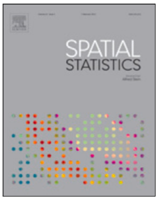
Spatial Statistics 49 (2022) 100526

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



Spatial-temporal generalized additive model for modeling COVID-19 mortality risk in Toronto, Canada

Cindy Feng

Department of Community Health and Epidemiology, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada, B3H 1V7

ARTICLE INFO

Article history:
Received 14 January 2021
Received in revised form 3 June 2021
Accepted 25 June 2021
Available online 6 July 2021

Keywords:
COVID-19 mortality
Generalized additive model
Space–time interaction
Thin plate spline
Tensor product smooth

ABSTRACT

This article presents a spatial–temporal generalized additive model for modeling geo-referenced COVID-19 mortality data in Toronto, Canada. A range of factors and spatial–temporal terms are incorporated into the model. The non-linear and interactive effects of the neighborhood-level factors, i.e., population density and average of income, are modeled as a two-dimensional spline smoother. The change of spatial pattern over time is modeled as a three-dimensional tensor product smoother. By fitting this model, the space–time effect can uncover the underlying spatial–temporal pattern that is not explained by the covariates. The performance of the modeling method based on the individual data is also compared to the modeling methods based on the aggregated data in terms of in-sample and out-of-sample predictive checking. The results suggest that the individual-level based analysis provided a better overall model fit and higher predictive accuracy for detecting epidemic peaks in this application as compared to the analysis based on the aggregated data.

1. Introduction

Coronavirus disease (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), presents an unprecedented threat to global health worldwide. Despite public health responses aimed at slowing the spread of the epidemic, confirmed case counts and hospitalizations continue to surge. To help healthcare professionals prioritize severely ill patients for providing the

E-mail address: cindy.feng@dal.ca.

https://doi.org/10.1016/j.spasta.2021.100526
2211-6753/© 2021 Elsevier B.V. All rights reserved.

best possible care for patients and mitigate the burden on the healthcare system, there is an urgency to understand who is most at risk of mortality, which requires appropriate statistical approaches for the timely analysis of large datasets.

For modeling COVID-19 mortality, most studies are limited to known risk factors such as older age with a high burden of co-morbid diseases (Meyerowitz-Katz and Merone, 2020; Zhou et al., 2020; Du et al., 2020). Nevertheless, COVID-19 mortality risk may also depend on the employment status, education level, income, and housing conditions (de Lusignan et al., 2020; Williamson et al., 2020), which could influence the ability to practice physical distancing measures and seek care. However, the socioeconomic factors remain understudied for modeling COVID-19 data, partly because of a lack of such information. Alternatively, neighborhood-level social factors from census data are often used as a proxy for individual-level social status. Previous research showed that neighborhoods with low socioeconomic status tend to experience higher rates of serious illness if infected due to economic and health inequities (Wadhera et al., 2020). Nevertheless, whether similar patterns have also emerged amid the COVID-19 pandemic in Canada is still understudied.

Further, spatial differences in disease risk are potential indicators of space-related disease factors such as environmental exposures, spatially aggregation of infected individuals and their nonrandom social interactions or availability of sufficient health care in certain areas (Real and Biek, 2007). Quantification of heterogeneity in disease risk patterns over geographical space is, therefore, of importance. While some previous research has noted the importance of spatial autocorrelation (Mollalo et al., 2020; Cordes and Castro, 2020; Zhang and Schwartz, 2020) for modeling risk of COVID-19, the research on modeling the potential changes in spatial patterns over time is still scarce. Providing more accurate time-varying risk maps offers a perspective about which areas may be more affected and when given the time to the central decision-makers to intervene on the local policies.

This study demonstrates a model developed in a generalized additive modeling (GAM) framework (Hastie and Tibshirani, 1990; Wood, 2004, 2017, 2011) is sufficiently flexible to model the spatial–temporal dynamics of COVID-19 mortality risk, which is computationally fast with good discrimination and calibration. Models with three types of link functions for the binomial regression were considered, since misspecification of the link function in binomial regression could result in substantial bias and increased mean squared error of parameter estimates and the predicted probabilities (Czado and Santner, 1992). The non-linear and interactive effects of the neighborhood-level factors, i.e., population density and the average of income, are modeled as a two-dimensional smoother. The spatial–temporal interaction is modeled as a tensor product between a two-dimensional thin-plate regression spline (TPRS) (Wood, 2003) for the centroid of the neighborhoods where the individuals are from and a one-dimensional cubic regression spline for the time variable. The spatial terms can provide insight on detecting high-risk areas not explained by the covariates and contribute to the elucidation of critical aspects of this outbreak, providing a useful perspective in the study region on how the pandemic spreads.

In addition, spatial modeling of COVID-19 data often focuses on analyzing aggregated area-level data to date. The aggregated data are inexpensive to obtain, particularly through data repositories. By combining the data at individual levels at varying levels of aggregation, the size of the data could also be reduced for computational convenience. However, data aggregation may lead to ecological bias when important confounders are missing (Greenland, 2001). In this case, conclusions based on a group-level analysis may differ from those that would have been drawn had an individual-level analysis (Harbarth et al., 2001). Moreover, analyzing aggregated count data can be inherently challenging due to overdispersion, heterogeneity, or the excessive number of zeros. The choice of distributional assumptions for modeling the count outcome variable is crucial in order to draw a valid statistical inference. To date, little work has been conducted in the context of COVID-19 modeling to compare individual-level analysis versus aggregate analysis. Our study aims to fill this gap by demonstrating the potential benefits of using individual-level data compared to using the aggregated data for predictive accuracy. The predictive accuracy of the proposed model is evaluated based on both in-sample and out-of-sample predictive checking.

This paper is organized as follows. In Section 2, details related to the database are provided. In Section 3, the general formulation of the spatial–temporal generalized additive model, as well as

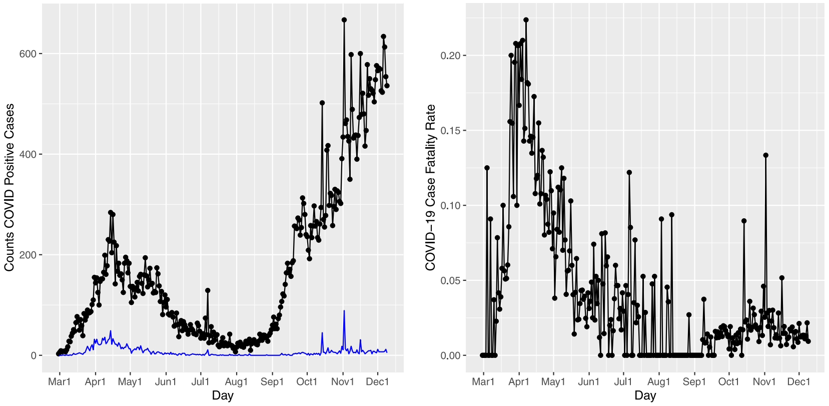


Fig. 1. Left panel: Daily counts of COVID-19 positive cases (black curve) and daily mortality counts of COVID-19 (blue curve). Right panel: Daily COVID-19 case fatality rate. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the model selection ad diagnosis methods, are presented. Section 4 presents the results obtained from the analysis. A comparison of analyses based on individual-level data and aggregated data are carried out with emphasis on predictive accuracy based on in-sample and various schemes of out-of-sample model validations. Finally, Section 5 contains the main conclusions and final remarks.

2. Data description

The City of Toronto in Canada contains 140 neighborhoods that were aggregated from census tracts and created by the Social Policy Analysis and Research Unit in the City's Social Development and Administration Division with assistance from Toronto Public Health.

Data on $n = 49{,}216$ COVID-19 confirmed cases from March 1, 2020, until December 10, 2020, in Toronto, Ontario, Canada, were retrieved from the Ontario Ministry of Health. Of those, 1938 (3.94%) died from COVID-19. The time variable provided in this dataset is a derived/combined variable that best estimates when the disease was acquired, which refers to the earliest available date from symptom onset (the first day that COVID-19 symptoms occurred), laboratory specimen collection date, or reported date. The other predictors include patient demographic (age and gender), ever hospitalized (cases that were hospitalized related to their COVID-19 infection), ever in ICU (cases that were admitted to the intensive care unit (ICU) related to their COVID-19 infection), and ever intubated (cases that were intubated related to their COVID-19 infection). Neighborhood level population density and median household income over the 140 neighborhoods in Toronto were obtained from the 2016 Canadian Census data.

The daily counts of COVID-19 positive cases and mortality counts are displayed in the left panel of Fig. 1, which indicates the COVID-19 infection peaked in April–May 2020, followed by a decline in the summertime and increased substantially in Nov–Dec, 2020. However, such a pattern was not observed in the daily COVID-19 mortality count. As shown in the right panel of Fig. 1, the risk of mortality is much higher in the first wave as compared to the second wave.

