Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Bayesian negative binomial regression with spatially varying dispersion: Modeling COVID-19 incidence in Georgia

Fedelis Mutiso a,⁎, John L. Pearce b, Sara E. Benjamin-Neelon c,d, Noel T. Mueller e,f, Hong Li a, Brian Neelon a,g

a Division of Biostatistics, Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA
b Division of Environmental Health, Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA
c Department of Health, Behavior and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
d Lerner Center for Public Health Promotion, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
e Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
f Welch Center for Prevention, Epidemiology and Clinical Research, Johns Hopkins University, Baltimore, MD, USA
g Charleston Health Equity and Rural Outreach Innovation Center (HEROIC), Ralph H. Johnson VA Health Care System, Charleston, SC, USA

A B S T R A C T

Overdispersed count data arise commonly in disease mapping and infectious disease studies. Typically, the level of overdispersion is assumed to be constant over time and space. In some applications, however, this assumption is violated, and in such cases, it is necessary to model the dispersion as a function of time and space in order to obtain valid inferences. Motivated by a study examining spatiotemporal patterns in COVID-19 incidence, we develop a Bayesian negative binomial model that accounts for heterogeneity in both the incidence rate and degree of overdispersion. To fully capture the heterogeneity in the data, we introduce region-level covariates, smooth temporal effects, and spatially correlated random effects in both the mean and dispersion components of the model. The random effects are assigned bivariate intrinsic conditionally autoregressive priors that promote spatial smoothing and permit the model components

⁎ Corresponding author.
E-mail address: fedelis@musc.edu (F. Mutiso).
to borrow information, which is appealing when the mean and dispersion are spatially correlated. Through simulation studies, we show that ignoring heterogeneity in the dispersion can lead to biased and imprecise estimates. For estimation, we adopt a Bayesian approach that combines full-conditional Gibbs sampling and Metropolis–Hastings steps. We apply the model to a study of COVID-19 incidence in the state of Georgia, USA from March 15 to December 31, 2020.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

In disease mapping studies, the Poisson distribution is commonly used to examine geographic variation in disease counts (Best et al., 2000; Wakefield, 2006). A well-known limitation of the Poisson distribution, however, is equidispersion, or equivalence of the mean and variance. In many applications, this assumption is violated and the data exhibit overdispersion in which the variance exceeds the mean. Less commonly, the data may be underdispersed whereby the variance is less than the mean. To address overdispersion, the Poisson mean is often allowed to vary across covariate groups or geographic units using fixed and random effects. For example, many disease mapping studies will allow the mean to vary according to a prior distribution that induces marginal overdispersion in the response. In the simplest case, the Poisson mean is assumed to follow a Gamma prior distribution, leading to a marginally overdispersed negative binomial (NB) distribution. More generally, it is possible to introduce spatial random effects into the Poisson mean to capture unobserved geographical variation in the mean and induce marginal overdispersion. Conditional on the random effects, however, the data are once again assumed to follow an equidispersed Poisson distribution.

In some cases, this conditional equidispersion assumption is itself limited, and the data are not only marginally overdispersed, but also conditionally overdispersed. Failing to account for this conditional overdispersion can lead to downwardly biased standard errors (Hilbe, 2011). One remedy is to assume a conditional NB (or related) distribution to accommodate residual overdispersion that may be present after conditioning on random effects. In particular, the NB model introduces a separate dispersion parameter that accounts for overdispersion in the data not explained by latent heterogeneity in the mean. Although this approach is sensible in many applications, most overdispersed count distributions, including NB, Conway–Maxwell–Poisson (Shmueli et al., 2005) or Poisson-lognormal (Stewart, 1994), constrain the dispersion parameter to be constant across covariate groups or, in the case of spatial data, across geographic areas. This assumption may be untenable in some applications. For example, in studies of COVID-19 incidence rates, we might expect the degree of conditional overdispersion to vary across covariate groups and across
geographic regions, such as counties or census tracts. In such cases, it is necessary to properly model not only the mean, but also the dispersion term itself as a function of covariates and potentially spatial random effects to appropriately capture the observed and unobserved heterogeneity in the data.

Several authors have explored the utility of varying dispersion parameters in Poisson-gamma models. Heydecker and Wu (2001) noted that the dispersion parameter could be modeled as a function of covariates in a study to identify sites for road accident remedial work. They postulated that the Poisson-gamma model with non-constant dispersion can better represent the nature of crash data than traditional Poisson-gamma model with constant dispersion parameter. Miaou and Lord (2003) proposed a Poisson-gamma model with a dispersion parameter dependent on traffic flows and flow ratios to study traffic crash-flow relationships for intersections. They found that assuming a fixed dispersion parameter could seriously bias both the empirical and full Bayes estimates for individual sites. Miranda-Moreno et al. (2005) compared the performance of alternative risk models for identifying hazardous sites and found that Poisson-gamma models with covariate-dependent dispersion parameters were superior to traditional models. More recently, Geedipally and Lord (2008) studied the effects of covariate-dependent dispersion parameters on estimation of confidence intervals in Poisson-gamma models. They noted that models with a non-constant dispersion parameter produced smaller confidence intervals and thus more precise estimates for both the gamma mean and the predicted response compared to models with a fixed dispersion parameter. Collectively, this work suggests that incorporating heterogeneous dispersion into count models can improve statistical inferences, particularly with respect to precision of parameter estimates. However, these studies did not incorporate spatial information, which is our focus here.

Our goal, therefore, is to extend this work to the spatial setting by modeling spatial heterogeneity not only in the mean of the NB distribution, but also the dispersion parameter. Again, this is relevant in cases where the degree of overdispersion varies across regions. To account for this heterogeneity, we introduce covariates and random effects for both the mean and dispersion components of the NB model. The random effects are chosen to have a spatial structure to induce smoothing and account for spatial dependence among neighboring geographic areas. We allow for correlation between the random effects in the mean and dispersion components, which is appealing in settings where the mean response is associated with the degree of overdispersion. For example, in our application examining daily COVID-19 incidence rates in state of Georgia in the United States (US), it is likely that counties with higher average COVID-19 case counts also have more heterogeneous count distributions due to the presence of extreme counts or outliers. It is therefore advantageous to build this dependence into the model, while reserving uncorrelated components as a special case. For inference, we adopt a Bayesian approach and develop an efficient Markov chain Monte Carlo (MCMC) algorithm that combines full-conditional Gibbs sampling and Metropolis–Hastings (MH) steps.

2. Case study

As a motivating application, we examined associations between county-level social vulnerability and daily COVID-19 incident case data for the state of Georgia from March 15 to December 31, 2020, for a total of 292 study days. The COVID-19 data were downloaded from the Johns Hopkins University (JHU) Center for Systems Science and Engineering (Johns Hopkins University, 2020). In Georgia, the first COVID-19 case was reported on March 2, 2020 (Georgia Department of Public Health, 2020) and there were relatively few cases in early March. As a result, we initiated our analysis on March 15. To examine associations between COVID-19 incidence and social vulnerability, we downloaded the Social Vulnerability Index (SVI) data from the CDC’s Agency for Toxic Substances and Disease Registry (Centers for Disease Control and Prevention, 2018). The SVI is a percentile-based measure of social vulnerability derived from four themes that explain different aspects of vulnerability: socioeconomic status; household composition; race, ethnicity and language; and housing and transportation availability. The four themes comprise a total of 15 variables obtained from the US Census Bureau’s American Community Survey (Centers for Disease Control and Prevention, 2018). These variables include the percent of county residents below poverty, percent
Fig. 1. Descriptive maps for incidence rate and overdispersion from intercept-only models for data from March 15, 2020 to December 31, 2020 for the 159 counties in Georgia. Panel (a): Predicted incidence. Panel (b): Overdispersion. Legends correspond to sample quintiles.

unemployed, percent of residents aged 65 and older, and crowding among others. Table S1 in Appendix A of the Supplementary material provides detailed information of the census variables used to construct each theme. Counties are ranked on each of the 15 items and the overall SVI rank for each county is formed by summing the individual rankings. To address potential confounding, we obtained several additional variables unrelated to the SVI from the Robert Wood Johnson Foundation’s 2019 County Health Rankings & Roadmaps: Rankings Data & Documentation (RWJ Foundation, 2020). These variables included the percentage of county residents in fair or poor health, the percentage of female residents in the county, the percentage of adult smokers in the county, and population density defined as the number of residents per square mile in the county. Data from various sources were at the county level, which allowed seamless merging using each county’s unique Federal Information Processing Standards (FIPS) code to create the final analysis dataset.

