta_{ij} = \frac{\Pr(y_{ij}=0|w_{ij}=1)\Pr(w_{ij}=1)}{\Pr(y_{ij}=0|w_{ij}=1)\Pr(w_{ij}=1)+\Pr(y_{ij}=0|w_{ij}=0)\Pr(w_{ij}=0)}$$
$$= \frac{q_{ij}^r(1-p_{ij})}{q_{ij}^r(1-p_{ij})+p_{ij}} \tag{11}$$

where q_{ij} and p_{ij} are defined in (8). Note that q_{ij} is the inverse logit of θ_{1ij}, and p_{ij} is the inverse logit of θ_{2ij}. To avoid numerical issue, we adjusted the sampled q_{ij} and p_{ij} to be within $(0.001, 0.999)$, respectively, in practice.

STEP2 Update $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$

To update $\boldsymbol{\beta}_1$, we draw a latent variable ξ_{1ij} from a Pólya-Gamma distribution $\text{PG}(1, \theta_{1ij})$ as shown in Polson et al. (2013). Given $\boldsymbol{w}$ and $\boldsymbol{\xi}_1$, the full conditional distribution of $\boldsymbol{\beta}_1$ is

$$\Pr(\boldsymbol{\beta}_1|\boldsymbol{w}, \boldsymbol{\xi}_1) \propto \pi(\boldsymbol{\beta}_1)\exp\left[-\frac{1}{2}(z_1 - \boldsymbol{X}\boldsymbol{\beta}_1)^T\Omega_1(z_1 - \boldsymbol{X}\boldsymbol{\beta}_1)\right] \tag{12}$$

where X is a $n \times p$ design matrix, $\pi(\boldsymbol{\beta}_1)$ the prior distribution $N_p(\beta_0, \Sigma_0)$, $z_1 = \frac{\boldsymbol{w}-1/2}{\boldsymbol{\xi}_1}$, and $\Omega_1 = \text{diag}(\boldsymbol{\xi}_1)$ an $n \times n$ precision matrix. Conditional on z_1, we update $\boldsymbol{\beta}_1$ from $N_p(\mu, \Sigma)$ where $\Sigma = \left(\Sigma_0^{-1} + X^T\Omega_1 X\right)^{-1}$, and $\mu = \Sigma\left(\Sigma_0^{-1}\beta_0 + X^T\Omega_1 z_1\right)$. We update $\boldsymbol{\beta}_2$ by similar process using the corresponding Pólya-Gamma distribution $\text{PG}(y_{ij} + r, \theta_{2ij})$ as shown in Pillow and Scott (2012).

STEP3 Update r

We can use a random-walk Metropolis–Hastings step illustrated in Neelon et al. (2019). We present Metropolis–Hastings method with uniform prior because of efficiency in computation time.

STEP4 Update $\boldsymbol{\phi}$, Γ

Let $\boldsymbol{\phi}_{11} = (b_{111}, \cdots, b_{1n1})^T$ be the $n \times 1$ vectors of random intercepts for the binary component, $\boldsymbol{\phi}_{12} = (b_{112}, \cdots, b_{1n2})^T$ be the $n \times 1$ vectors of random slopes for the binary component, $\boldsymbol{\phi}_{21} = (b_{211}, \cdots, b_{2n1})^T$ be the $n \times 1$ vectors of random intercepts for the count component, and $\boldsymbol{\phi}_{22} = (b_{212}, \cdots, b_{2n2})^T$ be the $n \times 1$ vectors of random slopes for the binary component. Then, $\boldsymbol{\phi} = (\boldsymbol{\phi}_{11}, \boldsymbol{\phi}_{12}, \boldsymbol{\phi}_{21}, \boldsymbol{\phi}_{22})^T$ is the $4n \times 1$ collection of all random effects by definition. Under the STZINB-NLT model illustrated in Sect. (3.1.3), the conditional prior for $\boldsymbol{\phi}_{11}$, for instance, is

$$\Pr(\boldsymbol{\phi}_{11}|\boldsymbol{\phi}_{12}, \boldsymbol{\phi}_{21}, \boldsymbol{\phi}_{22}, \Gamma) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\phi}_{11} - \boldsymbol{\mu}_{11})^T\Sigma_{11}(\boldsymbol{\phi}_{11} - \boldsymbol{\mu}_{11})\right] \tag{13}$$

where $\Sigma_{11} = \left[\Gamma_{11} - \Gamma_{1,-1}\Gamma_{-1,-1}^{-1}\Gamma_{-1,1}\right]^{-1} Q$, $\boldsymbol{\mu}_{11} = \left[\left(\Gamma_{1,-1}\Gamma_{-1,-1}^{-1}\right) \otimes I_n\right]\boldsymbol{\phi}_{(-1)}$, $\Gamma_{1,-1}$, $\Gamma_1 1$ denotes the first element of Γ, $\Gamma_{(1,-1)}$ is the 1×3 vector comprising the first row of Γ with element 1 removed, $\Gamma_{(-1,-1)}$ is the 3×3 sub-matrix of Γ after removing row 1 and column 1, $Q = M - A$, $M = \text{diag}(m_1, \cdots, m_n)$ an $n \times n$ matrix with diagonal elements equal to the number of neighbors for each spatial unit, A is an $n \times n$ adjacency

Table 1. Summary statistics for four regions: Mid-Atlantic (New Jersey, New York, and Pennsylvania), Pacific (California, Oregon, and Washington), South Atlantic (Florida, Georgia, North Carolina, South Carolina), and Midwest (Iowa, Kansas, Missouri, Nebraska, North Dakota, and South Dakota)