3. Methodology

3.1. Model formulation

Let Y_i denote the mortality status due to COVID-19 for the ith individual, $i = 1, \ldots, n$, which follows $Y_i \sim \text{Bernoulli}(\pi_i)$ where π_i denotes the probability of mortality. A spatial–temporal GAM for modeling the mortality risk of COVID-19 is formulated as follows,

$$\begin{aligned} g(\pi_i) = \alpha_0 + \boldsymbol{X}_i\boldsymbol{\beta} + f_t(\text{day}_i) + f_s(\text{lat}_i, \text{long}_i) + \\ f_{ngb}(\text{pop density}_i, \text{income}_i) + f_{st}(\text{lat}_i, \text{long}_i, \text{day}_i), \end{aligned} \tag{1}$$

where $g(\cdot)$ denotes the link function. Three link functions are considered including logistic link function, $g(\pi_i) = \log\{\pi_i/(1-\pi_i)\}$, that corresponds to the inverse cumulative distribution function (CDF) of the standard logistic distribution, probit link function $g(\pi_i) = \Phi^{-1}(\pi_i)$, where $\Phi(\cdot)$ is the standard normal CDF, and complementary log–log (cloglog) function $g(\pi_i) = \log\{-\log(1-\pi_i)\}$, that is formed from the inverse CDF of the Gumbel distribution. The intercept term is denoted as α_0; $\boldsymbol{X}_i$ is a row of the design matrix for any strictly parametric model components, such as age groups, gender and history health care use for COVID-19 condition and $\boldsymbol{\beta}$ is the corresponding parameter vector; $f_t(day_i)$ is modeled as a TPRS function. To simplify the notation, let $f_t(t)$ denote $f_t(day_i)$, which takes the form,

$$f_t(t) = \sum_i \delta_i |t - t_i|^3 + b_1 + b_2 t, \tag{2}$$

where δ_i and b_1 and b_2 are coefficients to be estimated, subject to the identifiability constraints $\sum_i \delta_i = \sum_i \delta_i t_i = 0$. In fitting regression splines, the choice of knot locations and the number of basis functions may have a substantial impact on modeling results. These problems can be avoided by using a relatively large number of basis functions and at the same time imposing a penalty that is designed to ensure that the fitted model is smooth; hence model flexibility is controlled by a smoothing parameter rather than the basis dimension (Ruppert, 2002; Wood, 2000, 2004, 2006, 2008). For the roughness penalty associated with this TPRS basis is,

$$J_t(f_t) = \int \left[\frac{df_t^2(t)}{dt}\right]^2 dt = \boldsymbol{\alpha}^T \boldsymbol{S_t} \boldsymbol{\alpha}. \tag{3}$$

where $\boldsymbol{S}_t$ contains known coefficients and $\boldsymbol{\alpha}$ are the parameters.

The term $f_s(\text{lat}_{ik}, \text{long}_{ik})$ denotes the two-dimensional TPRS applied to the geographical locations for the centroids of the neighborhoods where the individuals are from. For ease of presentation, let $f_s(u, v)$ denote $f_s(\text{lat}_i, \text{long}_i)$. A two-dimensional TPRS of $\boldsymbol{s} = (u, v)^T$ can be defined as

$$f_s(u, v) = \sum_i \delta_i \eta(\|\boldsymbol{s} - \boldsymbol{s}_i\|) + b_1 + b_2 u + b_3 v, \tag{4}$$

where δ_i, $i = 1, \ldots, n$, b_1, b_2 and b_3 are coefficients to be estimated, $\|\cdot\|$ denotes the Euclidean norm, $\eta(r) = \frac{1}{16\pi} r^2 \log(r^2)$ for $r > 0$ and 0 for $r = 0$, subject to identifiability constraints that $\sum_i \delta_i = \sum_i \delta_i u_i = \sum_i \delta_i v_i = 0$. The wiggliness penalty functional of $f_s(u, v)$ is defined as,

$$J_s(f_s) = \int\int \left\{\left(\frac{\partial^2 f_s}{\partial u^2}\right)^2 + 2\left(\frac{\partial^2 f_s}{\partial uv}\right)^2 + \left(\frac{\partial^2 f_s}{\partial v^2}\right)^2\right\} dudv = \boldsymbol{\gamma}^T \boldsymbol{S_s} \boldsymbol{\gamma}, \tag{5}$$

where $\boldsymbol{S}_s$ contains known coefficients and $\boldsymbol{\gamma}$ are the parameters. A major feature of a two-dimensional TPRS is the isotropy of the wiggliness penalty, i.e., smoothing in all directions is treated equally. The isotropy is often considered to be desirable for modeling the interaction between two variables on the same scale, such as geographic coordinates (Wood, 2017). Thin plate regression splines can be seen as a Gaussian process with generalized covariance (Cressie, 1993). Other spatial smoothing techniques could be considered, such as Markov random field (MRF) (Waller et al., 1997; Lawson, 2008; Knorr-Held, 2000). The Bayesian MRF is often fitted using Markov chain Monte Carlo (MCMC) techniques, which may exhibit slow mixing (Christensen et al., 2006). A thin-plate regression spline is applied in this study primarily for its computational efficiency with large datasets (Paciorek, 2007).

The interactive effect between neighborhood-level population density and average income, $f_{ngb}(\text{pop density}_i, \text{income}_i)$ is modeled as a tensor product smoother (Wood, 2006). Tensor product smoother often performs better than isotropic smoother, when the covariates of a smooth are not on the same scale. For ease of presentation, let $f_{p,q}(p, q)$ denote $f_{ngb}(\text{pop density}_i, \text{income}_i)$. Tensor product smoother is formulated by firstly representing smooth functions for each of the bivariate variables as $f_p(p) = \sum_{l_1=1}^{L_1} \zeta_{l_1} b_{l_1}(p)$ and $f_q(q) = \sum_{l_2=1}^{L_2} \delta_{l_2} b_{l_2}(q)$, where $b_{l_1}(p)$ and $b_{l_2}(q)$ are known basis functions and ζ_{l_1} and δ_{l_2} are corresponding basis coefficients. In this study, we used the cubic spline basis function. Creating a smooth function of p and q is achieved by allowing the parameter ζ_{l_1} vary smoothly with q, by representing $\zeta_{l_1}(q) = \sum_{l_2=1}^{L_2} \delta_{l_1 l_2} b_{l_2}(q)$, which gives

$$f_{pq}(p,q) = \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \delta_{l_1 l_2} b_{l_2}(q) b_{l_1}(p), \tag{6}$$

which allows for neighborhood average income to vary smoothly with population density. The penalty of the interaction term is composed of two components. One applies the penalties of the income to the varying coefficient of the marginal population density smooth $\zeta_{l_1}(q)$, $\sum_{l_1=1}^{L_1} J_q \left\{\zeta_{l_1}(q)\right\}$. The second component applies the penalties of the population density smooth to the varying coefficients of the marginal income smooth, $\delta_{l_2}(p)$, $\sum_{l_2=1}^{L_2} J_p \left\{\delta_{l_2}(p)\right\}$. Then, the roughness of $f_{p,q}(p,q)$ can be measured by the sum of the two penalties

$$J(f_{pq}) = \lambda_q \sum_{l_1=1}^{L_1} J_q \left\{\zeta_{l_1}(q)\right\} + \lambda_p \sum_{l_2=1}^{L_2} J_p \left\{\delta_{l_2}(p)\right\}. \tag{7}$$

where λ_q and space λ_p are the smooth parameters for q and p, respectively.

The space–time interaction is constructed as a three-dimensional tensor product smooth of space and time (Augustin et al., 2009), which involves firstly specifying the marginal smooth for time f_t and a two-dimensional marginal smooth for space f_s as $f_t(t) = \sum_{r=1}^{R} \alpha_r a_r(t)$ and $f_s(u,v) = \sum_{m=1}^{M} \beta_m b_m(u,v)$, where α_r and β_m are parameters and a_r and b_m are basis functions with $r = 1, \ldots$, and $m = 1, \ldots, M$. The interaction between space and time is constructed by allowing the temporal smooth f_t to varying smoothly within the space dimensions u and v by specifying $\alpha_r(u,v) = \sum_{m=1}^{M} \beta_{rm} b_m(u,v)$, which yields

$$f_{st}(u,v,t) = \sum_{r=1}^{R} \sum_{m=1}^{M} \beta_{rm} b_m(u,v) a_r(t), \tag{8}$$

where β_{rm} are the coefficients of basis functions of tensor product smooths. The penalty of the space–time interaction term is composed of two components. One applies the penalties of the spatial smooth to the spatially varying coefficient of the marginal temporal smooth $\alpha_r(u,v)$, $\sum_{r=1}^{R} J_s \{\alpha_r(u,v)\}$. The second component applies the penalties of the temporal smooth to the temporally varying coefficients of the marginal spatial smooth, $\beta_m(t)$, $\sum_{m=1}^{M} J_t \{\beta_m(t)\}$. Then, the roughness of $f_{st}(u,v,t)$ can be measured by the sum of the two penalties

$$J(f_{st}) = \lambda_s \sum_{r=1}^{R} J_s \{\alpha_r(u,v)\} + \lambda_t \sum_{m=1}^{M} J_t \{\beta_m(t)\}. \tag{9}$$

where λ_s and space λ_t are the smooth parameters for space and time, respectively.

3.2. Parameter estimation

The model can be written as $g(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\theta}$, where $g(\cdot)$ is the link function; $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ is a n-vector; $\mathbf{X}$ is the full design matrix and $\boldsymbol{\theta}$ contains all the parameters need to be estimated. To estimate the parameters $\boldsymbol{\theta}$, we maximize the penalized log-likelihood $l_p(\boldsymbol{\theta}|\mathbf{y}) = l(\boldsymbol{\theta}|\mathbf{y}) - \frac{1}{2}\sum_{j=1}^{J} \lambda_j J_j(f_j)$, conditional on smoothing parameters λ_j, operating on the penalty terms J_j, $j = 1, \ldots, J$, where J denotes the total number of penalty terms. Here $l(\boldsymbol{\theta})$ is the log-likelihood associated with the Bernoulli response: $l(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \Big\{ y_i \log\left[p(y_i = 1|\boldsymbol{\theta})\right] + (1 - y_i) \log\left[1 - p(y_i = 1|\boldsymbol{\theta})\right] \Big\}$. The smoothing parameters λ_j control the tradeoff between the goodness of fit of the model and model smoothness, with $J_j(f_j)$, $j = 1, \ldots, J$ penalizing sharp changes of the splines.

To estimate the parameters, penalized likelihood is maximized by a penalized iteratively reweighted least squares (P-IRLS) algorithm. Given the smoothing parameter, at the kth P-IRLS iteration, the following penalized sum of squares would be minimized with respect to $\boldsymbol{g} = X\boldsymbol{\theta}$ to find the $(k+1)$th estimate $\boldsymbol{g}^{[k+1]} = X\boldsymbol{\theta}^{[k+1]}$:

$$\sum_{i=1}^{n}\left\{w_i^{[k]}(z_i^{[k]}-g_i)\right\}^2+\sum_{j=1}^{J}\lambda_j J_j(f_j)\,, \tag{10}$$

where $z_i^{[k]} = g_i^{[k]} + g'(\mu_i^{[k]})(y_i - \mu_i^{[k]})$, and $w_i^{[k]} = 1/\sqrt{V(\mu_i^{[k]})g'(\mu_i^{[k]})^2}$ and $V_i^{[k]}$ is proportional to the variance of Y_i according to the current estimate $\mu_i^{[k]}$.