Fig. 1 presents a map of the incidence rate per 100 000 individuals (panel a) and overdispersion estimates (panel b) from separate intercept-only NB models for all 159 counties in Georgia over the course of the study period, from March 15 to December 31, 2020. The color shades correspond to the quintile distribution of the predicted incidence and estimated overdispersion. The white shade represents the lowest quintile, whereas the darker shade denotes the uppermost quintile. The predicted incidence rate ranged from 8.42 to 68.65 cases per 100 000 individuals while overdispersion ranged from 0.712 to 5.443. Fig. 1(a) suggests high incidence case rates for select counties in the southwest, southeast, and northern corners of Georgia. The majority of these counties are small and rural with high social vulnerability indices. In contrast, the unadjusted predicted incidence is lower in larger suburban counties such as those within the Atlanta area in the north central portion of the state. Fig. 1(b) suggests high overdispersion in select counties in the southwest, central and northern corners of the state. As with the incidence rates, most of the overdispersed counties are rural and highly vulnerable. The elevated overdispersion in the southwest counties may be due to presence of extreme counts emanating from COVID-19 outbreaks early in the pandemic (Galofaro, 2020). Overall, Fig. 1(b) indicates that there is substantial spatial variation in overdispersion across counties. This highlights the importance of properly modeling dispersion in infectious disease settings, where localized outbreaks can lead to spikes in overdispersion that vary over time and across geographic regions.
3. Spatial negative binomial model with heterogeneous dispersion

3.1. Model specification and Bayesian inference

The spatial NB model with heterogeneous dispersion extends the conventional NB model to accommodate covariates in the dispersion parameter, as well as spatial effects in both the mean and dispersion. Let \( y_{ij} \) denote the number of COVID-19 cases in county \( i \) \((i = 1, \ldots, n)\) on day \( j \) \((j = 1, \ldots, J)\). We assume that \( y_{ij} \) is distributed as NB(\( \alpha_{ij}, \psi_{ij} \)) with the following probability mass function

\[
p(y_{ij}|\alpha_{ij}, \psi_{ij}) = \frac{\Gamma(y_{ij} + 1/\alpha_{ij})}{\Gamma(1/\alpha_{ij})\psi_{ij}^{y_{ij}}(1 - \psi_{ij})^{1/\alpha_{ij}}}, \quad \alpha_{ij} > 0, \quad 0 < \psi_{ij} < 1
\]

(1)

The conditional mean and variance of \( y_{ij} \) are given by

\[
E(y_{ij}|\alpha_{ij}, \psi_{ij}) = \frac{\psi_{ij}}{\alpha_{ij}(1 - \psi_{ij})} = \mu_{ij}
\]

(2)

\[
\text{Var}(y_{ij}|\alpha_{ij}, \psi_{ij}) = \frac{\psi_{ij}}{\alpha_{ij}(1 - \psi_{ij})^2} = \mu_{ij}(1 + \alpha_{ij}\mu_{ij})
\]

where \( \alpha_{ij} \) captures the overdispersion in the data. As \( \alpha_{ij} \to 0 \), the mean and variance of the counts \( y_{ij} \) are approximately the same and a Poisson model could be assumed. As \( \alpha_{ij} \to \infty \), the counts of county \( i \) on day \( j \) become increasingly dispersed relative to the Poisson distribution. To facilitate Bayesian computation, we follow Pillow and Scott (2012) and model \( \psi_{ij} \) using an inverse-logit function:

\[
\psi_{ij} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}
\]

(3)

\[
\eta_{ij} = x_{ij}^T\beta_1 + f_1(t_j) + \phi_{1i} + \log P_i - \log C,
\]

where \( x_{ij} \) is a \( p \times 1 \) vector of county-level covariates, \( \beta_1 \) is a \( p \times 1 \) vector of fixed effect coefficients, \( f_1(t_j) \) is a smooth function of time evaluated at day \( t_j = j \), and \( \phi_{1i} \) is a county-level spatial random effect. The linear predictor, \( \eta_{ij} \), also includes two offset terms: \( P_i \), the population size for county \( i \), and \( C \), an additional rate constant that allows us to model the COVID-19 incident rate per \( C \) individuals (e.g. \( C=100 \ 000 \)) in a county of size \( P_i \). The above representation of NB allows us to exploit a key connection between the inverse-logit function for \( \psi_{ij} \) and the Pólya-Gamma (PG) density for efficient Bayesian inference. In particular, the parameterization in Eq. (3) enables us to introduce latent weight terms, \( \omega_{ij} \), that follow independent PG prior distributions. By conditioning on these weights, the NB GLM can be represented as a scale mixture of normal distributions with PG precision terms, which allows a straightforward and efficient Gibbs sampling update for NB regression parameters. Details on key properties of PG are provided in Appendix B of the Supplementary material. For additional details on PG data augmentation and the above parameterization of the NB model, see Neelon (2019) and Neelon et al. (2022).

To accommodate spatial and temporal variation in dispersion, we assume the following functional form for \( \alpha_{ij} \):

\[
\log \alpha_{ij} = x_{ij}^T\beta_2 + f_2(t_j) + \phi_{2i},
\]

(4)

where \( x_{ij} \) is a \( q \times 1 \) vector of county-level covariates, \( \beta_2 \) is a \( q \times 1 \) vector of fixed effect coefficients, \( f_2(t_j) \) is a smooth function of time evaluated at day \( t_j = j \), and \( \phi_{2i} \) is a county-level spatial random effect for overdispersion. In subsequent discussion, we assume that the fixed effect covariates for the mean and dispersion components are identical (i.e., \( x_{ij} = x_{ij} = x_{ij} \) and thus \( p = q \)) although in general this is not necessary. To allow for sufficient smoothness over time, we model \( f_1 \) and \( f_2 \) using cubic B-splines (Eilers and Marx, 1996) of the form

\[
f_h(t_j) = w_j^T \gamma_h = \sum_{k=1}^{\kappa} w_{jk} \gamma_{hk},
\]

(5)
where \( h = 1, 2 \) is the model component (1 for the mean and 2 for the dispersion component); \( \mathbf{w}_j \) is a \( K \times 1 \) vector of B-splines basis functions for time \( j \), where \( K = v + 3 \) and \( v \) is the number of interior knots; \( \mathbf{y}_h = (\mathbf{y}_{h1}, \ldots, \mathbf{y}_{hk})^T \) is a \( K \times 1 \) vector of fixed effect basis coefficients for component \( h \) that capture the overall temporal trend.

The non-intercept terms of \( \mathbf{\beta}_1 \) in Eq. (3), that is, \( \beta_{12}, \ldots, \beta_{1p} \), represent log rate ratios (RRs). Thus, for \( l = 2, \ldots, p \), \( \exp(\beta_{1l}) \) denotes a rate ratio. The parameter vector \( \mathbf{\beta}_2 \) in Eq. (4) denotes a vector of log overdispersion terms. For example, \( \beta_{2l} \) (\( l = 2, \ldots, p \)) represents the change in \( \log \alpha_{ij} \) for a one-unit change in \( \mathbf{x}_{2ij} \), holding other variables fixed. The random effects \( \phi_{1i} \) and \( \phi_{2i} \) capture, respectively, the variability in log rate ratio and log overdispersion across regions. Adjusted for observed covariates, large values of \( \phi_{1i} \) suggest a higher incidence of the outcome in region \( i \) compared to other regions while larger values of \( \phi_{2i} \) imply more overdispersion in counts for region \( i \) relative to other regions.