	Mid-Atlantic	Pacific	South Atlantic	Midwest
Number of counties	150	133	372	531
Average of zero death proportions (%)	55.0(9.3)	63.2(8.7)	53.7(13.5)	90.4(3.2)
Average of weekly death counts	16.4(169.0)	5.0(54.4)	2.6(9.1)	0.3(2.1)
Cumulative death counts by 2020/08/31	377.1(1028.4)	116.8(527.7)	59.4(168.8)	7.2(37.5)
Population size (thousands)	275(415)	373(1,014)	666(1,455)	29(79)
2000–2018 averaged ambient $PM_{2.5}$ $(\mu g/m^3)$	9.4(2.0)	6.2(3.0)	10.5(1.1)	6.8(1.8)
Poverty rate (%)	8.0(3.5)	9.3(4.3)	12.2(5.5)	9.6(5.4)
Population density (in sq mi)	2,525(8,703)	919(2,306)	390(732)	99(362)
Median house value[MHV] (in thousand $)	192.6(132.9)	275.7(172.9)	125.6(55.1)	99.5(31.1)
Median household income[MHI] (in thousand $)	60.2(17.0)	54.4(14.8)	43.2(9.8)	50.8(9.2)
Home owners rate (%)	74.9(10.5)	67.2(8.5)	71.0(8.1)	77.2(6.4)
Hispanic (%)	5.7(7.5)	16.7(15.1)	6.7(6.8)	3.7(5.4)
Less than high school education (%)	18.3(5.3)	15.0(8.6)	25.9(9.6)	16.3(7.4)
Black (%)	4.7(6.8)	1.6(2.1)	23.1(16.9)	1.1(3.3)
Older than age 65 (%)	15.7(2.5)	15.5(4.5)	15.1(4.6)	18.3(4.4)
Hospital beds per population (%)	0.33(0.28)	0.27(0.28)	0.32(0.34)	0.55(0.76)
Smoke rate (%)	47.5(8.0)	46.0(9.2)	47.5(9.5)	45.8(7.2)

matrix with $a_{ii} = 0$, $a_{il} = 1$ if county units i and l are neighbors, and $a_{il} = 0$ otherwise. We update each vector of $\boldsymbol{\phi} = (\boldsymbol{\phi}_{11}, \boldsymbol{\phi}_{12}, \boldsymbol{\phi}_{21}, \boldsymbol{\phi}_{22})^T$ from its normal full conditional distribution based on (13) applying sum-to-zero constraints as needed. To update Γ, we use its conjugate prior, an inverse-Wishart full conditional distribution.

4. RESULTS

4.1. OVERALL DESCRIPTION

Table 1 shows summary statistics of our data for the four regions in the USA: Mid-Atlantic (New Jersey, New York, and Pennsylvania), Pacific (California, Oregon, and Washington), South Atlantic (Florida, Georgia, North Carolina, South Carolina), and Midwest (Iowa, Kansas, Missouri, Nebraska, North Dakota, and South Dakota). For example, 55.0 for Mid-Atlantic in the row of the average of zero death proportions means on average 55.0% of counties in Mid-Atlantic region have zero deaths during the 23 weeks. In the row of the average of weekly death counts, 16.4 for Mid-Atlantic means on average 16.4 deaths for each county during a week.

The sample variance of COVID-19 weekly death counts exceeds the sample mean for all the four regions in Table 1. This indicates that the negative binomial distribution, which can accommodate overdispersion, would be suitable in modeling COVID-19 weekly death counts. High proportion of zero counts may not be fully explained by the negative binomial distribution alone. Therefore, zero-inflated negative binomial (ZINB) models could be suitable to account for the excess zero counts as well as the overdispersion. This modeling

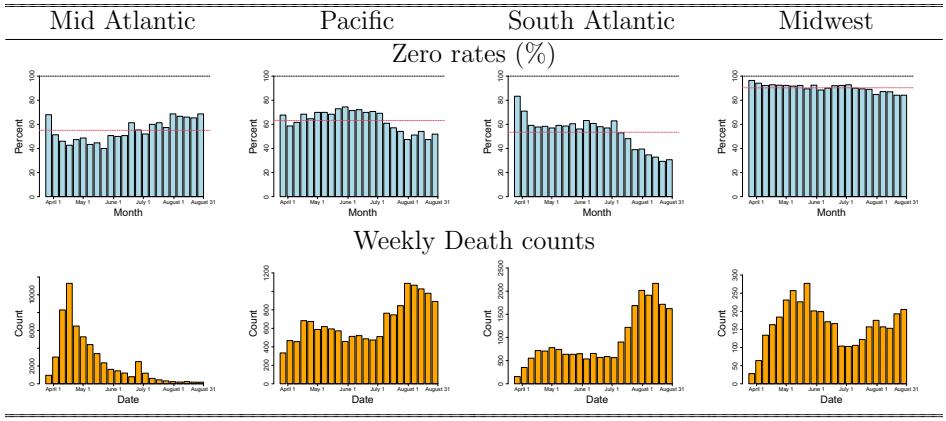

Figure 1. Bar plots of weekly COVID-19 death counts during the study period (2020/03/23–2020/08/31). In the first row, the height of each bar represents the proportion of counties with no death count (0–100%) in each week. A red dashed line is the global average of zero rates across time (week). In the second row, the height of each bar represents the weekly death counts averaged across all counties with regions. Please note that the ranges of y-axis are different among the four regions .

Table 2. Average and standard deviation (SD) of 23 correlation values calculated between logarithm of COVID-19 weekly death counts per capita and $PM_{2.5}$ over counties in each week for four regions. A zero count is replaced with 0.5 when calculating correlation. The row of percentage of significance provides the percentage of significant correlation values (different from zero) based on a t-test with a 5% significance level

	Mid-Atlantic	Pacific	South Atlantic	Midwest
Average	0.344	0.436	0.076	0.215
SD	0.081	0.108	0.053	0.051
Percentage of Significance	100%	96%	43%	96%

allows that the observed zero would come from two different sources: structural zero and zero from the negative binomial component.

The four regions have different characteristics of COVID-19 death counts until August. Figure 1 illustrates two types of bar plots representing the weekly rates of zeros and the weekly death counts averaged across counties within regions. Midwest has the highest zero rates over time. Pacific and South Atlantic regions have decreasing zero rates over time in general. The bar plot of the weekly death counts for both regions shows bimodal shapes with a large peak in early August. On the other hand, Midwest has bimodal shape but the peak is in early May. Different from the other three regions, Mid-Atlantic is left-skewed with a peak in April.

To investigate possible association between COVID-19 death counts and $PM_{2.5}$, we compute correlation between log of COVID-19 weekly death counts per capita and $PM_{2.5}$ over counties in each region for each week. With 23 weeks of consideration, we have 23 correlation values for each region. Note that all correlation values are positive. In Table 2, we provide average and standard deviation of these 23 values for each region. Also, we provide the percentage of significant correlation based on a t-test for nonzero correlation with 5%

significance level. Majority of weeks has statistically significant nonzero correlation values except South Atlantic. From this investigation, we hypothesized that the COVID-19 counts and long-term $PM_{2.5}$ exposure were positively associated.