The smoothing parameters λ_j are estimated by minimizing the generalized cross-validation score (Craven and Wahba, 1979; Wahba, 1985) for each working penalized linear model of the P-IRLS iteration. The score has the following form:

$$\mathrm{GCV} = \frac{n \parallel \sqrt{\mathbf{W}}(\mathbf{z}-\mathbf{X}\boldsymbol{\theta}) \parallel^2}{[n-\mathrm{tr}(\mathbf{A})]^2}\,, \tag{11}$$

where $\mathbf{A} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X}+\mathbf{S})^{-1}\mathbf{X}^T\mathbf{W}$ is the influence matrix (Hastie and Tibshirani, 1990) and $\mathrm{tr}(\mathbf{A})$ is the effective degrees of freedom, or the effective number of parameters, of the model; $\mathbf{z}$ is a vector of pseudodata z_i and $\mathbf{W}$ is a diagonal matrix with elements w_i, defined previously, all of these values would be calculated based on estimates of related quantity at each iteration (Hastie and Tibshirani, 1990; Wood, 2004).

Defining $\boldsymbol{v} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z}$, based on large sample approximation, the distribution of $\boldsymbol{\beta}$ is derived as,

$$\boldsymbol{\beta}|\boldsymbol{v} \sim N([\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}+\sum_j \lambda_j \boldsymbol{S}_j]^{-1}\boldsymbol{v}, [\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}+\sum_j \lambda_j \boldsymbol{S}_j]^{-1}\phi) \tag{12}$$

where ϕ stands for scale parameter. In logistic regression, $\phi = 1$. The uncertainty or confidence interval of $\boldsymbol{\beta}$ can be estimated by plugging in the $\boldsymbol{v}$ and $\boldsymbol{W}$ estimates at convergence of R-IRLS algorithm, along with the smoothing parameters. For more details of estimation for generalized additive models, refer to Hastie and Tibshirani (1990) and Wood (2004).

3.3. Model selection and diagnosis

3.3.1. Model selection

Models with different link functions were considered, including logistic, probit, and cloglog link functions. Under each model, model selection was carried out by adding an extra penalty to each smooth term that penalizes functions in the null space of the penalty matrices for each smooth (Marra and Wood, 2011). This is carried out by using the argument `select = TRUE` in the `gam` function in `R`. After model selection, the choice of the link function was evaluated by checking the model fit based on AIC (Akaike, 1973), deviance, percentage of deviance explained, area under the ROC curve (AUC) (Harrell, 2015) and Brier's score (Rufibach, 2010; Harrell, 2015). AUC measures the discrimination ability of the model. Higher values of the AUC indicate better model discrimination. AUC can examine the ability of the method to distinguish the patients who have the outcome have higher risk predictions than those who do not, but cannot account for calibration, i.e., the magnitude of the disagreement between the observed and predicted responses (Harrell, 2015). To quality how close the predictions are to the actual outcome, Brier's score (Rufibach, 2010; Harrell, 2015), the mean squared prediction error, is defined as, $1/n\sum_{i=1}^{n}(Y_i-\hat{\pi}_i)^2$, where Y_i and $\hat{\pi}_i$ denote the actual outcome of the event and the predicted probability of the event, respectively. Lower Brier's scores indicate greater model accuracy. After determining the link function, the relative contribution of each factor for predicting the COVID-19 mortality risk is evaluated by examining the change of model fit by excluding each model term from the full model one at a time.

3.3.2. Residual diagnosis

Residual plots are often used to check the assumptions of regression models; however, residual plots from logistic regression based on deviance or Pearson residuals are generally not informative due to the discreteness of the data (Gelman et al., 2000). Randomized quantile residuals (RQRs) (Dunn and Smyth, 1996) were proposed in the literature to diagnose models for discrete outcomes. However, the power of RQRs for diagnosing discrete models with a limited number of unique values in the outcome variable, such as binary outcome, tends to be low due to the randomness introduced to the estimated cumulative distribution function (Feng et al., 2020). To overcome these challenges, the simulated envelope is recommended to be added to the quantile–quantile (QQ) plot of deviance or Pearson residuals to detect overall departures from the model assumptions as well as outlying observations of the fitted model. For a well-fitted model, most of the residuals are expected to fall within the simulated envelope. The graphical display of residuals versus fitted values or predictors can provide further focused diagnostics. An alternative standard way to make discrete residual plots interpretable is to work with binned residuals (Gelman et al., 2000), which divides the data into categories (bins) of equal size based on their fitted values. Then, the average residual is plotted against the average fitted value for each bin. If the model were true, the residuals should be closer to symmetric about zero, and one would expect about 95% of the residuals to fall inside the error bounds.

To examine if there is spatial autocorrelation remained in the residuals, a permutation test for Moran's I statistic (Moran, 1950; Cliff and Ord, 1981) is used using the spdep package (Bivand and Wong, 2018) in R, calculated by using 999 random permutations of the data over the study region, to establish a distribution of simulated Moran's I based on a null hypothesis of spatial randomness. The observed value of Moran's I is then compared with the simulated distribution. The temporal autocorrelation in the residuals was assessed by autocorrelation function (ACF) plot (Box and Jenkins, 1976) and partial autocorrelation (PACF) plots (Venables and Ripley, 2002), which compute autocorrelations for data values at varying time lags. If random, such autocorrelations should be near zero for any time-lags. If non-random, one or more of the autocorrelations will be significantly non-zero.

3.4. Predictive accuracy assessment

Using modeling techniques to predict the COVID-19 mortality could help effectively guide public health policy-making to optimize the allocation of the limited resources. In literature, some modeling methods are based on individual-level data (Das et al., 2020; Rosenthal et al., 2020; Xu et al., 2020; Yu et al., 2020) and some are based on aggregated data (Talukder, 2020; Ujiie et al., 2020; Okeahalam et al., 2020; Chaudhry et al., 2020; Karmakar et al., 2021; Lima et al., 2021). The predictive accuracy of the models based on individual versus aggregated data has not been well understood. As a result, this study also examined the accuracy and precision of predictions by comparing the selected best-fitting model based on the individual-level data against commonly used count regression models for modeling the aggregated daily COVID-19 mortality count.

3.4.1. Modeling aggregated daily COVID-19 mortality count

To avoid the potential loss of critical information in the process of data aggregation, the individual-level data were aggregated according to the unique combinations of the levels of the predictors considered in this study, i.e., day, neighborhoods, age groups, gender, ever been hospitalized, ever been in ICU and ever been intubated. The neighborhood-level characteristics, i.e., population density and average income, were linked to this data by neighborhood. By doing so, all the predictors considered in the analysis based on the individual-level data are included in the regression models for modeling the aggregated daily death counts. The aggregation units without COVID-19 positive cases were removed. The final dataset contains 40,882 aggregate units. The outcome variable is the number of deaths within each aggregate unit, which ranges from 0 to 19. The logarithm of the number of positive cases within each aggregate unit is included as an offset term in the count regression models.

Modeling count data involves many challenges, especially when the data exhibits a preponderance of zeros and overdispersion. Although alternatives exist for modeling overdispersed count data with an excess of zeros, they are not widely and effectively applied and implemented in standard software. In this study, Poisson, Negative Binomial (NB), and zero-inflated Poisson (ZIP) (Lambert, 1992) models are considered. One major limitation of the Poisson regression is that it assumes the mean and variance conditional on the covariates are the same, which can be restrictive in some applications. NB extends the Poisson by positing that the conditional mean of the response variable is not only determined by the covariates but also a latent component independent of the covariates. Although the NB model could model the overdispersed Poisson data, it is not appropriate for modeling the data with a high percentage of zero counts as in the current context. ZIP is a mixture of two statistical processes, with one always generating "structural" zero counts and the other both "sampling" zero and positive counts. That is, it assumes that each observation comes from one of two potential distributions, with one consisting of a constant zero while the other following Poisson. In a ZIP model, a logit model is typically used to model the probability of the excessive zeros, while the Poisson regression models the rest of the count data.

3.5. *Evaluation of the prediction performance*

The predictive accuracy of the models is evaluated based on both in-sample and out-of-sample prediction in terms of Root Mean Squared Prediction Error (RMSPE) as a measure of the agreement between the observed binary outcome and the predicted probability of the outcome. In-sample analysis means to estimate the model using all available data, and then compare the model's fitted values to the actual realizations. The out-of-sample predictions are evaluated based on (1) K-fold cross-validation, (2) moving window forecasting, (3) leave-one-location-out cross-validation, and (4) leave-one-location-out moving window forecasting cross-validation.

3.5.1. *In-sample predictive accuracy check*

To ensure the individual and aggregate analyses are comparable, the predicted probabilities of COVID-19 mortality based on the individual-level data analysis for the individuals from the same aggregate unit are summed together, which gives the estimated counts of mortality in that aggregate unit.

The RMSPE for the in-sample predictive accuracy is defined as, i.e.,

$$\mathrm{RMSPE}^{I} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(Z_j - \hat{Z}_j)^2}, \tag{13}$$

where Z_j and $\hat{Z}_j$ represent the observed and predicted death counts in the jth aggregate unit, respectively, $j = 1, \ldots, J$ with J denoting the total number of aggregate units. For the analysis based on the individual analysis, the observed and predicted death count in the jth aggregate unit are $Z_j = \sum_{i \in A_j} Y_i$ and $Z_j = \sum_{i \in A_j} \hat{Y}_i$, respectively, where A_j denotes the set of individual observations belonging to the jth aggregate unit.