The spatial random effects account for unmeasured regional level factors that may influence the incidence and dispersion in related ways. For instance, regions with a high incidence may exhibit more overdispersion, possibly due to presence of extreme counts. To allow this association and to encourage spatial smoothing and borrowing of information across neighboring areas, we assume a bivariate intrinsic CAR (BICAR) prior distribution for \( \phi_1 = (\phi_{11}, \phi_{21})^T \) (Mardia, 1988):

\[
\phi_i | \phi_{(-i)}, \Sigma \sim N_2 \left( \frac{1}{m_i} \sum_{l \in \hat{a}_i} \phi_l, \frac{1}{m_i} \Sigma \right),
\]

where \( m_i \) represents the number of neighbors of region \( i \), \( \hat{a}_i \) is the set of neighbors for region \( i \) and \( \Sigma \) is a \( 2 \times 2 \) variance–covariance matrix of \( \phi_i \) conditional on the remaining spatial random effects, \( \phi_{(-i)} \). In presence of a fixed effect intercept, a sum-to-zero constraint must be applied to \( \{\phi_1, \ldots, \phi_n\} \) to ensure the model is identifiable.

The BICAR prior implies that the conditional mean of \( \phi_i \) is the average of neighboring regions' spatial effects, and thus incorporates information from the neighbors by allowing adjacent regions to borrow information. The sharing of information can lead to more reliable random-effect estimates for regions with small sample sizes. The covariance matrix \( \Sigma \) scaled by the number of neighbors for region \( i \) allows for spatial smoothing since, as the number of neighbors \( m_i \) increases, more information is borrowed in predicting \( \phi_i \), and hence more prior confidence that \( \phi_i \) is conditionally similar to the average of adjacent regions. The off-diagonal element of \( \Sigma \), denoted \( \Sigma_{12} \), is the covariance between \( \phi_{1i} \) and \( \phi_{2i} \) and controls the degree of association between the two model components. When \( \Sigma_{12} > 0 \), regions with higher incidence may tend to have more overdispersed counts. When \( \Sigma_{12} = 0 \), the two models are uncorrelated; thus, we retain separate components as a special case, and this assumption can be evaluated by inspecting whether the credible interval for \( \Sigma_{12} \) includes 0.

The joint prior distribution for the \( 2n \times 1 \) spatial random effects matrix \( \Phi = (\phi_1^T, \ldots, \phi_n^T)^T \) follows from Brook's lemma (Banerjee et al., 2014) and is proportional to the multivariate normal distribution given by

\[
\Phi | \Sigma \propto \exp \left\{ -\frac{1}{2} \Phi^T \left[ (M - A) \otimes \Sigma^{-1} \right] \Phi \right\},
\]

where \( M = \text{diag}(m_1, m_2, \ldots, m_n) \) and \( A \) is \( n \times n \) adjacency matrix with \( a_{ii} = 0, a_{ij} = 1 \) if regions \( i \) and \( j \) are neighbors and \( a_{ij} = 0 \) otherwise. Since \( M - A \) is singular, the joint prior distribution in (7) is improper, but the posterior of \( \Phi \) is itself proper.

To complete Bayesian model specification, we assign proper but diffuse priors to the respective model parameters. For the mean component, we conveniently combine the fixed effect and spline parameters, \( \mathbf{\beta}_1 \) and \( \mathbf{y}_1 \), into a single parameter vector, say \( \mathbf{\theta}_1 = (\mathbf{\beta}_1^T, \mathbf{y}_1^T)^T \) of dimension \((p + K)\times 1\), for a single Gibbs update from the full conditional distribution of \( \mathbf{\theta}_1 \). We assume an exchangeable normal prior for \( \mathbf{\theta}_1 \). That is, \( \mathbf{\theta}_1 \sim N_{p+K}(\mathbf{0}, \mathbf{\Sigma}_0) \), where \( \mathbf{\Sigma}_0 = 10I_{p+K} \) is the default choice leading to independent priors for the fixed effects and spline coefficients. To update the parameters for dispersion component, \( \mathbf{\beta}_2 \) and \( \mathbf{y}_2 \), we take advantage of the correlation between these two parameter vectors to efficiently update them as a block, say \( \mathbf{\theta}_2 = (\mathbf{\beta}_2^T, \mathbf{y}_2^T)^T \) of dimension \((p + K)\times 1\),
using a Metropolis–Hastings step. Similar to \( \theta_1 \), we assume an exchangeable normal prior for \( \theta_2 \). That is, \( \theta_2 \sim N_{p+K}(\mathbf{0}, \Sigma_0) \). For the spatial covariance matrix, \( \Sigma \), we assume an inverse-Wishart \((\nu, \Omega)\) prior with default \( \nu = 3 \) degrees of freedom and \( \Omega \) as the default choice for \( \Sigma_0 \).

For posterior computation, we implement an MCMC algorithm that combines full-conditioned Gibbs steps with Metropolis–Hastings updates. For \( \theta_1 \), we adopt the computationally efficient data-augmented Gibbs sampler proposed by Pillow and Scott (2012) for posterior computation of \( \theta_1 \). The sampler introduces latent weights, \( \omega_{ij} \) for all \( i \) and \( j \), that are scale mixtures of normal with independent PG precision terms (Polson et al., 2013a). For \( b > 0 \) and \( c \in \mathbb{R} \), \( \omega \) is said to have a PG distribution if

\[
\omega \sim \text{PG}(b, c) = \frac{1}{2\pi} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)},
\]

where the \( g_k \)'s are independently distributed according to a Gamma\((b, 1)\) distribution. If we draw \( \omega_{ij} \) independently from \( \text{PG}(y_{ij} + 1/\alpha_{ij}, \eta_{ij}) \), where \( \eta_{ij} \) is defined in (3), then conditional on \( \omega_{ij} \), the latent response variable \( z_{ij} = \frac{y_{ij} - 1/\alpha_{ij}}{\omega_{ij}} \) is distributed according to \( N(\mathbf{u}_i^T \theta_1 + \phi_{ij} + \log P_i - \log C_i, \omega_{ij}^{-1}) \) where \( \mathbf{u}_i \) is the combined \((p + K) \times 1\) vector of fixed-effect covariates and the \( B \)-splines basis functions. This leads to a convenient data-augmentation scheme for updating \( \theta_1 \): (1) for all \( i, j \), draw \( \omega_{ij} \) from \( \text{PG}(y_{ij} + 1/\alpha_{ij}, \eta_{ij}) \); (2) form \( z_{ij} \); (3) assuming a \( N_{p+K}(\theta_0, \Sigma_0) \) prior, update \( \theta_1 \) from its closed-form \( N_{p+K}(\mu, \mathbf{V}) \) full conditional, where

\[
\mathbf{V} = (\Sigma_0^{-1} + \mathbf{U}^T \Omega \mathbf{U})^{-1},
\]

\[
\mu = \mathbf{V} [\Sigma_0^{-1} \theta_0 + \mathbf{U}^T \Omega (\mathbf{z} - \phi_1 - \log \mathbf{P} + \log C)]
\]