4.2. ESTIMATION RESULTS OF THE MODELS ON COVID-19 MORTALITY

Among the Bayesian spatiotemporal models such as STNB, STZINB-LT, and STZINB-NLT, one model that shows the lowest Deviance information criterion (DIC) is chosen for each of the four regions. STZINB-NLT is chosen for Mid-Atlantic and Pacific regions, while STZINB-LT is chosen for South-Atlantic and Midwest regions. Note that we include a nonzero-inflated model for comparison and the zero-inflated models were selected for all the regions we considered in this study with given study period. The DIC values for the models we considered are provided in Tables 4, 6, 8 and 10 in the Supplementary document.

Table 3 describes the estimated coefficients and their 95% credible intervals of the covariates on COVID-19 weekly death counts under the selected model. The results for the four regions indicate that long-term exposure to $PM_{2.5}$ is positively associated with the expected COVID-19 weekly death counts, but only Pacific region shows lack of significance based on the 95% credible interval. The point estimate of $\beta_{2,2}$ that corresponds to $PM_{2.5}$ for the Mid-Atlantic region is equal to 0.069, which is the change in the expected COVID-19 weekly death counts in log scale by a unit change in $PM_{2.5}$. That is, an increase of 1 $\mu g/m^3$ in the long-term $PM_{2.5}$ level is associated with a $e^{0.069} - 1 \simeq 7.1\%$ increase in the expected COVID-19 weekly death counts per county after controlling all the confounding factors. Similarly, we have 2.3% for Pacific, 3.1% for South Atlantic and 9.2% for Midwest of increments in the expected COVID-19 weekly death counts per county.

Increases in MHI, Hispanic population and number of beds are associated with the increases in the expected COVID-19 death counts, although some are not significant based on 95% credible intervals. MHV is negatively associated except Mid-Atlantic region, while Black population and Smoking rate are mostly positively associated except Pacific region for Black population and South Atlantic for smoking rate. The other confounders show mixed directions of the effects.

Table 4 illustrates the estimated coefficients and their 95% credible intervals of the covariates on q_{ij}, the probability that belongs to the negative binomial component under the selected model. The estimated coefficients for long-term ambient $PM_{2.5}$ are positive for all regions except for Pacific region. This implies that the increase in the level of long-term ambient $PM_{2.5}$ can potentially increase the chance of COVID-19 death since q_{ij} is the probability that belongs to the negative binomial component which models nonzero death counts. The effect of the long-term ambient $PM_{2.5}$ on q_{ij} is statistically significant for Mid-Atlantic and Midwest regions. If the long-term ambient $PM_{2.5}$ increases in 1 $\mu g/m^3$ adjusting the other confounding factors in Midwest region, for example, the log odds ratio for q_{ij} increases by 0.355. This implies that if the long-term ambient $PM_{2.5}$ changes from 8 $\mu g/m^3$ to 9 $\mu g/m^3$, q_{ij} will increase in $\exp(9 \cdot 0.355)/[1 + \exp(9 \cdot 0.355)] - \exp(8 \cdot 0.355)/[1 + \exp(8 \cdot 0.355)] \simeq 0.016$. On the last week of the study period, q_{ij} in Midwest region ranges from 0.164 (Jewell County) to 0.999 (Johnson County).

Table 3. Point estimates and 95% credible intervals of the fixed effects for the expected COVID-19 death counts of the model chosen by DIC in each region. Results are obtained by fitting a STZINB-NLT given in (8) for Mid-Atlantic and Pacific and by fitting a STZINB-LT given in (10) for South Atlantic and Midwest. The bold numbers indicate statistical significance

		Mid-Atlantic		Pacific		South Atlantic		Midwest	
		STZINB-NLT		STZINB-NLT		STZINB-LT		STZINB-LT	
$\beta_{2,2}$	$PM_{2.5}$	**0.069**	(0.02, 0.13)	0.023	(−0.03, 0.17)	**0.031**	(0.01, 0.05)	**0.088**	(0.04, 0.14)
$\beta_{2,3}$	Poverty	−0.210	(−0.32, 0.18)	**0.152**	(0.01, 0.49)	0.008	(−0.03, 0.04)	**0.191**	(0.02, 0.37)
$\beta_{2,42}$	Population density2	1.153	(−1.35, 5.85)	**−2.394**	(−4.65, −1.43)	**−1.107**	(−1.51, −0.8)	0.381	(−0.21, 1.05)
$\beta_{2,43}$	Population density3	2.252	(−0.14, 6.93)	**−2.445**	(−5.42, −1.06)	**−1.599**	(−2.03, −1.28)	**0.915**	(0.52, 1.27)
$\beta_{2,44}$	Population density4	**2.894**	(0.47, 7.43)	**−3.509**	(−8.22, −1.53)	**−1.682**	(−2.13, −1.32)	**0.678**	(0.25, 1.11)
$\beta_{2,5}$	MHV	**0.419**	(0.24, 1.33)	**−0.901**	(−2.03, −0.43)	**−0.095**	(−0.13, −0.06)	**−0.490**	(−0.71, −0.28)
$\beta_{2,6}$	MHI	0.095	(−0.14, 0.39)	**0.852**	(0.41, 1.85)	**0.098**	(0.05, 0.14)	**0.466**	(0.22, 0.76)
$\beta_{2,7}$	Home owners rate	−0.042	(−1.98, 0.06)	0.098	(−0.06, 0.52)	**−0.081**	(−0.11, −0.05)	**0.305**	(0.11, 0.49)
$\beta_{2,8}$	Hispanic	0.257	(−0.37, 0.37)	**0.446**	(0.27, 1.12)	**0.088**	(0.07, 0.1)	**0.183**	(0.09, 0.28)
$\beta_{2,9}$	Education	0.078	(−0.03, 0.19)	−0.088	(−0.7, 0.12)	**0.05**	(0.02, 0.08)	−0.128	(−0.29, 0.05)
$\beta_{2,10}$	Black	0.167	(−0.4, 0.23)	**−0.112**	(−0.18, 0.00)	**0.243**	(0.21, 0.27)	**0.203**	(0.13, 0.28)
$\beta_{2,11}$	Older than age 65	0.010	(−0.14, 0.1)	**−0.299**	(−0.65, −0.01)	**0.236**	(0.21, 0.26)	−0.139	(−0.32, 0.05)
$\beta_{1,12}$	Number of beds	0.093	(−0.11, 0.17)	**0.864**	(0.48, 1.69)	**0.045**	(0.02, 0.07)	0.016	(−0.17, 0.23)
$\beta_{1,13}$	Mean value of smoke rate	**0.285**	(0.20, 0.40)	0.024	(−0.34, 0.21)	**−0.037**	(−0.07, −0.01)	0.043	(−0.22, 0.29)