3.5.2. *K-fold out-of-sample cross-validation*

In general, k-fold cross validation can be defined as follows:

- Randomly divide data into K equal groups. Let A_k denote the set of data points $(Z_j, \boldsymbol{X}_j)$ placed into the k'th fold.
- For $k = 1, \ldots, K$, train model was fitted on all except kth fold. Let $\hat{f}^{-k}$ denote the resulting predicted probability of the fitted model based on the training dataset A_k, $k = 1, \ldots, K$.
- RMSPE^{CV} can be then written as

$$RMSPE^{CV} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{J/K}\sum_{j \in A_k}(Z_j - \hat{f}^{-k}(\boldsymbol{X}_j))^2}. \tag{14}$$

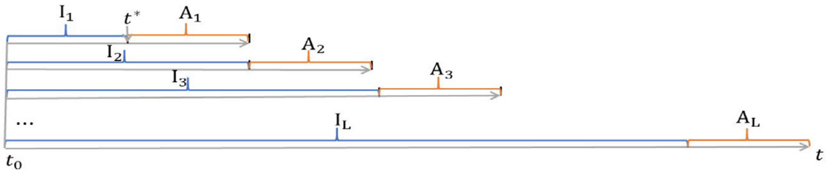


Fig. 2. An illustration of the moving window forecast process, where I_l and A_l, $l = 1, \ldots, L$ denote the training and testing datasets, respectively.

3.5.3. *Moving window forecasting predictive check*

The predictive ability of the model is further evaluated by its forecasting accuracy, which is evaluated by comparing the forecasts against future realizations of the data. Fig. 2 illustrates the moving window forecast process considered in this study, with $l = 1, \ldots, L$, indexing the forecast window, where L denotes the number of forecast windows during the study period. I_l denotes the lth training dataset, including all observations from the beginning observation period up till the starting time point of the lth forecast window. The starting time point of forecasting, t^*, is set as 120 days since the beginning of the study period allowing for enough observations for reliably fitting models to the training dataset. The lth moving window forecast dataset is defined as A_l. The forecast horizon, i.e., length of time of forecasting into the future, ρ is set as 7, 14, 21, and 28 days, respectively.

The RMSPE at the lth forecasting window is defined as

$$RMSPE_l^F = \sqrt{\frac{1}{n_l}\sum_{j \in A_l}[Z_j - \hat{f}(\boldsymbol{X}_{I_l})]^2}, \tag{15}$$

where n_l represents the number of observation units in A_l and $\hat{f}(\boldsymbol{X}_{I_l})$ represent the predicted death count for window A_l based on the train dataset I_l. The average RMSPE over all the L forecast windows can be then defined as $RMSPE^F = \frac{1}{L}\sum_{l=1}^{L} RMSPE_l^F$.

3.5.4. *Leave-one-location-out cross validation*

As the data are spatially correlated, leave-one-location-out (LOLO) cross-validation is also used to predict the time series at an unobserved neighborhood. The average RMSPE for LOLO CV across all the neighborhoods is defined as,

$$RMSPE^{LOLO} = \frac{1}{S}\sum_{s=1}^{S}\sqrt{\frac{1}{n_s}\sum_{j \in A_s}[Z_j - \hat{f}(\boldsymbol{X}_{I_{(-s)}})]^2}, \tag{16}$$

where A_s denotes the testing data, which is the entire time series of observations at the $s$the neighborhood, $s = 1, \ldots, S$, and n_s represents the total number of aggregate units at the sth neighborhood. $I_{(-s)}$ denotes the training dataset for the sth neighborhood, which includes the observations from all neighborhoods except for the sth neighborhood. Defining each CV fold as the observations taken at one neighborhood ensures that predictions are always being evaluated at a neighborhood excluded from the data on which the prediction was trained.

3.5.5. *Leave-one-location-out moving window forecasting cross validation*

In Section 3.5.4, the forecasting was made for the entire time series of a location, which does not effectively evaluate the model performance at predicting the future. The RMSPE for evaluating spatial–temporal extrapolation error is therefore examined, which is defined as

$$RMSPE^{F-LOLO} = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{S}\sum_{s=1}^{S}\left\{\frac{1}{n_{sl}}\sum_{j \in A_{sl}}\left[Z_j - \hat{f}(\boldsymbol{X}_{I_{(-s)l}})\right]^2\right\}}, \tag{17}$$

Table 1
Comparison of the model fits in terms of AIC, deviance (dev), percentage of deviance explained (dev.expl), AUC, and Brier's score.

	Logit	Probit	Cloglog
AIC	8175.273	8105.895	8293.470
dev	7942.698	7878.747	8055.724
dev.expl	51.4%	51.8%	50.7%
AUC	0.9668	0.9672	0.9661
Brier	0.0251	0.0251	0.0252

Table 2
Evaluation of the contribution of each predictor for modeling the COVID-19 mortality risk by excluding each predictor from the full model, in terms of AIC, deviance, percentage of deviance explained (dev.expl), AUC and Brier's score.

Model	AIC	Deviance	dev.expl	AUC	Brier's
Full	8105.895	7878.747	51.8%	0.967	0.025
-pop × income	8145.149	7936.670	51.4%	0.967	0.025
-space × time	8184.995	8087.861	50.5%	0.965	0.026
-gender	8191.274	7971.685	51.2%	0.966	0.025
-space	8252.505	8176.207	49.9%	0.964	0.026
-time	8352.776	8120.091	50.3%	0.964	0.026
-health condition	9518.019	9355.916	42.7%	0.947	0.028
-age	11 115.384	10 817.703	33.8%	0.925	0.031

where n_{sl} denotes the total number of aggregate units at the sth neighborhood and the lth forecast window; A_{sl} denotes the lth forecast window at the sth neighborhood and $I_{(-s)l}$ represents the training dataset, which contains the observations prior to the lth forecast window after excluding the observations from the sth neighborhood, and $\hat{f}(\boldsymbol{X}_{I_{(-s)l}})$ denotes the predicted disease counts at the sth neighborhood and the lth forecast window.

4. Results

4.1. Model selection and diagnosis

4.1.1. Model selection

After the model selection, all the predictors remain significant regardless of which link function is applied. The results of the model comparison (Table 1) indicate that the model with the probit link function outperforms the models with the logit or cloglog link functions, which yielded the lowest AIC and deviance, highest percentage of deviance explained, highest AUC and lowest Brier's score.

Evaluation of the relative contribution of each predictor for modeling the COVID-19 mortality risk is also conducted by excluding each predictor from the full model. The results (Table 2) indicate that age or history health conditions for COVID-19 are the strongest predictors for mortality status since excluding either of the two variables results in the most substantial change in the measures of model fit. Space, time, and space–time interaction terms also improve the model fit. The model with continuous space–time interaction performs better compared with the model with an additive space–time effect. In addition, accounting for the neighborhood characteristics improves the predictive ability of the model.

4.1.2. Residual diagnosis

The QQ plots of the deviance residuals for the logistic, probit, and cloglog models with simulated envelopes are displayed in Fig. 3. As displayed, for the logistic and cloglog models, a few residuals at a higher range of the residuals fall outside of the simulated envelopes. By contrast, the points in the QQ-plots based on the probit model mostly fall within the simulated envelopes, which indicates the probit model fits the data well.

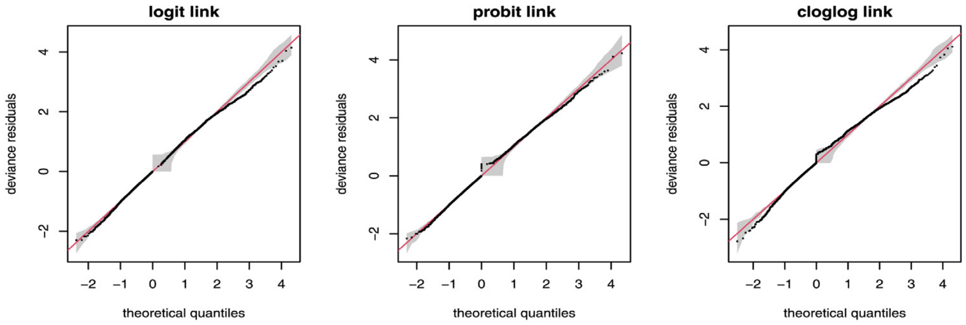


Fig. 3. QQ plots of the deviance residuals for the logit, probit and cloglog models. The gray areas are the simulated envelopes based on 1000 simulated data from the fitted model. For a well-fitted model, most of the residuals are expected to fall within the simulated envelope.

The binned residual plots against the fitted values and the predictors are displayed in Fig. 4. The top panels are the average residuals versus average fitted values for each bin under the models with the different link functions. The results revealed that the probit model provided an adequate model fit with almost all the residuals falling within the simulated envelope. By contrast, for logistic and cloglog models, quite a few residuals fall outside of the simulated envelopes. A closer examination of the binned residual plots against the continuous predictors indicates all the models underpredict the mortality risk at one-time point during the second wave, but not substantially. Both the QQ plots and the binned residual plots suggest the probit model provides a better fit to the data as compared to the logistic and cloglog models.

The examination of the temporal and spatial autocorrelation in the residuals of the space–time probit model is displayed in Fig. 5. The ACF and PACF plots do not reveal any significant autocorrelation in the residuals, as shown in the first two panels of Fig. 5. The third panel of Fig. 5 suggests no spatial correlation remained in the residuals with a pseudo p-value of 0.953.

4.2. Estimated model components

The results of the parametric terms for the full model are presented in Table 3 and the non-parametric spline terms for the full model are presented in Figs. 6 and 7. As shown in Table 3, the risk of mortality has no significant difference for age groups below 40–49 years, but the risk increases drastically for the age groups 50–59 years and onward. The mortality risk is significantly higher among males and those who had ever been hospitalized, ever in ICU, or ever intubated.