Here, \( \mathbf{U} \) is an \( n \times (p + K) \) design matrix of fixed-effect covariates and the \( B \)-splines basis functions, \( \Omega \) is an \( n \times n \) diagonal matrix with diagonal entries \( \{\omega_{ij}\} \), \( \mathbf{z} = (z_{11}, \ldots, z_{1J}, \ldots, z_{n1}, \ldots, z_{nJ}) \) is an \( n \times 1 \) vector of latent responses, \( \phi_1 = (\phi_{11}, \ldots, \phi_{n1}, \ldots, \phi_{nJ}) \) is an \( n \times 1 \) vector of spatial random effects, and \( \mathbf{P} = (P_{11}, \ldots, P_{1J}, \ldots, P_{n1}, \ldots, P_{nJ}) \) is an \( n \times 1 \) vector of population offsets. The proof of (8) is provided in Appendix B of the Supplementary material. To update the fixed and spline effects for the dispersion, \( \theta_2 \), we adopt a random walk MH steps with symmetric multivariate \( t \) proposal density centered at the previous value of \( \theta_2 \) and acceptance ratio

\[
\rho_{\theta_2} = \frac{p(\theta_{2}^{s} | \mathbf{y}, \theta_1, \phi_i)}{p(\theta_{2}^{s} | \mathbf{y}, \theta_1, \phi_i)} = \frac{\prod_{i=1}^{n} \prod_{j=1}^{J} \text{NB}(y_{ij} | \theta_{2}^{s}, \phi_{ij}, \phi_{ij}) \times N_{p+K}(\theta_{2}^{s}; \mathbf{0}, \Sigma_0)}{N_{p+K}(\theta_{2}; \mathbf{0}, \Sigma_0)}.
\]

where \( \theta_2^s \) and \( \theta_2^{s} \) are the proposed and current values of \( \theta_2 \) at iteration \( s \), respectively, and \( \text{NB}(y_{ij} | \theta_{2}^{s}, \phi_{ij}, \phi_{ij}) \) and \( N_{p+K}(\theta_{2}^{s}; \mathbf{0}, \Sigma_0) \) are the probability distribution functions for NB and the \( p + K \)–variate normal prior distribution with mean \( \mathbf{0} \) and covariance \( 100I_{p+K} \) evaluated at \( \theta_2^s \).

Similarly, to update the \( 2 \times 1 \) random effect vector \( \phi_i \) \((i = 1, \ldots, n)\), we use a random walk MH step with a symmetric bivariate \( t \) proposal density centered at the previous \( \phi_i \) and acceptance ratio

\[
\rho_{\phi_1} = \frac{p(\phi_{1}^{s} | \mathbf{y}, \theta_1, \phi_i)}{p(\phi_{1}^{s} | \mathbf{y}, \theta_1, \phi_i)} = \frac{\prod_{i=1}^{n} \prod_{j=1}^{J} \text{NB}(y_{ij} | \phi_{1}^{s}, \phi_{ij}, \phi_{ij}) \times \pi(\phi_{1}^{s})}{\pi(\phi_{1})},
\]

where \( \phi_{1}^{s} \) and \( \phi_{1}^{s} \) are the proposed and current values of \( \phi_1 \), respectively and \( \pi(\phi_{1}) \) is distributed according to the conditional bivariate CAR prior distribution of \( \phi_1 | \phi_{(-1)} \) given in (6). Lastly, assuming an \( \text{IW}(v_0, S_0) \) prior, we update the random effects covariance from a conjugate inverse-Wishart (IW) distribution

\[
\Sigma | \Phi \sim \text{IW}(v_0 + n - 1, S_0 + S_{\Phi}),
\]

where \( S_{\Phi} = \Phi^T (M - A) \Phi^* \) and \( \Phi^* \) is the \( n \times 2 \) random effects matrix centered at its mean. The first column of \( \Phi^* \) is equal to \( \phi_{11} = (\phi_{11}, \ldots, \phi_{1n})^T \) and the second column equal to \( \phi_{21} = (\phi_{21}, \ldots, \phi_{2n})^T \).
The Gibbs sampler for posterior inference can be summarized as follows:

1. For county $i$ on day $j$, draw $\omega_{ij}$ from $\text{PG}(y_{ij} + 1/\alpha_{ij}, \eta_{ij})$ using the accept–reject algorithm implemented using the R package BayesLogit (Polson et al., 2013b)
2. For all $i$ and $j$, create the latent response variable $z_{ij} = y_{ij} - 1/\alpha_{ij}^2$ $\omega_{ij}$
3. Update $\theta_1$ from its conjugate $N(\mu, V)$ full conditional, where $\mu$ and $V$ are provided in (8)
4. Update $\theta_2$ as a block using a random walk MH step with acceptance ratio given in (9)
5. For $i = 1, \ldots, n$, update the vector $\phi_i = (\phi_{1i}, \phi_{2i})^T$ using a MH step with acceptance ratio given in (10)
6. Finally, update the random effects covariance matrix, $\Sigma$, from a conjugate $\text{IW}$ distribution given in (11).

As a sensitivity analysis, we consider an alternative first-order random walk (RW1) prior for the spline coefficients $\gamma_1$ and $\gamma_2$ of the form

$$
\pi(\gamma_h | \tau_{\gamma_h}) \propto \exp \left( -\frac{\tau_{\gamma_h}}{2} \gamma_h^T K \gamma_h \right), \ h = 1, 2
$$

where $\tau_{\gamma_h}$ is a prior precision term and $K$ is a penalty matrix of rank $K - 1$ with entries

$$
k_{ij} = \begin{cases} 
1, & i = j = 1 \text{ or } K; \\
2, & 2 \leq i = j \leq K - 1; \\
-1, & |i - j| = 1; \\
0, & |i - j| > 1.
\end{cases}
$$

This prior introduces autoregressive dependence and creates additional smoothing which may be desirable in some cases particularly for the mean portion of the model. The Gibbs and MH update steps under this alternative RW1 prior for $\gamma_h (h = 1, 2)$ are outlined in Appendix C of the Supplementary material. As described below in Section 5, an additional sensitivity analysis was performed using this prior with COVID-19 case data from USAFacts (USAFacts, 2022).

MCMC convergence is monitored using standard Bayesian diagnostic approaches such as trace plots and Geweke (1992) statistics, which we implement using the coda package (Plummer et al., 2006). To compare different models, we use the “widely applicable information criterion” or WAIC (Watanabe, 2010) Bayesian measure, which combines a measure of predictive accuracy of the fitted model with a penalty for model complexity.

### 3.2. Parameter interpretation for a simple model

In this section, we illustrate parameter interpretation for a simple model. Consider now the following fixed effects version of the model in Eq. (1), where

$$
\psi_{ij} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}
$$

$$
\eta_{ij} = \beta_{10} + \beta_{11} x_{ij} + \log P_i - \log C
$$

$$
\log \alpha_{ij} = \beta_{20} + \beta_{21} x_{ij}.
$$

Under this specification, the model parameters have the following interpretations:

1. $\exp(\beta_{10})$ represents the incidence rate (i.e., cases per $C$ individuals) when $x_{ij} = 0$ and $\alpha_{ij} = 1$.
2. $\exp(\beta_{11})$ denotes the incidence rate ratio corresponding to a one-unit change in $x_{ij}$, holding the dispersion parameter fixed.
3. If both $\beta_{20} \ll 0$ and $\beta_{21} \ll 0$, then $\exp(\beta_{20}) \approx \exp(\beta_{21}) \approx 0$ and hence $\alpha_{ij} \approx 0$. In this case, we have essentially no overdispersion and the model reduces to a Poisson-like equidispersed model in which the variance in Eq. (2) reduces to the mean $\mu_{ij}$. Thus, the Poisson model arises as a special case of the proposed model.
4. If $\exp(\beta_{20}) \neq 0$ but $\exp(\beta_{21}) \approx 0$, then we have a conventional NB model with constant $\alpha = \exp(\beta_{20})$. Hence, the NB model arises as a special case of the proposed model.
5. Finally, if $\exp(\beta_{20}) \neq 0$ and $\exp(\beta_{21}) \neq 0$, then the degree of overdispersion varies as a function of $x_{ij}$. For example, if $x_{ij}$ is a binary indicator for the presence of a hospital facility in the county, this would imply different levels of overdispersion across counties based on presence or absence of a hospital facility.

More generally, the model proposed in Eqs. (1) and (3) permits the dispersion to vary flexibly over space and time through the inclusion of spatiotemporal fixed and random effects. At the same time, it includes the Poisson and NB as special cases, and this assumption can be assessed using model comparison metrics such as WAIC (Watanabe, 2010). Thus, the model accommodates flexible dispersion structures but can reduce to more basic models if the data support a lower degree of overdispersion.