STZINB-NLT : Spatiotemporal zero-inflated negative binomial (nonlinear time trend)

STZINB-LT : Spatiotemporal zero-inflated negative binomial (linear time trend)

MHV : Median House Value

MHI : Median Household Income

Table 4. Point estimates and 95% credible intervals of the fixed effects on q_{ij}, (the probability that belongs to the negative binomial component). Results are obtained by fitting a STZINB-NLT given in (8) for Mid-Atlantic and Pacific and by fitting a STZINB-LT given in (10) for South-Atlantic and Midwest. The bold numbers indicate statistical significance

		Mid-Atlantic		Pacific		South Atlantic		Midwest	
		STZINB-NLT		STZINB-NLT		STZINB-LT		STZINB-LT	
$\beta_{1,2}$	$PM_{2.5}$	**1.404**	(0.82, 2.21)	−0.026	(−0.12, 0.07)	0.952	(−0.28, 2.13)	**0.355**	(0.02, 1.07)
$\beta_{1,3}$	Poverty	**2.438**	(1.21, 4.34)	**−0.252**	(−0.45, −0.05)	−1.098	(−4.86, 0.6)	**−0.65**	(−1.08, −0.29)
$\beta_{1,42}$	Population density2	17.658	(−12.46, 41.34)	0.394	(−0.24, 0.97)	**17.491**	(4.87, 29.63)	−1.829	(−3.3, 0.03)
$\beta_{1,43}$	Population density3	6.455	(−6.21, 18.43)	0.329	(−0.48, 1.17)	**36.055**	(23.61, 51.14)	**−2.569**	(−5.11, −1.01)
$\beta_{1,44}$	Population density4	**−13.404**	(−23.6, −1.96)	**1.575**	(0.41, 2.66)	**33.161**	(14.23, 49.45)	0.941	(−4.66, 8.13)
$\beta_{1,5}$	MHV	**−9.428**	(−16.61, −5.23)	−0.258	(−0.57, 0.06)	0.07	(−2.61, 3.18)	0.903	(−0.07, 2.48)
$\beta_{1,6}$	MHI	**10.00**	(4.76, 19.78)	0.157	(−0.2, 0.52)	0.997	(−3.02, 5.11)	−0.194	(−1.23, 0.72)
$\beta_{1,7}$	Home owners rate	0.338	(−0.54, 1.93)	−0.192	(−0.5, 0.08)	−1.866	(−6.89, 1.88)	**−1.474**	(−2.76, −0.65)
$\beta_{1,8}$	Hispanic	0.045	(−1.18, 1.35)	0.214	(−0.1, 0.52)	1.208	(−1.05, 4.39)	0.221	(−0.23, 0.77)
$\beta_{1,9}$	Education	0.14	(−0.86, 1.48)	0.209	(−0.06, 0.49)	−0.885	(−3.95, 1.63)	0.601	(−0.19, 1.33)
$\beta_{1,10}$	Black	0.264	(−0.77, 1.98)	−0.022	(−0.28, 0.27)	1.26	(−2.22, 5.54)	**−0.736**	(−1.44, −0.21)
$\beta_{1,11}$	Older than age 65	**3.029**	(0.81, 7.47)	0.151	(−0.23, 0.58)	**−2.737**	(−5.64, −0.05)	**0.685**	(0.12, 1.16)
$\beta_{1,12}$	Number of beds	0.855	(−0.75, 3.28)	**−0.361**	(−0.69, −0.05)	0.214	(−2.18, 2.97)	0.227	(−0.54, 1.03)
$\beta_{1,13}$	Mean value of smoke rate	−0.056	(−2.44, 1.46)	−0.01	(−0.25, 0.24)	1.575	(−2.09, 4.41)	0.153	(−1.11, 2.7)

STZINB-NLT : Spatiotemporal zero-inflated negative binomial (nonlinear time trend)

STZINB-LT : Spatiotemporal zero-inflated negative binomial (linear time trend)

MHV : Median house value

MHI : Median household income

The most estimates for MHI, Hispanic population and number of beds are positive, which are similar to the results in Table 3, but they are mostly not significant. For the other confounders, the directions of the effects for the confounders are rather different and compared to the results in Table 3, and the effects are mostly not significant.

Figure 2 shows the time-averaged q_{ij}. Note that q_{ij} represents the probability of the i^{th} county being in the negative binomial component at the j-th week. The dark colored counties for each region indicate a high time-averaged probability in the first row. In the second row, we show 12 counties whose estimated q_{ij}s correspond to the $\frac{j}{11}$th quantiles for $j = 0, 1, \cdots, 11$. The differences between the first and the second counties are relatively larger in Mid-Atlantic and Pacific regions than those in the other two regions. The decreasing rate over counties is slower in South Atlantic and Midwest regions compared to the other regions.

The estimated Γ, the covariance matrix of four spatial random effects in each region shows that off-diagonal entries are close to zero (Table 3 in the Supplementary document). Recall that these random effects are random intercepts and random slopes for time in the count component and excess zero component. Thus, estimated zero implies these random effects are not correlated to each other, although each random effect is spatially dependent.

Figure 3 shows the estimated nonlinear mixed time effects in the negative binomial component (the COVID-19 weekly death counts). Each of the four regions shows different nonlinear temporal patterns in the negative binomial component. We selected five representative counties out of twelve appeared in Fig. 2 considering the rank of the estimated q_{ij}, and they show distinctive patterns over study periods since we allow county-specific random effects in the intercept and linear components. The temporal patterns are overall similar to the patterns we observed in Fig. 1. This is expected since the covariates in the models are static so that the time effect components in the models try to capture the temporal patterns unexplained by the static covariates. The estimated effects of the models were not much sensitive by a different starting day of a week when aggregating COVID-19 daily death counts to weekly death counts (results not shown).