The left panel of Fig. 6 displays the estimated smooth term of the reporting date of COVID-19 on the logit of mortality probability. The plot indicates that the mortality risk increased sharply and peaked around April–May 2020, followed by a decline during summer and then increased again in fall, but the peak is lower than the first peak. The right panel of Fig. 6 displays the two-dimensional tensor product smoother of the population density and average income, which clearly shows their non-linear interactive effect. More specifically, the risk of mortality tends to be much higher for neighborhoods with larger population sizes and lower average incomes. The risk of mortality is the lowest for the neighborhoods with very high average income and moderately low population density.

The spatial–temporal interaction terms (Fig. 7) reflect how the spatial pattern of mortality risk evolves over time after accounting for known risk factors. The results indicate localized differences in the underlying attributes of neighborhoods can have a statistically significant impact on COVID-19 mortality risk. The shifting of the high-risk regions might be attributable to the interventions such as lockdowns and social distancing practices, which reduced the community spread of infection in some areas. Additionally, geographic disparity of health care capability may exist, such as access

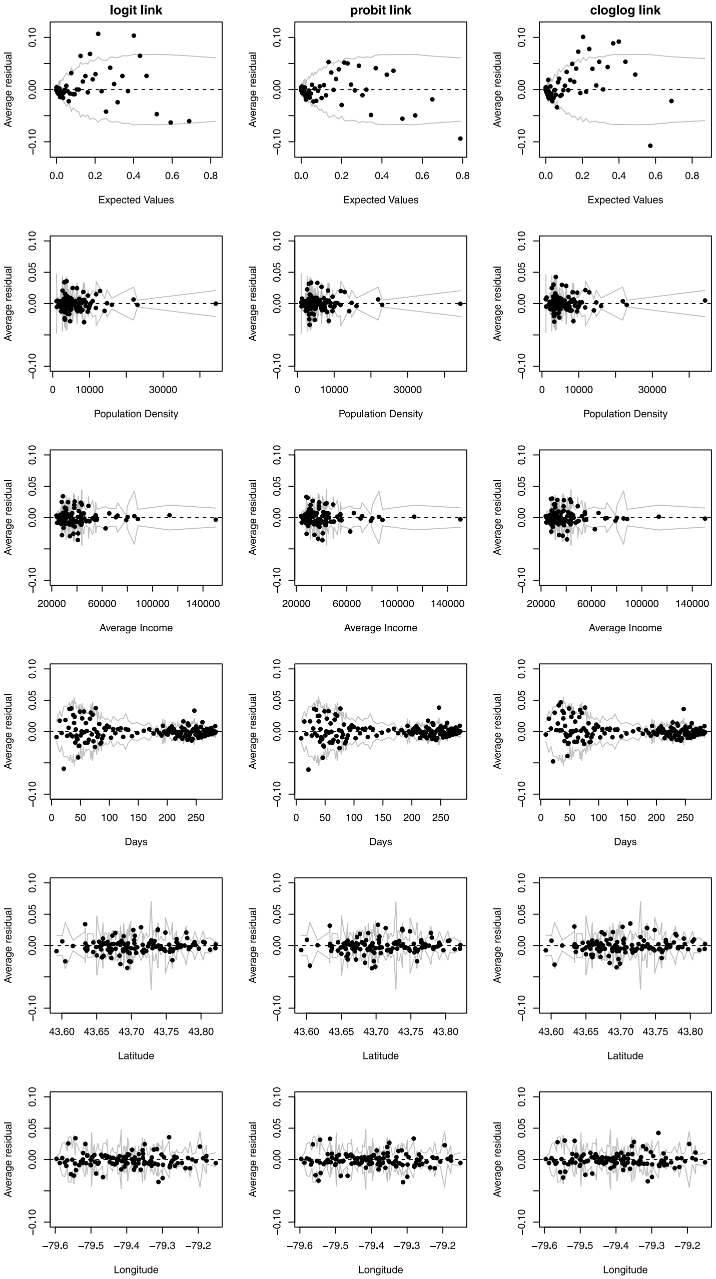


Fig. 4. Binned residual plots of the full models with logit, probit and cloglog link functions, respectively. The gray lines indicate plus and minus 2 standard-error bounds, within which one would expect about 95% of the binned residuals to fall, if the model were true.

to clinics, ICU beds, ventilators, or drugs and equipment to effectively treat people most severely affected by COVID-19 infection complications.

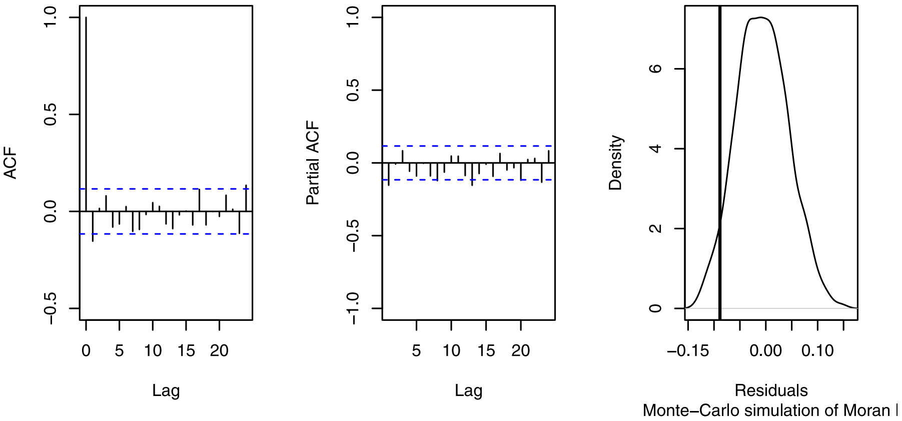


Fig. 5. Autocorrelation (ACF) and partial autocorrelation (PACF) plots for examining temporal autocorrelation in residuals and the Monte-Carlo simulation of Moran's I statistic for examining spatial autocorrelation in residuals (p-value = 0.953).

Table 3
Significance of estimates from the space–time probit model. SE stands for the standard error of the parameter estimate. "edf" represents the effective degrees of freedom of the functional parameters.

Parametric terms	Estimate	SE	p-value
Age (Reference: 19 years and below)			
20 to 29 Years	−0.448	0.438	0.306
30 to 39 Years	−0.132	0.359	0.714
40 to 49 Years	0.119	0.315	0.706
50 to 59 Years	0.724	0.291	0.013
60 to 69 Years	1.278	0.287	0.000
70 to 79 Years	1.889	0.286	0.000
80 to 89 Years	2.397	0.286	0.000
90 and older	2.828	0.286	0.000
Gender (Reference: Females)			
Male	0.314	0.034	0.000
Others	0.152	0.143	0.288
Health history (Yes vs. No)			
Ever Hospitalized	0.938	0.037	0.000
Ever in ICU	0.623	0.096	0.000
Ever Intubated	0.549	0.112	0.000
Smooth terms	**edf**		**p-value**
f_t(day)	8.184		<0.001
f_{ngb}(pop density, income)	12.994		<0.001
f_s(lat, long)	25.327		<0.001
f_{st}(lat, long, day)	53.069		0.0028

4.3. Predictive accuracy assessment

4.3.1. Predictive models based on aggregated data

This section compares the proposed model based on the individual-level data with the count regression models based on the aggregated data constructed based on the individual-level data. Fig. 8 presents the observed versus predicted daily mortality count based on the probit model for modeling the individual-level data and Poisson, NB, and ZIP models for modeling the aggregated data. The results clearly show that the count regression models severely underestimate the higher

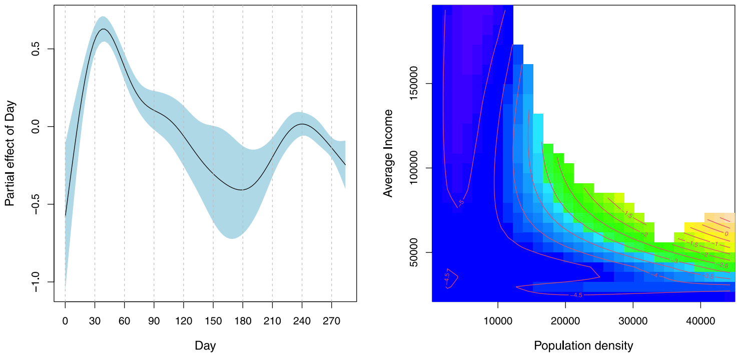


Fig. 6. The left panel displays the smooth function of day and the right panel shows the smooth function of the population density and average income of the space–time probit model.

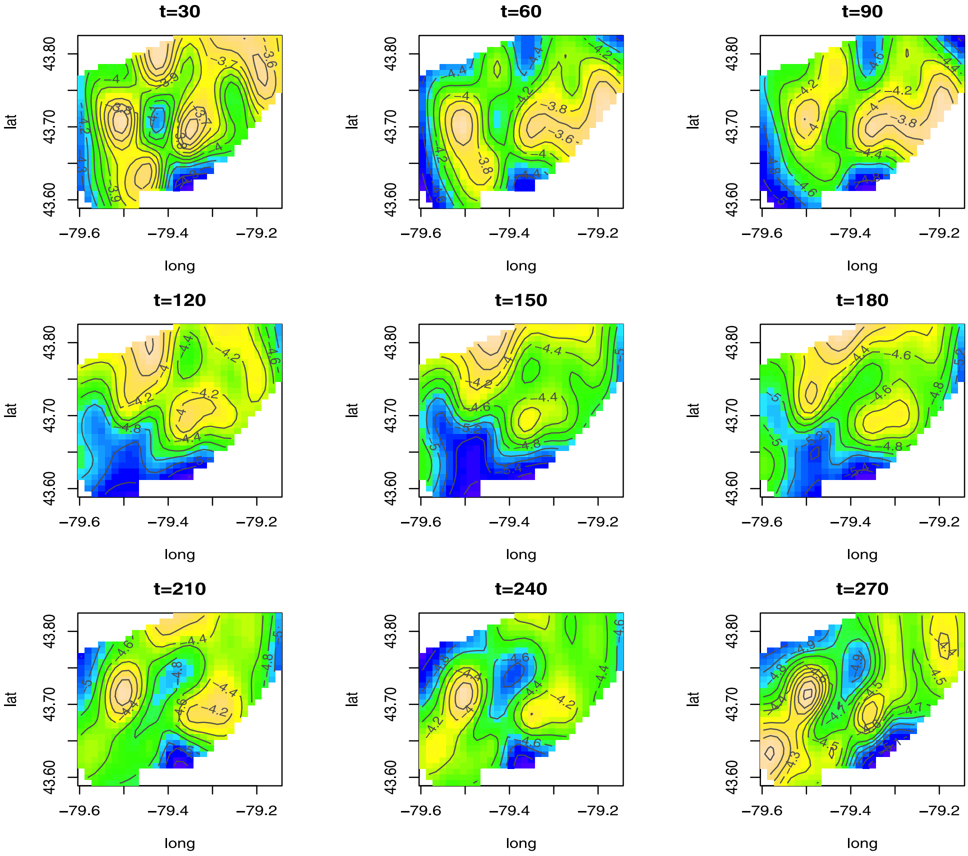


Fig. 7. The estimated partial effect of space–time interaction term at specified times *t* on the logit of probability of COVID-19 mortality.