4. Simulation study

We conducted a simulation study to examine the performance of the model and to explore the effect of ignoring spatial heterogeneity during model fitting. Specifically, we compared estimates under the proposed model to estimates from three misspecified models: (1) a model that included fixed and uncorrelated effects rather than BICAR spatial effects for the two components; (2) a model that included fixed and spatial effects for the mean component but only fixed effects for overdispersion parameter; and (3) a model with fixed and spatial effects for the mean and a constant overdispersion parameter.

To mimic the spatial structure of our application, we used the US Census county-level adjacency matrix for Georgia. This matrix has $n = 159$ counties and 431 unique adjacencies. We generated the outcome data, $y_{ij}$, under the following model comparable to one given in (1):

$$y_{ij} \sim \text{NB}(\alpha_{ij}, \psi_{ij})$$

$$p(y_{ij}|\alpha_{ij}, \psi_{ij}) = \frac{\Gamma(y_{ij} + 1/\alpha_{ij})(1 - \psi_{ij})^{1/\alpha_{ij}}\psi_{ij}^{y_{ij}}}{\Gamma(1/\alpha_{ij})y_{ij}!}$$

$$\psi_{ij} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}$$

$$\eta_{ij} = \beta_{10} + \beta_{11}x_{1ij} + \beta_{12}x_{2ij} + f_1(t_j) + \phi_{1i} + \log P_i - \log C$$

$$\log \alpha_{ij} = \beta_{20} + \beta_{21}x_{1ij} + \beta_{22}x_{2ij} + f_2(t_j) + \phi_{2i}$$

$$\phi_{i|\phi_{[-i]}, \Sigma} \sim N_2\left(\frac{1}{m_i} \sum_{t \in d_i} \phi_{t}, \frac{1}{m_i} \Sigma\right),$$

where for $i = 1, \ldots, 159$ and $j = 1, \ldots, 292$, $y_{ij}$ denotes incident COVID-19 case count for county $i$ on day $j$; $X_{1ij} \sim \text{N}(0, 1)$ is a county level covariate such as the percent of county designated as rural; $X_{2ij} \sim \text{Bern}(0.5)$ is a county level binary covariate; $f_1(t_j)$ and $f_2(t_j)$ are smooth functions of time, $t_j$, for the mean and dispersion components; and $\phi_i = (\phi_{1i}, \phi_{2i})^T$ is a vector of spatial intercepts that is assumed to follow a BICAR prior distribution. For component $h = 1, 2$, we approximate $f_h(t_j)$ using cubic B-splines as described in Eq. (5) with 20 interior knots placed 14 days apart for a total $20 + 3 = 23$ spline coefficients. For the county sample sizes, $P_i$, we used the 2021 US Census county population estimates for the state of Georgia (US Census Bureau, 2021), resulting in a range of 1 537 to 1 063 937. We assumed a rate constant $C = 10^3$ in order to model incidence rates per 100 000 individuals. The true values for the model parameters (excluding splines) are given in Table 1. Figure S1 in Appendix D of the Supplementary material presents a partial histogram of the counts. We assumed independent $\text{N}(0, 100)$ priors for the fixed effects for both the count and dispersion components of the model. To complete prior specification, we assigned an IW $(3, I_2)$ prior to $\Sigma$. We set the initial values for the fixed effects of the mean, $\theta_1$, at the maximum likelihood estimates from a fixed effect NB model while the fixed effects for dispersion component, $\theta_2$, were initialized to 0. We sampled the initial values for $\phi_{1i}$ and $\phi_{2i}$ from independent standard normal distributions, and $\Sigma$ was initialized to $I_2$. 


Table 1
Results from simulation study comparing a spatially correlated NB model with three misspecified models: Model 1: Spatially correlated random intercepts in the mean and dispersion components. Model 2: Separate spatial random intercepts in the mean and dispersion components. Model 3: Spatial random intercept in mean component and only covariates in the dispersion component. Model 4: Spatial random intercept in mean component and a constant dispersion parameter. Note that $\Sigma_{11} = 1/\tau_1$ in models 2, 3, and 4 while $\Sigma_{22} = 1/\tau_2$ in model 2.

| Model | Component | Parameter | Truth | Post. Mean (95% CrI) | Bias |
|-------|-----------|-----------|-------|----------------------|------|
| Model 1 (Simulated Model) | Count | $\beta_{10}$ | 0.50 | 0.42 (0.28, 0.57) | -0.08 |
| | | $\beta_{11}$ | -0.25 | -0.25 (-0.32, -0.18) | 0.003 |
| | | $\beta_{12}$ | 0.50 | 0.41 (0.28, 0.53) | -0.09 |
| | Dispersion | $\beta_{20}$ | -0.50 | -0.54 (-0.59, -0.47) | -0.04 |
| | | $\beta_{21}$ | 0.50 | 0.50 (0.46, 0.57) | 0.001 |
| | Random effects | $\Sigma_{11}$ | 0.70 | 0.61 (0.44, 0.85) | -0.09 |
| | | $\Sigma_{12}$ | 0.20 | 0.16 (0.05, 0.30) | -0.04 |
| | | $\Sigma_{22}$ | 0.50 | 0.45 (0.31, 0.63) | -0.05 |
| Model 2 | Count | $\beta_{10}$ | 0.50 | 0.42 (0.25, 0.59) | -0.08 |
| | | $\beta_{11}$ | -0.25 | -0.22 (-0.29, -0.15) | 0.03 |
| | | $\beta_{12}$ | 0.50 | 0.46 (0.35, 0.57) | -0.04 |
| | Dispersion | $\beta_{20}$ | -0.50 | -0.52 (-0.61, -0.43) | -0.02 |
| | | $\beta_{21}$ | 0.50 | 0.53 (0.48, 0.58) | 0.03 |
| | Random effects | $\Sigma_{11}$ | 0.70 | 0.48 (0.35, 0.64) | -0.22 |
| | | $\Sigma_{22}$ | 0.50 | 0.38 (0.28, 0.51) | -0.12 |
| Model 3 | Count | $\beta_{10}$ | 0.50 | 0.16 (0.00, 0.33) | -0.34 |
| | | $\beta_{11}$ | -0.25 | -0.36 (-0.43, -0.29) | -0.11 |
| | | $\beta_{12}$ | 0.50 | 0.36 (0.21, 0.50) | -0.14 |
| | Dispersion | $\beta_{20}$ | -0.50 | -0.77 (-0.83, -0.71) | 0.23 |
| | | $\beta_{21}$ | 0.50 | 0.39 (0.36, 0.42) | -0.11 |
| | Random effects | $\Sigma_{11}$ | 0.70 | 0.74 (0.59, 0.93) | 0.04 |
| Model 4 | Count | $\beta_{10}$ | 0.50 | 0.11 (-0.04, 0.25) | -0.39 |
| | | $\beta_{11}$ | -0.25 | -0.75 (-0.80, -0.69) | -0.50 |
| | | $\beta_{12}$ | 0.50 | 0.26 (0.14, 0.38) | -0.24 |
| | Dispersion | $\alpha$ | - | 2.40 (2.33, 2.48) | - |
| | Random effects | $\Sigma_{11}$ | 0.70 | 0.74 (0.59, 0.94) | 0.04 |

For comparison, we fitted three misspecified models (labeled Models 2–4). Model 2 assumed uncorrelated random intercepts, $\phi_{1i}$ and $\phi_{2i}$, in Eq. (14). Hence, this model assumes that the process governing random variation in incidence is unrelated to the process governing random variation in the dispersion. Consequently, we assumed univariate ICAR distribution for the spatial random effects for each component with conditional precision $\tau_h$ ($h = 1, 2$). We assigned independent conjugate Gamma(0.001, 0.001) priors to the conditional precision parameters $\tau_h$ for both components. We set the initial values for the fixed effects of the mean, $\theta_1$, at the maximum likelihood estimates from a fixed effect NB model while the fixed effects for dispersion component, $\theta_2$, were initialized to 0. We sampled the initial values for the random effects from independent standard normal distributions, and the precision terms $\tau_h$ ($h = 1, 2$) were initialized to one. To update $\phi_{1i}$ and $\phi_{2i}$, we used separate MH steps with a univariate $t$ proposal centered at the previous $\phi_{hi}$ ($h = 1, 2$).