4.3. COMPARISON BY MODELING TECHNIQUES

In this section, we compare our spatiotemporal models with a spatial negative binomial model (SNB) with cumulative death counts. The results are provided in Table 5. Since the models are different as well as the response variables are different, we cannot directly compare the results between the SNB model and spatiotemporal models. However, we can assess the direction of associations and whether they are consistent or not between these models. Under the SNB model, we found that the long-term exposure to $PM_{2.5}$ is positively associated with the expected cumulative death counts for COVID-19 for all four regions. This is consistent with the results from the spatiotemporal models provided in the previous subsection and also with the results by Wu et al. (2020). This supports the hypothesis of positive association between the long-term ambient $PM_{2.5}$ and the expected COVID-19 death counts. On the other hand, the results are different in some aspects. Note that the effect size of long-term ambient $PM_{2.5}$ estimated from the spatial model exceeds those from spatiotemporal models. We suspect that this could be due to the lack of temporal components

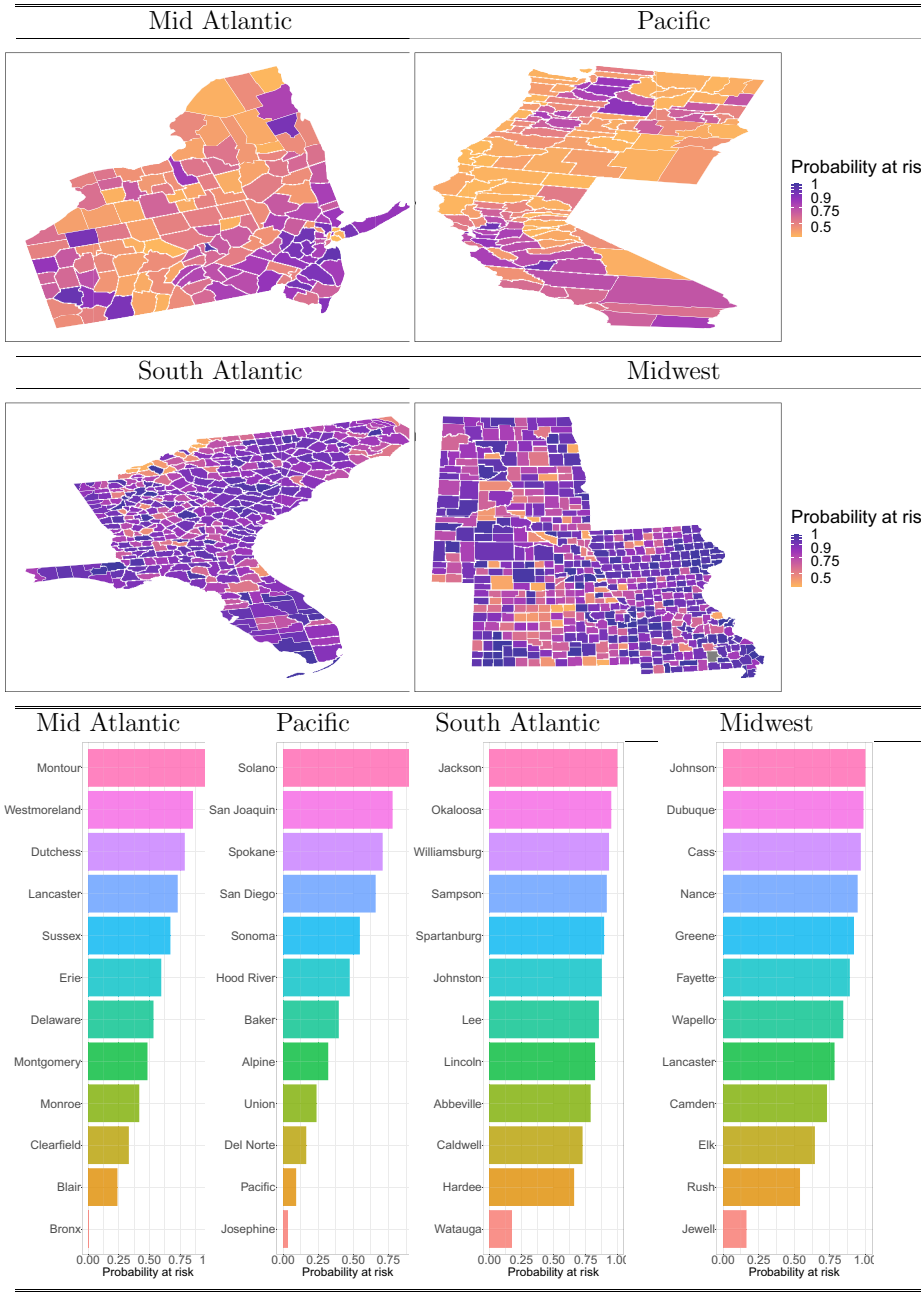

Figure 2. Visualization of time-averaged q_{ij} using the selected models for the four regions. The first two rows show the time-averaged q_{ij} over counties. The last row show twelve counties whose estimated q_{ij}s correspond to the $\frac{j}{11}$th quantiles for $j = 0, 1, \cdots, 11$.