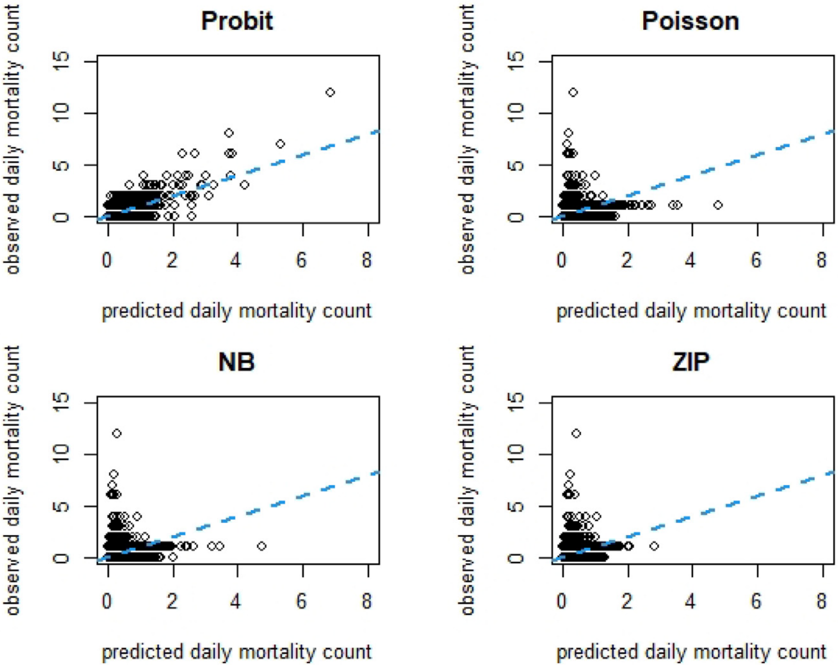


Fig. 8. Predicted versus observed daily mortality counts based on the probit, Poisson, NB and ZIP models, respectively.

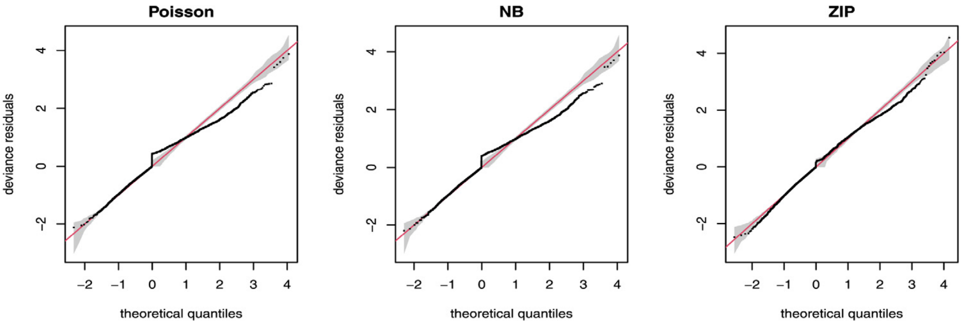


Fig. 9. QQ-plots of deviance residuals for Poisson, NB and ZIP models, respectively. The gray areas are the simulated envelopes based on 1000 replicates from the fitted models.

values of daily mortality count. In other words, count regression models could adequately detect epidemic peaks. In contrast, the probit model provides a much better fit to the data, with the predicted mortality counts much closer to the observed mortality counts. The probit model also under-predicts the higher values of daily mortality counts, but the underestimation is not as pronounced as the count regression models. Fig. 9 presents the QQ-plots of the deviance residuals of the count regression models with simulated envelopes, which confirmed the inadequacy of the count regression models with larger values of residuals falling outside of the simulated envelopes. This phenomenon is more pronounced for Poisson and NB models.

Table 4 presents the parameter estimates of the ZIP model. Similar results were observed based on the Poisson and NB regression, which are not presented here. The results show that all the

Table 4
Model estimates of the ZIP model for modeling the daily COVID-19 mortality count. SE stands for the standard error of the parameter estimate. "edf" represents the effective degrees of freedom of the functional parameters.

Parametric terms	Estimate	SE	p-value
Age (Reference: 19 years and below)			
20 to 29 Years	−0.652	1.293	0.614
30 to 39 Years	0.129	1.120	0.908
40 to 49 Years	1.677	0.955	0.079
50 to 59 Years	2.993	0.922	0.001
60 to 69 Years	4.137	0.918	0.000
70 to 79 Years	5.084	0.917	0.000
80 to 89 Years	5.788	0.916	0.000
90 and older	6.362	0.916	0.000
Gender (Reference: Females)			
Male	0.354	0.046	0.000
Others	0.159	0.190	0.402
Health History (Yes vs. No)			
Ever Hospitalized	1.143	0.050	0.000
Ever in ICU	0.731	0.112	0.000
Ever Intubated	0.628	0.127	0.000
Smooth terms	edf		p-value
f_t(day)	7.525		<0.001
f_{ngb}(pop density, income)	7.548		<0.001
f_s(lat, long)	21.434		<0.001
f_{st}(lat, long, day)	18.886		0.1

significant predictors in analysis based on the individual-level data remain significant in the ZIP model, except for the space–time interaction term. As a result, the analysis based on aggregated data cannot reliably elucidate the space–time interactive effect, further demonstrating the potential benefits of the individual-level analysis for gains in power and efficiency for detecting space–time interaction. The inferior performance of the count regression models based on the aggregated data compared to the probit model based on the individual-level data likely stems from the assumptions of distributions for modeling count data.

4.3.2. Evaluation of predictive performance

Table 5 shows that RMSPE based on the in-sample, 10-fold CV, and moving window forecast, leave-one-location out CV and leave-one-location out moving window forecast CV for the probit, Poisson, NB, and ZIP models. The results indicate the probit model had the smallest overall RMSPE as compared to the count regression models. Among the count regression models, the ZIP model had better predictive accuracy than the other methods in most of the cases, except for RMSPE^F and RMSPE^{F-LOLO}. This implies the performance of the count regression models heavily depends on the schemes of the model validation.

For completeness, we also compared the probit and count regression models for RMSPE^F and RMSPE^{F-LOLO} over the moving forecasting windows. As displayed in Fig. 10, both RMSPE^F and RMSPE^{F-LOLO} for all models decrease during summer 2020, followed by an increase during the second wave. The RMSPE^F and RMSPE^{F-LOLO} for count regression models tend to be higher than the probit model during the second peak time. Such differences increases as the forecasting window size increases. The results imply the count regression models are less accurate for predicting the peak mortality risk as compared to the probit model in this investigation.

5. Discussion

This study demonstrated a spatial–temporal generalized additive model for modeling and predicting COVID-19 mortality risk in the study region, which could help improve resource allocation

Table 5
RMSPE of the probit model based on the individual-level data and the count regression models (Poisson, NB and ZIP models) based on the aggregated data.

	Probit	Poisson	NB	ZIP
RMSPE^I	0.1815	0.2381	0.2388	0.2336
RMSPE^{CV}	0.1629	0.2354	0.2453	0.2299
RMSPE^{LOLO}	0.1697	0.1816	0.1902	0.1841
RMSPE^F				
7 days	0.1422	0.1507	0.1491	0.1511
14 days	0.1456	0.1586	0.1572	0.1586
21 days	0.1512	0.1622	0.1602	0.1583
28 days	0.1459	0.1768	0.1752	0.1728
RMSPE^{F-LOLO}				
7 days	0.1162	0.1202	0.1182	0.1189
14 days	0.1257	0.1342	0.1324	0.1341
21 days	0.1360	0.1443	0.1416	0.1422
28 days	0.1374	0.1543	0.1526	0.1532

RMSPE^I: RMSPE for in-sample prediction
RMSPE^{CV}: RMSPE for 10-fold CV
RMSPE^{LOLO}: RMSPE for leave-one-location-out CV
RMSPE^F: RMSPE for moving window forecast
RMSPE^{F-LOLO}: LOLO CV for moving window forecast.

across sub-populations. The study adds to the evidence showing the age of 50 years, and onward, males and those who had ever been hospitalized, in ICU, or intubated are at a significantly higher mortality risk due to COVID-19 at the study region during the study period. Neighborhood level population density and average income also have a significant interactive effect on COVID-19 mortality risk. Moreover, we demonstrated a powerful and computationally efficient approach for uncovering the complexity of the inter-correlated nature of space–time COVID-19 data. The identified spatial–temporal effect may be driven by the neighborhood-specific characteristics, such as social network and interactions, regulations and compliances of social-distancing policies, or other socio-demographic information that are not captured by the neighborhood population density and average income.

In this study, the model diagnostic indicated the independent error structure was adequate, and no spatial–temporal heterogeneity was identified in the residuals of the best fitting model. However, in some applications, if residuals exhibit spatial or temporal autocorrelation, generalized additive mixed-effect model (GAMM) (Wood, 2006) including random effects with spatial or temporal autocorrelation error structures should be considered. The random effect terms could properly quantify the variability that was unexplained by the predictors in the model. However, the computation for GAMM is typically slower than GAM, and not as numerically robust (Wood, 2006).