Model 3 assumed fixed and spatial random effects for the mean but only fixed effects in the dispersion, thus ignoring random spatial variation in this component. We assumed a univariate ICAR distribution for the spatial random effect, $\phi_{3i}$, for the mean with conditional precision $\tau_1$. We assigned an independent conjugate Gamma(0.001, 0.001) prior to $\tau_1$. We set the initial values for the fixed effects of the mean, $\theta_1$, at the maximum likelihood estimates from a fixed effect NB model while the fixed effects for dispersion component, $\theta_2$, were initialized to 0. We sampled the initial values for $\phi_{1i}$ for the mean from independent standard normal distribution, and the precision term $\tau_1$ was initialized to one. As with Model 2, a MH step with a univariate $t$ proposal centered at the previous $\phi_{1i}$ was used to update $\phi_{1i}$. 
Model 4 assumed fixed and spatial random effects for the mean but assumed constant dispersion \((\alpha_{ij} = \alpha)\) for all \(i\) and \(j\). Similar to Models 2 and 3, we assumed a univariate ICAR distribution for the spatial random effects for the mean with conditional precision \(\tau_1\). We assigned independent conjugate Gamma(0.001, 0.001) prior to \(\tau_1\). We set the initial values for the fixed effects of the mean, \(\theta_1\), at the maximum likelihood estimates from a fixed effect NB model. We sampled the initial values for the random effects for the mean from independent standard normal distribution, and the precision term \(\tau_1\) was initialized to one. As with Models 2 and 3, a MH step with a univariate \(t\) proposal was used to update \(\phi_{1i}\). We initialized the overdispersion parameter, \(\alpha\), to 1.2. To update \(\alpha\), we used a MH step with a zero-truncated normal proposal with variance 0.0018.

Model 4 assumed fixed and spatial random effects for the mean but assumed constant dispersion \((\alpha_{ij} = \alpha)\) for all \(i\) and \(j\). Similar to Models 2 and 3, we assumed a univariate ICAR distribution for the spatial random effects for the mean with conditional precision \(\tau_1\). We assigned independent conjugate Gamma(0.001, 0.001) prior to \(\tau_1\). We set the initial values for the fixed effects of the mean, \(\theta_1\), at the maximum likelihood estimates from a fixed effect NB model. We sampled the initial values for the random effects for the mean from independent standard normal distribution, and the precision term \(\tau_1\) was initialized to one. As with Models 2 and 3, a MH step with a univariate \(t\) proposal was used to update \(\phi_{1i}\). We initialized the overdispersion parameter, \(\alpha\), to 1.2. To update \(\alpha\), we used a MH step with a zero-truncated normal proposal with variance 0.0018.

We conducted the simulations in R version 4.0.3 (R Core Team, 2021) and ran all the models for 30 000 iterations each with a burn-in of 15 000, which was sufficient in all cases to ensure convergence based on trace plots and Geweke diagnostics p-values. To reduce auto-correlation, we retained every 10th observation. Across all models, we observed adequate MH acceptance ratios ranging from 35% to 43%.

The simulation results are detailed in Table 1. The columns present the true parameter values, the posterior means, 95% credible intervals (Crls), and the estimated bias for the proposed (Model 1) and misspecified models (Models 2–4). The results for the fixed effect parameters for Models 1 and 2 were similar and well estimated. The random effect variance components of Model 1 were all close to their true values. In Model 2, the bias for \(\text{Var}(\phi_{1i}), \Sigma_{11}\), was \(-0.22\), which is substantially smaller than the bias for \(\Sigma_{11}\) from Model 1 (\(-0.09\)). In addition, the 95% Crl for \(\Sigma_{11}\) from Model 2 did not include the true value. Most parameters for Model 3 were poorly estimated. Likewise, all the parameters for Model 4 exhibited large bias and poor Crl coverage.

As expected, the proposed model performed best, followed by Models 2 and 3. These findings suggest that ignoring or oversimplifying spatial heterogeneity in the dispersion component has an impact on the mean. This highlights a critical point that proper specification for dispersion is needed for correct model inferences, even when interest is in the mean. Table 2 summarizes the goodness of fit statistics for different models using WAIC. Rows indicate the model while the first three columns represent various components of the WAIC statistic (likelihood, penalty, and Total WAIC). The last column represents the successive differences in WAIC values from the previous model. Model 1 slightly outperformed Model 2; Models 3 and 4 showed poor fit. These results again suggest that ignoring the spatial effects in dispersion (Model 3) or assuming constant dispersion (Model 4) can lead to poor model fit when the mean and dispersion processes do, in fact, vary spatially in a correlated manner.

Figs. 2(a) and 2(b) present, respectively, the true and estimated incidence rate and overdispersion trends over time. In particular, the mean trend followed a sinusoidal pattern with higher counts around days 18, 90, 185, and 250. A similar sinusoidal pattern for overdispersion was observed with higher overdispersion around days 30, 120, 230, and 280. Overall, the estimated time trends closely mimic the simulated trends suggesting the proposed model properly fits the data. If the effect of time on incidence case rates or overdispersion was negligible, the line trend for \(f_1(t)\) or \(f_2(t)\) would be horizontal through zero; in that case, the spline function for that component would be unnecessary.

Figs. 3(a)–3(d) present a map of the true values and posterior mean predictions for \(\phi_{1i}\) and \(\phi_{2i}\) \((i = 1, \ldots, n)\) under Model 1. In both cases, the spatial pattern for the estimated effects closely resembles the true spatial distribution, suggesting that the proposed model precisely recovered the underlying spatial pattern in the data. Of note, the model was able to accurately capture the spatial pattern for the dispersion, which can in turn lead to improved model fit.

### Table 2
Model comparison results for Simulation Study.

| WAIC estimate | Likelihood | Penalty | Total | Successive difference |
|---------------|------------|---------|-------|-----------------------|
| Model 1       | −75701.53  | 261.44  | 151925.90 | –                     |
| Model 2       | −75714.45  | 259.09  | 151947.10 | 21.20                 |
| Model 3       | −76294.23  | 202.50  | 152993.40 | 1046.30               |
| Model 4       | −77208.52  | 187.39  | 154791.80 | 1798.42               |

Figs. 2(a) and 2(b) present, respectively, the true and estimated incidence rate and overdispersion trends over time. In particular, the mean trend followed a sinusoidal pattern with higher counts around days 18, 90, 185, and 250. A similar sinusoidal pattern for overdispersion was observed with higher overdispersion around days 30, 120, 230, and 280. Overall, the estimated time trends closely mimic the simulated trends suggesting the proposed model properly fits the data. If the effect of time on incidence case rates or overdispersion was negligible, the line trend for \(f_1(t)\) or \(f_2(t)\) would be horizontal through zero; in that case, the spline function for that component would be unnecessary.
Fig. 2. (a): Average time trend for the mean in simulation study. (b): Average time trend for dispersion in simulation study.