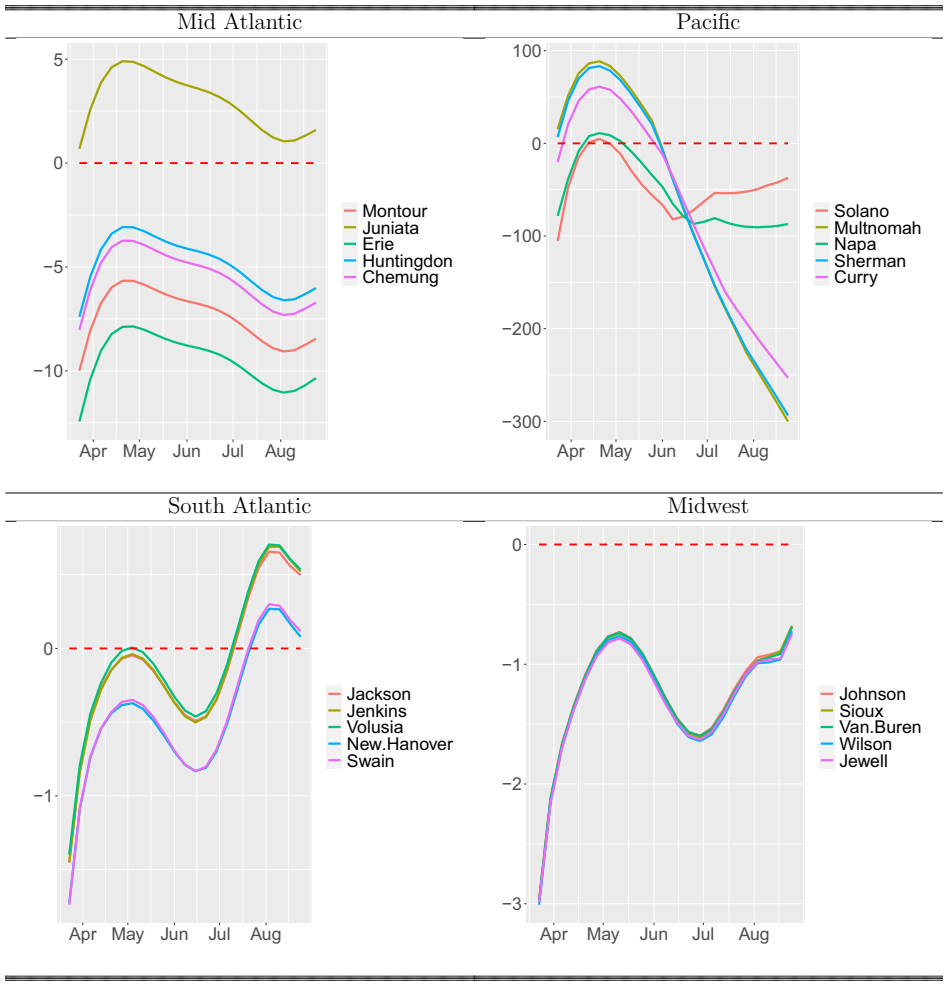

Figure 3. The median value of estimated nonlinear mixed time effect in the negative binomial component (the COVID-19 death counts) using the selected model structure. The x-axis represents timeline (weeks). We show five representative counties in each region for visualization. A red dash line is a zero reference. The ranges of y-axis are different among the four regions (Color figure online).

in the model. Also, the directions of the association for some confounders are not the same and there are less number of significant effects.

The estimated effects are rather sensitive to the length of the period for aggregating the outcome variable (results not shown), which could be an issue to consider a SNB model for COVID-19 death counts. In addition to this issue, a SNB model ignores temporal characteristics of the data. Thus, it is natural to consider a spatiotemporal model, but we should be careful in interpreting the results from the spatiotemporal models since the model complexity can result in an overfitting or unstable estimation.

Table 5. The estimation results of effects on the cumulative death counts of COVID-19 using the spatial negative binomial model (SNB model; Sect. 3.1.1) for the four regions (Mid-Atlantic, Pacific, South Atlantic, and Midwest). The counts are aggregated from March 23, 2020 to August 31, 2020. The bold numbers indicate statistical significance

		Mid-Atlantic		Pacific		South Atlantic		Midwest	
$\beta_{2,2}$	$PM_{2.5}$	**0.165**	(0.01, 0.33)	0.055	(−0.07, 0.19)	**0.249**	(0.13, 0.37)	**0.274**	(0.16, 0.39)
$\beta_{2,3}$	Poverty	−0.038	(−0.36, 0.31)	0.023	(−0.22, 0.26)	−0.091	(−0.21, 0.04)	0.062	(−0.14, 0.28)
$\beta_{2,42}$	Population density2	4.216	(−1.5, 10.5)	0.426	(−0.39, 1.2)	0.295	(−0.17, 0.75)	0.211	(−0.26, 0.68)
$\beta_{2,43}$	Population density3	5.612	(−0.04, 12.03)	**0.909**	(0.25, 1.61)	**0.938**	(0.45, 1.41)	**0.961**	(0.51, 1.45)
$\beta_{2,44}$	Population density4	**5.894**	(0.51, 12.6)	**1.834**	(1.11, 2.58)	**1.439**	(0.92, 2)	9.939	(7.82, 12.25)
$\beta_{2,5}$	MHV	0.209	(−0.47, 0.89)	−0.247	(−0.65, 0.14)	−0.149	(−0.31, 0.01)	−0.063	(−0.3, 0.18)
$\beta_{2,6}$	MHI	**1.040**	(0.28, 1.85)	0.337	(−0.11, 0.73)	−0.020	(−0.19, 0.16)	**0.425**	(0.14, 0.69)
$\beta_{2,7}$	Home owners rate	**−1.159**	(−1.78, −0.57)	−0.142	(−0.54, 0.24)	0.040	(−0.08, 0.16)	−0.051	(−0.28, 0.2)
$\beta_{2,8}$	Hispanic	−0.050	(−0.51, 0.52)	0.198	(−0.22, 0.61)	**0.239**	(0.13, 0.35)	0.001	(−0.16, 0.17)
$\beta_{2,9}$	Education	−0.080	(−0.45, 0.29)	0.256	(−0.08, 0.66)	−0.098	(−0.23, 0.04)	0.012	(−0.16, 0.2)
$\beta_{2,10}$	Black	−0.061	(−0.36, 0.25)	−0.050	(−0.32, 0.24)	**0.324**	(0.21, 0.44)	**−0.596**	(−0.76, −0.44)
$\beta_{2,11}$	Older than age 65	0.140	(−0.2, 0.43)	**−0.759**	(−1.17, −0.38)	**0.210**	(0.1, 0.33)	**−0.653**	(−0.89, −0.41)
$\beta_{1,12}$	Number of beds	0.057	(−0.24, 0.4)	−0.247	(−0.72, 0.15)	0.008	(−0.1, 0.11)	−0.096	(−0.31, 0.11)
$\beta_{1,13}$	Mean value of smoke rate	**0.378**	(0.12, 0.61)	**0.321**	(0.05, 0.6)	−0.028	(−0.13, 0.07)	0.055	(−0.17, 0.3)

MHV : Median House Value

MHI : Median Household Income

5. CONCLUSION AND DISCUSSION

We investigate the relationship between long-term exposure to $PM_{2.5}$ and county-level COVID-19 weekly death counts by implementing and comparing several spatiotemporal negative binomial models with/without a zero-inflated component. These associations were adjusted by social and environmental factors. We also considered county-level random effects that account for spatiotemporal interaction in both the structural zero component and the negative binomial component via an ICAR model. We considered possible nonlinear time effects in both components as well.