The space–time smoother is set up using the multi-dimensional tensor product smoother, with the marginal spatial and temporal bases being scale-invariant. This provides flexibility by allowing for different degrees of smoothness for each dimension. The alternative approach of modeling the spatial correlation could also be considered, such as using Markov Random Field (MRF), also called conditional autoregressive prior for spatial random effect term, which is often used for analyzing spatially aggregated data (Waller et al., 1997; Lawson, 2008; Knorr-Held, 2000). Such models are often implemented in the Bayesian framework, which offers the advantage of using prior information to improve inference. In fact, MRF approximation to a thin plate spline that involves only nearby grid cells as neighbors (Rue and Held, 2005; Yue and Speckman, 2010). However, the areal unit problem may arise from using political boundaries that are arbitrarily related to public health. Bayesian models are also more difficult to specify, especially for investigating space–time interaction on a daily basis over a long period. Bayesian estimation methods also require examining prior distributions through a sensitivity analysis. The fitting process may fail in ways that are

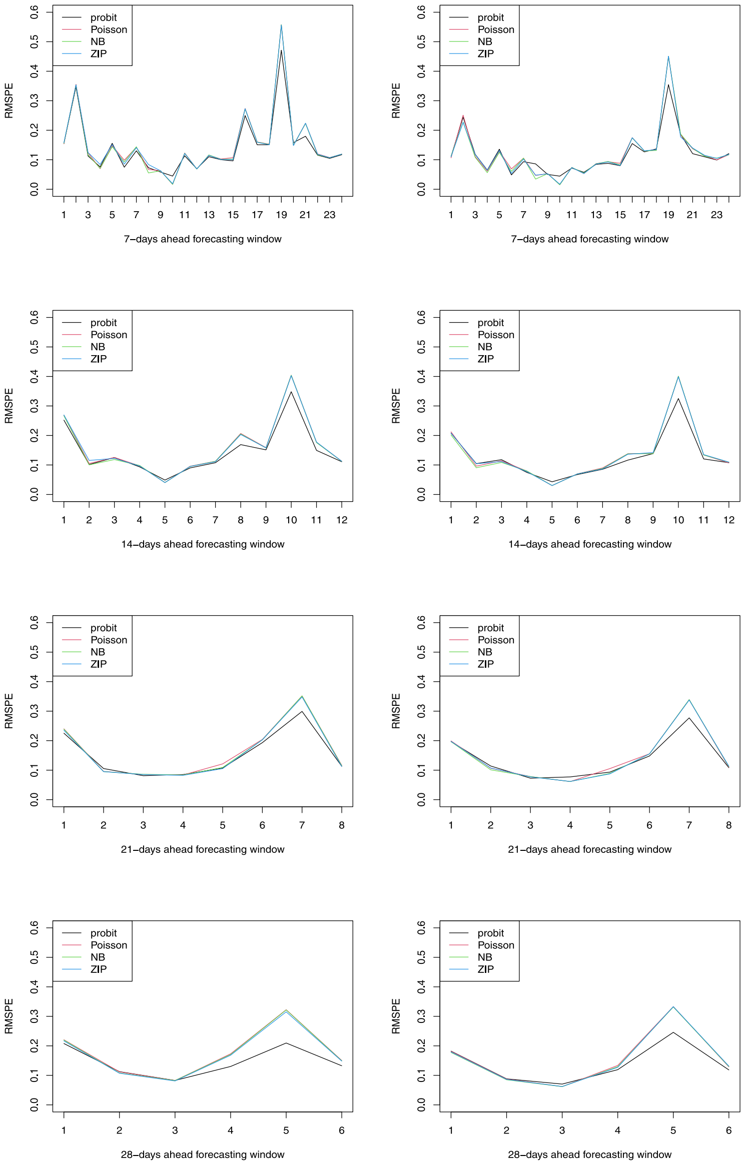


Fig. 10. RMSPE^{LOLO} (left panels) and RMSPE^{F-LOLO} right panels) for comparing the predictive accuracy between the probit and count regression models for 7 days, 14 days, 21 days and 28 days ahead prediction. Note that in most cases, the RMSPEs based on the count regression models differ only minimally, hence they can hardly be distinguished.

difficult to diagnose and rectify. The computation can also be intensive or even prohibitive in fitting the model to large datasets. One way to remedy the computational challenge of implementing the Markov Chain Monte Carlo method is to use the Integrated Nested Laplace Approximation (INLA) framework (Rue et al., 2009). INLA uses analytic Laplace approximation and efficient numerical integration schemes to achieve a highly accurate analytical approximation of posterior quantities of interest with relatively short computing times. However, it could be challenging to know when such approximations are adequate. For these reasons, the spatial–temporal generalized additive model by specifying space–time smooth as tensor product smoother offers computational advantage and also significant flexibility for modeling space–time interaction.

This study also demonstrates that the individual-level analysis provided a better fit for the data than the analysis based on the aggregated data. The inferior performance of the latter one could be due to the inappropriate distributional assumption of the daily mortality count variable. Generalized additive models for location, scale, and shape (GAMLSS) (Rigby and Stasinopoulos, 2005; Stasinopoulos and Rigby, 2007) are an extension to GAMs that allows all parameters of a certain response distribution to be modeled separately. GAMLSS provides a general framework for fitting regression type models where the distribution of the response variable does not have to belong to the exponential family and includes highly skewed and kurtotic continuous and discrete distribution (Rigby and Stasinopoulos, 2005). Comparing analyses based on individual versus aggregated data based on GAMLSS is interesting to explore as a new project. Further, this study considered three commonly used link functions for binomial regression. Both logit and probit links are symmetric links, which assume that the probability of a given binomial response approaches 0 at the same rate as it approaches 1. The cloglog link function is asymmetric, with the probability of a given binomial response increasing slowly from 0 but then tapering off more quickly as it approaches 1. The results of the model validation and residual diagnosis indicated model with the probit link function provided an adequate fit to the data considered in this study. However, in some other applications, these popular links may not always provide the best fit for a given dataset. Other flexible link function for binomial regression can be chosen, such as a two-parameter class of generalized logistic models (Stukel, 1988; Ghosh and Alzaatreh, 2018), skewed t-link function (Kim et al., 2007), skewed probit link function (Chen et al., 1999). However, these models may not be straightforward to be implemented. Future studies to develop spatial–temporal GAM or GAMLSS with more flexible link functions is therefore warranted.

In addition, previous research showed the overall coverage of the GAM model is reasonably close to the normal level, and the component-wise intervals can only be used as a rough guide to the uncertainty of the components, since the intervals are conditional on the smoothing parameters (Wood, 2017; Marra and Wood, 2012). One way to improve the performance of the intervals is to account for the smoothing parameter uncertainty. However, obtaining the distribution of the smoothing parameter is computationally intensive, which requires MCMC simulations or using parameter bootstrap approximation of its sampling distribution (Wood, 2017). Despite the underperformance of component-wise intervals, the performance of all confidence intervals in a GAM setting improves as sample size increases (Wood, 2017; Marra and Wood, 2012). Further, the theoretical argument suggests that component-wise confidence intervals could achieve nominal coverage probability provided that heavy smoothing and collinearity among predictors are avoided. As a result, the empirical coverages of the confidence intervals are expected to be close to the nominal level in this study, since the sample size is fairly large and no heavy smoothing and collinearity among the predictors were identified. However, more rigorous investigations in the context of spatial–temporal GAM through extensive simulation studies would be needed.

COVID mortality data may also exhibit a more complex spatial–temporal dependence structure or even discontinuities due to lockdown or distancing policies. Such complexity may not be adequately reflected using tensor product smoothing spline, which may smooth out the abrupt changes. Locally adaptive spatial models proposed in a Bayesian framework (Lee and Mitchell, 2013) could be useful for capturing the spatial discontinuities. Nevertheless, the main constraint of applying the Bayesian model in disease mapping (Lawson, 2008; Knorr-Held, 2000; Lee, 2011) is the computational complexity for large-scale data, which may lead to infeasible computational times. A computationally efficient algorithm for modeling the discontinuity of spatially and temporally correlated infectious disease surveillance data could be developed.

Several limitations of this study need to be acknowledged. First, no information on the individual-level socio-economic status, human mobility pattern, and social restrictions at a local level are available, which may contribute to explaining the spatial–temporal pattern identified in the study. Second, survival models to model the time from the date of infection or symptom onset until death or censoring would be a more appropriate approach than logistic regression for modeling mortality risk. However, the time variable provided in the dataset used in this study is a derived/combined variable that best estimates when the disease was acquired and refers to the earliest available date from symptom onset (the first day that COVID-19 symptoms occurred), laboratory specimen collection date, or reported date. Another variable is available in the dataset is called reported date, which is the date on which the case was reported to Toronto Public Health. However, 4513 out of 49,216 (9.2%) of the subjects have episode dates equal to or greater than the reported dates, reflecting the imperfect and incomplete information on the time variables. In addition, death occurrences are subject to reporting delay, and the analysis of deaths by reported date is therefore inevitably distorted by such delay. The lack of accurate information on the dates of symptom onset and mortality thus limits the capacity of carrying out survival analysis in this study.

Acknowledgments

This research was supported by the discovery grant from the Natural Sciences and Engineering Research Council of Canada. The author is also grateful to the Editor and two anonymous referees for their very valuable and constructive comments, which greatly helped to improve the quality of this paper.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Selected Papers of Hirotugu Akaike. Springer New York, New York, NY, pp. 199–213.

Augustin, N.H., Musio, M., von Wilpert, K., Kublin, E., Wood, S.N., Schumacher, M., 2009. Modeling spatiotemporal forest health monitoring data. J. Amer. Statist. Assoc. 104 (487), 899–911.

Bivand, R., Wong, D.W., 2018. Comparing implementations of global and local indicators of spatial association. TEST 27 (3), 716–748.

Box, G.E.P., Jenkins, G., 1976. Time Series Analysis: Forecasting and Control, Holden-Day.