5. Analysis of Georgia COVID-19 data

To analyze the Georgia COVID-19 data, we follow the same setup described in Section 4 and fit the following spatial NB model with correlated random intercepts for the mean and dispersion:

\[
\eta_{ij} = \beta_{10} + \beta_{11}SVI_i + \beta_{12}PctPoorHealth_i + \beta_{13}PctFemale_i + \beta_{14}PctSmoke_i \\
+ \beta_{15}PopDensity_i + f_1(t_j) + \phi_{1i},
\]

\[
\log \alpha_{ij} = \beta_{20} + \beta_{21}SVI_i + \beta_{22}PctPoorHealth_i + \beta_{23}PctFemale_i + \beta_{24}PctSmoke_i \\
+ \beta_{25}PopDensity_i + f_2(t_j) + \phi_{2i},
\]
Fig. 3. Simulated and predicted random effects for the mean and dispersion components for Model 1: spatial NB model with correlated random intercepts. Panel (a): Simulated random intercept for the mean component. Panel (b): Predicted random intercept for the mean component. Panel (c): Simulated random intercept for dispersion component. Panel (d): Predicted random intercept for dispersion component. Legends correspond to sample quintiles.

where SVI is the social vulnerability index for county $i$, PctPoorHealth is the percentage in fair or poor health in county $i$, PctFemale is the percentage of female residents in county $i$, PctSmoke is the percentage of adult smokers in county $i$, and PopDensity is the population density of county $i$. As in the simulation, we model time as a smooth function, $f_h(t_j) (h = 1, 2)$, which we approximate by cubic $B$-splines with 20 interior knots placed 14 days apart. For SVI, we considered three functional forms: linear, quadratic, and cubic $B$-splines. However, because the WAIC values for all
Table 3: Posterior mean rate ratios and 95% posterior intervals for the COVID-19 study from the spatial NB model with correlated intercepts (excluding spline parameters for time).

| Model component | Variable               | Rate ratio (95% CrI)       |
|-----------------|------------------------|---------------------------|
| Mean            | SVI                    | 1.20 (1.07, 1.32)         |
|                 | % fair or poor health  | 1.06 (0.94, 1.22)         |
|                 | % female               | 0.82 (0.78, 0.86)         |
|                 | % of adult smokers     | 0.98 (0.84, 1.13)         |
|                 | Population density     | 0.80 (0.75, 0.85)         |
| Dispersion      | SVI                    | 1.12 (1.05, 1.15)         |
|                 | % fair or poor health  | 0.94 (0.89, 1.00)         |
|                 | % female               | 0.85 (0.83, 0.89)         |
|                 | % of adult smokers     | 1.07 (0.99, 1.13)         |
|                 | Population density     | 0.83 (0.80, 0.86)         |
| Random effects  | var(φ₁₁)               | 0.82 (0.63, 1.06)         |
|                 | cov(φ₁₁, φ₂₁)          | 0.54 (0.40, 0.72)         |
|                 | var(φ₂₁)               | 0.52 (0.41, 0.67)         |
|                 | corr(φ₁₁, φ₂₁)         | 0.82 (0.75, 0.87)         |

The three models were nearly identical, we opted for the more parsimonious model with linear SVI trend. Since population density provides a measure of rurality, we did not include rurality as an independent predictor in the model. Moreover, because the final SVI measure is composed of variables such as percent poverty, percent minority, and percent age 65 and older, we did not include these variables as separate independent variables in the models.

Similar to Section 4, we assigned weakly informative normal priors to both the fixed effect and spline regression coefficients. We assigned a BICAR prior for φ with an IW(3, I₂) prior for the spatial covariance Σ. We ran the MCMC algorithm for 20 000 iterations with a burn-in of 10 000 and a thin of 5. Model diagnostics such as the trace plots and Geweke diagnostics p-values greater than 0.05 suggest the MCMC chains achieved convergence.

Posterior mean RRs (exponentiated coefficients) for the mean and dispersion components are presented in Table 3. The SVI was positively associated with COVID-19 incident cases in the adjusted model, while the percentage of female residents and population density exhibited a negative association with COVID-19 incident cases. The SVI estimate suggests that each unit increase in SVI was on the average associated with a 20% increase in COVID-19 incidence (RR 1.20; 95% CrI: 1.07, 1.32). Additionally, these results confirm the somewhat surprising descriptive finding noted earlier in Section 2 about a possible inverse relationship between COVID-19 incidence and population density in the state of Georgia. Specifically, results from the adjusted model suggest that each unit increase in population density was on the average associated with a 20% decrease in COVID-19 incidence (RR 0.80; 95% CrI: 0.75, 0.85). When compared to the mean component, most estimates for the dispersion component preserve the same direction. The SVI was significant and positively associated with overdispersion while the percentage of female residents was significantly negatively associated with decreased overdispersion. In particular, each unit increase in SVI was associated with a 12% increase in overdispersion of COVID-19 incident case counts per 100 000 individuals (RR 1.12; 95% CrI: 1.05, 1.15).

The variance component estimates suggest a moderate degree of variability in the random effects across counties, with greater heterogeneity in the mean than in the dispersion component (Σ₁₁ : 0.82 vs Σ₂₂ : 0.52). This is not surprising since we expect counties to differ more in terms of underlying factors that affect the incidence rates than in terms of factors that affect overdispersion. The estimate for the random effect correlation ρ was 0.82 (95% CrI: 0.75, 0.87), suggesting a strong positive correlation between the mean and dispersion components, and further supporting the appropriateness of a model with spatially correlated random intercepts.

Figs. 4(a) and 4(b) display the estimated average trends across counties for incident cases and overdispersion, respectively. The observed daily mean incidence rates across counties are overlaid in Fig. 4(a). The average daily incidence increased steadily from March 15 to mid-April before leveling
off in May to early June. This plateau may be the result of social distancing mandates put into effect in April, 2020. From mid June to July, the incidence rose steadily and remained elevated through August. The incidence tapered off in September and early October, before a precipitous increase through the end of year as a “third wave” of infections set in. Fig. 4(b) suggests an increasing rate of overdispersion from March 15 through the first week of April. This increase in overdispersion could be due to lack of incident case reports across many counties early on in the pandemic, and hence the data were possibly zero-inflated. The overdispersion trend leveled-off between mid-April to late May before steadily decreasing until early August. From September through the end of study period, the overdispersion trend appeared to flatten, possibly suggesting that a model with a constant time trend would provide adequate fit to the data during this time interval. Overall, however, the non-constant overdispersion trend further highlights the model fitting flexibility afforded by the proposed model.
**Fig. 5.** Model-based spatial random effect estimates for the COVID-19 study. Panel (a): Posterior mean predicted random effects for the mean component ($\phi_1$). Panel (b): Posterior mean predicted random effects for the dispersion component ($\phi_2$). Legends correspond to sample quintiles.

**Fig. 5** presents the model-based spatial random effect predictions for the mean and dispersion components. Both figures suggest a substantial variation in the spatial distribution of random effects within each component but similar spatial patterns across components. The between-component similarities in spatial patterns support the hypothesis of increased overdispersion in counties with higher incident rates.

Counties with darker shades in panel (a) exhibit higher expected incidence compared to typical (i.e., $\phi_1 = 0$) counties while those in white shade have lower expected incident rates. Similarly, the dark shaded counties in panel (b) suggest higher overdispersion compared to typical (i.e., $\phi_2 = 0$) counties while those in white shade have lower overdispersion. Specifically, the maps indicate high incidence rates and levels of overdispersion among rural counties, particularly those in the southwest that were home to severe outbreaks early in the pandemic (Galofaro, 2020). This result corroborates the negative association for population density that we observed in Table 3. In effect, rural areas were prone to occasional outbreaks that caused extreme spikes in incidence given the small population sizes. This in turn led to increased overdispersion in these rural counties. The random effects for counties within large suburban areas such as Atlanta (north central portion) and Augusta (central eastern border) are lightly or moderately shaded suggesting lower incidence and overdispersion relative to more rural and less populated counties such as in southwest Georgia.

As a sensitivity analysis, we refit the model using the RW1 prior for the spline coefficients, as described in Section 3. The results are presented in Supplementary Figures S2 and S3 and closely mirror those from our primary analysis assuming independent normal priors for the spline coefficients.

Finally, because COVID-19 case data can be susceptible to various types of measurement error, we conducted a second sensitivity analysis using data from USAFacts (USAFacts, 2022) to determine if our primary results based on JHU data were consistent across different publicly available databases. USAFacts has been used widely in COVID-19 studies (Millett et al., 2020; Whaley et al., 2021; Madlock-Brown et al., 2022). Results using the USAFacts data with RW1 prior are presented in Supplementary Figures S2 and S3 and are nearly identical to those based on the JHU data, suggesting that the findings are robust to different sources of publicly available data.
6. Discussion

This paper proposed a NB model with spatially correlated random effects for the mean and dispersion components. Both components consist of region level predictors, temporal trends, and random effects that are modeled via a bivariate CAR prior. The choice of this prior is appealing since it induces correlation between the mean and dispersion components of the model, an attractive feature if regions with high incidence rates also exhibit more overdispersion. Results from our simulation study suggest that ignoring the spatial correlation when the two components are truly correlated leads to biased fixed effect and covariance estimates. Moreover, modeling the overdispersion parameter as a function of region level covariates (e.g. SVI) and temporal B-splines provides additional model flexibility, as it is uncommon for overdispersion to remain constant over space and time in disease-mapping applications such as the COVID-19 study presented here.