We hypothesize potential heterogeneity in the effects of long-term exposure to $PM_{2.5}$ and other social and environmental factors on the COVID-19 weekly death counts across divisions of the USA, likely due to different region-specific sociocultural, behavioral and healthcare system as well as COVID-19 policies by assuming that nearby states have similar characteristics. Thus, we consider four geographically different regions (Mid-Atlantic, Pacific, South Atlantic, and Midwest) and applied the spatiotemporal models to each region.

Based on model comparison by DIC, we selected zero-inflated models for all four regions. Within zero-inflated models, the linear time trend model for the probability that belongs to the negative binomial component was chosen for South Atlantic and Midwest regions, while the nonlinear time trend model was chosen for the other two regions. Note that we assumed nonlinear time trend for the negative binomial component for all four regions.

We compare the results obtained from these spatiotemporal models with the results obtained from a spatial negative binomial model that completely ignores the temporal information. The spatial negative binomial model was applied to the cumulative death counts until the date we considered (August 31, 2020). Because the spatiotemporal model uses weekly COVID-19 counts as outcome and the spatial model uses cumulative death counts for the whole study period as outcome, the results obtained under the two models have different interpretations. Still we found that the direction and the strength of the associations between $PM_{2.5}$ and COVID-19 death counts are consistent.

The estimated coefficients associated with long-term exposure to $PM_{2.5}$ from the selected models are mostly positive and statistically significant for the regions under this study after adjusting the nonlinear time trend with a county-specific random slope, spatial dependence, and many other measured confounders. The directions of association for COVID-19 weekly death counts, although not significant for Pacific, are consistent with the result of the previous study (Wu et al. 2020). Note that Wu et al. (2020) did not consider spatial and temporal dependence in the model. We also checked the effects of long-term exposure to $PM_{2.5}$ for all the models introduced in Sect. 3.1 and the model from Wu et al. (2020) using the data used in this study. All the models show the positive association. The results are given in Tables 5, 7, 9 and 11 in the Supplementary document.

Some of confounders may affect on COVID-19 weekly death counts through an interaction effect. Thus, we investigate effects by an interaction term between $PM_{2.5}$ and other confounders, but they are not critical for the models and data sets we considered. These findings add evidence of the increase in risk of death for COVID-19 by the long-term exposure to air pollution into the literature.

By aggregating the data over time or region, we may encounter ecological fallacy and need to be careful in interpreting the result. That is, the increased effects on COVID-19 death counts by $PM_{2.5}$ that we observed from our analysis may not imply an increased risk of individual's death by COVID-19 since the analysis is based on a county-level weekly aggregated death counts. Our results only imply that long-term exposure to $PM_{2.5}$ may increase COVID-19 weekly death counts at a county level.

By applying the models to each region, separately, we are able to see different association patterns among regions. Although the effect of the long-term exposure to $PM_{2.5}$ on the COVID-19 weekly death counts is in the same direction for all four regions, the size of the effects is different. Also, the effects of the some other confounders show different directions by region. The proposed models, spatiotemporal zero-inflated negative binomial models with nonlinear time effects, spatial random effects and spatiotemporal interaction random effects, are very flexible and capture the spatiotemporal characteristics of the data. On the other hand, the complexity of the model could lead to an increased variability in estimation due to the large number of parameters to estimate. This issue could be alleviated by controlling the prior distribution with the information from the previous study.

To support our claim about our methodology and modeling strategy, we have done extensive simulation studies which mimics our specific spatiotemporal data structure with nonlinear temporal effects. Full details of our simulation studies are included in the Supplementary document. We have developed a user-friendly R tool. All model-related R codes and software can be downloaded from GitHub repository (https://github.com/junpeea). We are also in the process of developing a R-Shiny-based application for cloud-based deployment and interactive interface for non-statistician's easy use and access.

As the spread of COVID-19 is ongoing, geographically different regions have different dynamics of the disease spread and a spatiotemporal model is flexible enough to handle different types of dynamics. As the surge of COVID-19, the zero-inflated model might not be suitable for many states anymore. However, by investigating and comparing several spatiotemporal models including nonzero-inflated models, we can find a reasonable model that explains the characteristics of the data. The risk factor and confounders we used are not temporally varying although the response variable, COVID-19 weekly death counts, is varying over time. Temporal variations in the response variable would be due to unavailable temporal covariates such as policy changes by county health department on COVID-19 and policy changes for school opening and business operation. To handle temporal variations without such temporal covariates, we introduced nonlinear time effects with county-specific random slopes. Once this additional information is available, we can easily accommodate it into the model.

One can consider applying the spatiotemporal model to county-level COVID-19 weekly death counts for the entire USA. This can be doable but a single model may not be able to capture heterogeneous effects that we found in this study. Also, handling a spatial dependence model with a large number of spatial regions brings an additional computational burden. On the other hand, we can extend the current model with a spatiotemporally varying coefficient model to capture heterogeneous effects by states or by county so that we can investigate the whole data into one model framework. A modified spatial dependence modeling such as

allowing spatial dependence only within state to increase computational efficiency can be also considered. We plan to investigate such extension as a future study.

The effect of $PM_{2.5}$ on individuals' health has been investigated in both long-term and short-term directions in the literature. COVID-19 changed economic environment as well as people's life styles in many ways, which cause large variations (either up or down) in PM levels during a short-term period (Wu et al. 2020; Venter et al. 2020) As our focus is long-term effect of $PM_{2.5}$ on COVID-19 death counts, these short-term changes of PM levels were not considered in our analysis. Thus, our findings are restricted to the association between long exposure to $PM_{2.5}$ and COVID-19 death counts, in particular, aggregated death counts over counties and weeks.