Chaudhry, R., Dranitsaris, G., Mubashir, T., Bartoszko, J., Riazi, S., 2020. A country level analysis measuring the impact of government actions, country preparedness and socioeconomic factors on COVID-19 mortality and related health outcomes. EClinicalMedicine 25, 100464.

Chen, M.-H., Dey, D.K., Shao, Q.-M., 1999. A new skewed link model for dichotomous quantal response data. J. Amer. Statist. Assoc. 94 (448), 1172–1186.

Christensen, O.F., Roberts, G.O., Sköld, M., 2006. Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. J. Comput. Graph. Statist. 15 (1), 1–17.

Cliff, A.D., Ord, J.K., 1981. Spatial Processes: Models and Applications. Pion, London.

Cordes, J., Castro, M.C., 2020. Spatial analysis of COVID-19 clusters and contextual factors in new york city. Spat. Spatio-Temporal Epidemiol. 34, 100355.

Craven, P., Wahba, G., 1979. Smoothing noisy data with spline functions. Numer. Math. 31, 377–403.

Cressie, N., 1993. Statistics for Spatial Data, revised ed. Wiley-Interscience, New York.

Czado, C., Santner, T.J., 1992. The effect of link misspecification on binary regression inference. J. Statist. Plann. Inference 33 (2), 213–231.

Das, A.K., Mishra, S., Saraswathy Gopalan, S., 2020. Predicting CoVid-19 community mortality risk using machine learning and development of an online prognostic tool. PeerJ 8.

Du, R.-H., Liang, L.-R., Yang, C.-Q., Wang, W., Cao, T.-Z., Li, M., Guo, G.-Y., Du, J., Zheng, C.-L., Zhu, Q., Hu, M., Li, X.-Y., Peng, P., Shi, H.-Z., 2020. Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: A prospective cohort study. Eur. Respir. J..

Dunn, P.K., Smyth, G.K., 1996. Randomized quantile residuals. J. Comput. Graph. Statist. 5 (3), 236–244.

Feng, C., Li, L., Sadeghpour, A., 2020. A comparison of residual diagnosis tools for diagnosing regression models for count data. BMC Med. Res. Methodol. 20 (175), 1–21.

Gelman, A., Goegebeur, Y., Tuerlinckx, F., Van Mechelen, I., 2000. Diagnostic checks for discrete data regression models using posterior predictive simulations. J. R. Stat. Soc. Ser. C. Appl. Stat. 49 (2), 247–268.

Ghosh, I., Alzaatreh, A., 2018. A new class of generalized logistic distribution. Comm. Statist. Theory Methods 47 (9), 2043–2055.

Greenland, S., 2001. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. Int. J. Epidemiol. 30 (6), 1343–1350.

Harbarth, S., Harris, A.D., Carmeli, Y., Samore, M.H., 2001. Parallel analysis of individual and aggregated data on antibiotic exposure and resistance in gram-negative bacilli. Clin. Infect. Dis. 33 (9), 1462–1468.

Harrell, F.E., 2015. Regression Modeling Strategies: With Applications To Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall.

Karmakar, M., Lantz, P.M., Tipirneni, R., 2021. Association of social and demographic factors with COVID-19 incidence and death rates in the US. JAMA Netw. Open 4 (1), e2036462.

Kim, S., Chen, M.-H., Dey, D.K., 2007. Flexible generalized t-link models for binary response data. Biometrika 95 (1), 93–106.

Knorr-Held, L., 2000. Bayesian Modelling of inseparable space-time variation in disease risk. Stat. Med. 19 (17–18), 2555–2567.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.

Lawson, A., 2008. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. Chapman and Hall, New York.

Lee, D., 2011. A comparison of conditional autoregressive models used in Bayesian disease mapping. Spat. Spatio-Temporal Epidemiol. 2 (2), 79–89.

Lee, D., Mitchell, R., 2013. Locally adaptive spatial smoothing using conditional auto-regressive models. J. R. Statist. Soc. Ser. C (Appl. Statist.) 62 (4), 593–608.

Lima, E.E.C.d., Gayawan, E., Baptista, E.A., Queiroz, B.L., 2021. Spatial pattern of COVID-19 deaths and infections in small areas of Brazil. PLoS One 16 (2), 1–12.

de Lusignan, S., Dorward, J., Correa, A., Jones, N., Akinyemi, O., Amirthalingam, G., Andrews, N., Byford, R., Dabrera, G., Elliot, A., Ellis, J., Ferreira, F., Lopez Bernal, J., Okusi, C., Ramsay, M., Sherlock, J., Smith, G., Williams, J., Howsam, G., Zambon, M., Joy, M., Hobbs, F.D.R., 2020. Risk factors for SARS-CoV-2 among patients in the oxford royal college of general practitioners research and surveillance centre primary care network: a cross-sectional study. Lancet Infect. Dis. 20, 1034–1042.

Marra, G., Wood, S.N., 2011. Practical variable selection for generalized additive models. Comput. Statist. Data Anal. 55 (7), 2372–2387.

Marra, G., Wood, S.N., 2012. Coverage properties of confidence intervals for generalized additive model components. Scand. J. Stat. 39 (1), 53–74.

Meyerowitz-Katz, G., Merone, L., 2020. A systematic review and meta-analysis of published research data on COVID-19 infection fatality rates. Int. J. Infect. Dis. 101, 138–148.

Mollalo, A., Vahedi, B., Rivera, K.M., 2020. GIS-Based spatial modeling of COVID-19 incidence rate in the continental United States. Sci. Total Environ. 728, 138884.

Moran, P.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37 (1/2), 17–23.

Okeahalam, C., Williams, V., Otwombe, K., 2020. Factors associated with COVID-19 infections and mortality in africa: a cross-sectional study using publicly available data. BMJ Open 10 (11).

Paciorek, C.J., 2007. Computational techniques for spatial logistic regression with large data sets. Comput. Statist. Data Anal. 51, 3631–3653.

Real, L.A., Biek, R., 2007. Spatial dynamics and genetics of infectious diseases on heterogeneous landscapes. J. R. Soc. Interface 4, 935–948.

Rigby, R.A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape. J. R. Stat. Soc. Ser. C. Appl. Stat. 54 (3), 507–554.

Rosenthal, N., Cao, Z., Gundrum, J., Sianis, J., Safo, S., 2020. Risk factors associated with in-hospital mortality in a US national sample of patients with COVID-19. JAMA Netw. Open 3 (12), e2029058.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. Chapman and Hall, Boca Raton.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2), 319–392.

Rufibach, K., 2010. Use of brier score to assess binary predictions. J. Clin. Epidemiol. 63 (8), 938–939.

Ruppert, D., 2002. Selecting the number of knots for penalized splines. J. Comput. Graph. Statist. 11, 735–757.

Stasinopoulos, D.M., Rigby, R.A., 2007. Generalized additive models for location scale and shape (GAMLSS) in r. J. Statist. Softw. 23 (7), 1–46, Articles.

Stukel, T.A., 1988. Generalized logistic models. J. Amer. Statist. Assoc. 83 (402), 426–431.

Talukder, A., 2020. Effect of age on death due to coronavirus disease 2019 (COVID-19): Application of Poisson regression model. Int. J. Clin. Pract. 74 (11), e13649.

Ujiie, M., Tsuzuki, S., Ohmagari, N., 2020. Effect of temperature on the infectivity of COVID-19. Int. J. Infect. Dis. 95, 301–303.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Springer-Verlag.

Wadhera, R.K., Wadhera, P., Gaba, P., Figueroa, J.F., Joynt Maddox, K.E., Yeh, R.W., Shen, C.Y., 2020. Variation in COVID-19 hospitalizations and deaths across new york city boroughs. JAMA 323(21), 2192–2195.

Wahba, G., 1985. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. Ann. Statist. 13, 1378–1402.

Waller, L.A., Carlin, B.P., Xia, H., Gelfand, A.E., 1997. Hierarchical spatio-temporal mapping of disease rates. J. Amer. Statist. Assoc. 92 (438), 607–617.

Williamson, E.J., Walker, A., Goldacre, B., 2020. Factors associated with COVID-19-related death using opensafely. Nature 584, 430–436.

Wood, S.N., 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. J. R. Stat. Soc. Ser. B Stat. Methodol. 62, 413–428.

Wood, S.N., 2003. Thin plate regression splines. J. R. Stat. Soc. Ser. B Stat. Methodol. 65 (1), 95–114.

Wood, S.N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Amer. Statist. Assoc. 99 (467), 673–686.

Wood, S.N., 2006. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. Biometrics 62 (4), 1025–1036.

Wood, S.N., 2008. Fast stable direct fitting and smoothness selection for generalized additive models. J. R. Stat. Soc. Ser. B Stat. Methodol. 70, 495–518.

Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (1), 3–36.

Wood, S., 2017. Generalized Additive Models: An Introduction with R. CRC Press, Boco Raton.

Xu, K., Zhou, M., Yang, D., Ling, Y., Liu, K., Bai, T., Cheng, Z., Li, J., 2020. Application of ordinal logistic regression analysis to identify the determinants of illness severity of COVID-19 in China. Epidemiol. Infect. 148, e146.

Yu, C., Lei, Q., Li, W., Wang, X., Liu, W., Fan, X., Li, W., 2020. Clinical characteristics, associated factors, and predicting COVID-19 mortality risk: A retrospective study in wuhan, China. Am. J. Prev. Med. 59(2), 168–175.

Yue, Y., Speckman, P.L., 2010. Nonstationary spatial Gaussian Markov random fields. J. Comput. Graph. Statist. 19 (1), 96–116.

Zhang, C.H., Schwartz, G.G., 2020. Spatial disparities in coronavirus incidence and mortality in the United States: An ecological analysis as of may 2020. J. Rural Health 36(3), 433–445.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y.M., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., Tu, S., Zhang, Y., Chen, H., Cao, B., 2020. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in wuhan, China: a retrospective cohort study. Lancet 395, 1054–1062.