Finally, as we have mentioned earlier, this study contributes toward the possibility of the hazards of $PM_{2.5}$ on COVID-19-positive cases and related mortality. Specifically, we believe that as we have incorporated the "zero-inflation" part to the model, it will allow to make inference in early stages of the pandemic of this kind. Nevertheless, we assent that the claim on the higher risk of a COVID-19 death in polluted counties is open to debate as our findings (positive association) do not imply causation. It is also acknowledged that the findings in this paper are limited as they depend on several factors such as the data structure, models and inference methods. There have been about only two years since the coronavirus outbreak began; even so a heavy contribution from several genre of research related to this pandemic—which is a great reflection of the fact that researchers around the world are trying to infer on the connection between the air pollution and its association with the deaths related to COVID-19. However, we are in a urgent need of more research to be done, at the granular level with more complexity in the data and in the modeling. As we are working on this problem, as a statistician and data scientist, it is believed that there is an immediate need of more patient-level data (which is not easy to access because of the issues related to data confidentiality, data sharing, and other related things) than the kind of data publicly available these days. That also makes us believe that as more similar findings are accumulated by other researchers, our claim would become much more stronger.

ACKNOWLEDGEMENTS

Dominici was supported by National Institute of Health (NIH) grants, R01ES026217, R01MD012769, R01ES028033, P30ES000002, and the 2020 Starr Friedman Award. Lim was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (Nos. 2019R1A2C1002213, 2020R1A4A1018207).

[Received July 2021. Revised December 2021. Accepted January 2022. Published Online January 2022.]

REFERENCES

Abdelzaher H et al (2020) COVID-19 Genetic and environmental risk factors: a look at the evidence. Front Pharmacol 11:1528 (**ISSN: 1663-9812**)

Ali N, Islam F (2020) The effects of air pollution on COVID-19 infection and mortality–a review on recent evidence. Front Pub Health 8:779 (**ISSN: 2296-2565**)

Atmospheric Composition Analysis Group http://fizz.phys.dal.ca/~atmos/martin/

Banerjee S, Carlin BP, Gelfand AE (2014) Hierarchical modeling and analysis for spatial data. CRC Press, USA

Bhaskar A, Chandra J, Braun D, Cellini J, Dominici F (2020) Air pollution, SARSCoV-2 transmission, and COVID-19 outcomes: A state-of-the-science review of a rapidly evolving research area. medRxiv

Chakrabarty RK et al (2020) Ambient PM2. 5 exposure and rapid spread of COVID-19 in the United States. Sci Total Environ 760:143391

Comunian S, Dongo D, Milani C, Palestini P (2020) Air pollution and Covid-19: the role of particulate matter in the spread and increase of Covid-19's morbidity and mortality. Int J Environ Res Pub Health 17:4487

Conticini E, Frediani B, Caro D (2020) Can atmospheric pollution be considered a cofactor in extremely high level of SARS-CoV-2 lethality in Northern Italy? Environ Pollut 261:114465

Fiasca F et al (2020) Associations between COVID-19 incidence rates and the exposure to PM2.5 and NO2 a nationwide observational study in Italy. Int J Environ Res. https://doi.org/10.3390/ijerph17249318

Gelman A et al (2013) Bayesian data analysis. CRC Press, USA

Jiang Y, Xu J (2021) The association between COVID-19 deaths and short-term ambient air pollution/meteorological condition exposure: a retrospective study from Wuhan, China. Air Qual, Atmos Health 14:1–5

Jiang X-Q, Mei X-D, Feng D (2016) Air pollution and chronic airway diseases: what should people know and do? J Thorac Dis 8:E31

Karmakar M, Lantz PM, Tipirneni R (2021) Association of social and demographic factors with COVID-19 incidence and death rates in the US. JAMA Netw Open 4:e2036462–e2036462 (**ISSN: 2574-3805**)

Lelieveld J, Münzel T (2019) Air pollution, chronic smoking, and mortality. Euro Heart J 40:3204–3204

Liu P, Beeler P, Chakrabarty RK (2020) Dynamic interplay between social distancing duration and intensity in reducing COVID-19 US hospitalizations: a law of diminishing returns. Chaos: An Interdiscip J Nonlin Sci 30:071102

Neelon B et al (2019) Bayesian zero-inflated negative binomial regression based on pólyagamma mixtures. Bayesian Anal 14:829–855

Ogen Y (2020) Assessing nitrogen dioxide (NO2) levels as a contributing factor to the coronavirus (COVID-19) fatality rate. Sci Total Environ 726:138605

Paolo Giani et al (2020) Short-term and long-term health impacts of air pollution reductions from COVID-19 lockdowns in China and Europe: a modelling study. The Lancet Planet Health. https://doi.org/10.1016/S2542-5196(20)30224-2

Pillow J, Scott J (2012) Fully Bayesian inference for neural models with negative-binomial spiking. Adv Neural Inform Process Syst 25:1898–1906

Polson NG, Scott JG, Windle J (2013) Bayesian inference for logistic models using Pólya-Gamma latent variables. J Am Statist Assoc 108:1339–1349

Van Donkelaar A, Martin RV, Li C, Burnett RT (2019) Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. Environ Sci Technol 53:2595–2611

Venter ZS, Aunan K, Chowdhury S, Lelieveld J (2020) COVID-19 lockdowns cause global air pollution declines. PNAS USA 117:18984–18990

Webb Hooper M, Nápoles AM, Pérez-Stable EJ (2020) COVID-19 and Racial/Ethnic disparities. JAMA 323:2466–2467 (**ISSN: 0098-7484**)

Wu X, Nethery RC, Sabath BM, Braun D, Dominici F (2020) Air pollution and COVID-19 mortality in the United States: strengths and limitations of an ecological regression analysis. Sci Adv 6:eabd4049

Wu X, Nethery RC, Sabath BM, Braun D, Dominici F (2020) Exposure to air pollution and COVID-19 mortality in the United States. medRxiv

Yancy CW (2020) COVID-19 and African Americans. JAMA 323:1891–1892 (**ISSN: 0098-7484**)

Yongjian Z, Jingu X, Fengming H, Liqing C (2020) Association between short-term exposure to air pollution and COVID-19 infection: evidence from China. Sci Total Environ 727:138704