Our COVID-19 study for incident cases in Georgia revealed several important findings. First, counties with high SVI had significantly higher incident case rates and more overdispersion. Second, results suggest that incident case rates and overdispersion were generally higher in counties with low population density. This finding suggests that the pandemic has had a disproportionate effect on the incident rate in more vulnerable, rural communities. This may point to inadequate access to health care or lack of appropriate health resources in rural areas. The maps of model-based spatial random effect estimates in Fig. 5 accurately highlight several rural counties in southwest Georgia that reported COVID-19 outbreaks arising from superspreader events such as funerals (Galofaro, 2020). Thus, the proposed model could be a useful tool to inform policy decisions related to COVID-19 or future pandemics. In particular, the model could be used to ensure effective distribution of health resources and personnel in more vulnerable communities in Georgia, throughout the US, and in other parts of the world faced with emerging infectious diseases that may exhibit spatially and temporarily varying overdispersion.

There are some limitations to our analysis. In particular, measurement error (due, for example, to underreporting) poses a challenge in analyzing COVID-19 case data. While our sensitivity analysis using the USAFacts data does not address underreporting per se, it lends some confidence that our analyses yield consistent results across different publicly available and widely used data sources. Future work might tackle the issue of measurement error more directly using recent approaches proposed by Zhang and Yi (2022) and de Oliveira et al. (2022). Our model could also be applied to COVID-19 deaths, which are less prone to underreporting; however, zero-inflated models may be warranted in this case due to the high proportion of zeros present in COVID-19 fatality data. Future work could also permit county-specific time trends by incorporating space–time interactions into the model, as discussed in Napier et al. (2016), Rushworth et al. (2014), and Lee et al. (2018). Finally, following work of Sahu and Böhning (2021) and Neelon et al. (2022), the proposed model could also be extended to jointly model cases and deaths as part of future work.

The methods discussed in this paper can be applied to other research areas, such as the analysis of the human microbiome. This work could be also extended to address underdispersed data through the development of generalized Poisson (Consul, 1989) models with spatially varying dispersion parameters. Underdispersion has increasingly been recognized as an important feature of COVID-19 data (Kobak, 2022). More generally, our approach should prove useful when interest lies in modeling spatial count data with non-constant dispersion.

Funding and disclaimer

This work was supported by grant HX002299-01A2 from the Department of Veterans Affairs, USA Health Services Research and Development (HSR&D) program and in part by the Biostatistics Shared Resource, Hollings Cancer Center, Medical University of South Carolina, USA (P30 CA138313). This article represents the views of the authors and not those of the Department of Veterans Affairs, HSR&D, the United States government, or the Biostatistics Shared Resource, Hollings Cancer Center. The funders had no influence on the study design, implementation, or findings.
Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spasta.2022.100703.

References

Banerjee, S., Carlin, B., Gelfand, A., 2014. Hierarchical Modeling and Analysis for Spatial Data, second ed. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Best, N.G., Ickstadt, K., Wolpert, R.L., 2000. Spatial Poisson regression for health and exposure data measured at disparate resolutions. J. Amer. Statist. Assoc. 95 (452), 1076–1088. http://dx.doi.org/10.1080/01621459.2000.10474304.

Centers for Disease Control and Prevention, 2018. Social vulnerability index 2018 database US. URL https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html.

Consul, P., 1989. Generalized Poisson Distributions: Properties and Applications. Marcel Dekker, New York, Basel.

Eilers, P.H.C., Marx, B., 1996. Flexible smoothing with B-splines and penalties. Statist. Sci. 11 (2), http://dx.doi.org/10.1214/ss/1038425655.

Galofaro, C., 2020. ‘It’s gone haywire’: When COVID-19 arrived in rural america. URL https://apnews.com/article/new-york-race-and-ethnicity-fl-state-wire-ny-state-wire-virus-outbreak-b2a2add19ce7f4f75f42b29331034706. (Accessed 05 April 2021).

Geedipally, S., Lord, D., 2008. Effects of varying dispersion parameter of Poisson–Gamma models on estimation of confidence intervals of crash prediction models. Transp. Res. Rec. J. Transp. Res. Board 2061 (1), 46–54. http://dx.doi.org/10.3141/2061-06.

Georgia Department of Public Health, 2020. Gov. Kemp, officials confirm two cases of COVID-19 in Georgia. URL https://dph.georgia.gov/press-releases/2020-03-02/gov-kemp-officials-confirm-two-cases-covid-19-georgia. (Accessed 05 April 2021).

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics. Vol. 4. Clarendon Press, Oxford, pp. 169–193.

Heydecker, B.G., Wu, J., 2001. Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference. Adv. Eng. Softw. 32 (10–11), 859–869. http://dx.doi.org/10.1016/s0965-9978(01)00037-0.

Hilbe, J., 2011. Negative Binomial Regression. Cambridge University Press.

Johns Hopkins University, 2020. COVID-19 case counts. URL https://www.arcgis.com/home/item.html?id=62857697b2d4dce4c3f9df45ae1843. (Accessed 05 April 2021).

Kobak, D., 2022. Underdispersion: A statistical anomaly in reported Covid data. Significance 19 (2), 10–13. http://dx.doi.org/10.1111/1740-9713.01627.

Lee, D., Rushworth, A., Napier, G., 2018. Spatio-temporal areal unit modeling in R with conditional autoregressive priors using the CARBayesST package. J. Stat. Softw. 84 (9). http://dx.doi.org/10.18637/jss.v084.i09.

Madlock-Brown, C., Wilkens, K., Weiskopf, N., Cesare, N., Bhattacharyya, S., Riches, N.O., Espinoza, J., Dorr, D., Goetz, K., Phuong, J., Sule, A., Kharrazi, H., Liu, F., Lemon, C., Adams, W.G., 2022. Clinical, social, and policy factors in COVID-19 cases and deaths: methodological considerations for feature selection and modeling in county-level analyses. BMC Public Health 22 (1), http://dx.doi.org/10.1186/s12889-022-13168-y.

Mardia, K.V., 1988. Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. J. Multivariate Anal. 24 (2), 265–284.

Miaou, S.-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: Dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transp. Res. Rec. J. Transp. Res. Board 1840 (1), 31–40. http://dx.doi.org/10.3141/1840-04.

Millett, G.A., Jones, A.T., Benkeser, D., Baral, S., Mercer, L., Beyrer, C., Honermann, B., Lankiewicz, E., Mena, L., Crowley, J.S., Sherwood, J., Sullivan, P.S., 2020. Assessing differential impacts of COVID-19 on black communities. Ann. Epidemiol. 47, 37–44. http://dx.doi.org/10.1016/j.annemepidemiol.2020.05.003.

Miranda-Moreno, L.F., Fu, L., Saccomanno, F.F., Labbe, A., 2005. Alternative risk models for ranking locations for safety improvement. Transp. Res. Rec. J. Transp. Res. Board 1908 (1), 1–8. http://dx.doi.org/10.3141/0361198105190800101.

Napier, G., Lee, D., Robertson, C., Lawson, A., Pollock, K.G., 2016. A model to estimate the impact of changes in MMR vaccine uptake on inequalities in measles susceptibility in Scotland. Stat. Methods Med. Res. 25 (4), 1185–1200. http://dx.doi.org/10.1177/096228021660420.

Neele, B., 2019. Bayesian zero-inflated negative binomial regression based on pólya-Gamma mixtures. Bayesian Anal. 14 (3), http://dx.doi.org/10.1214/18-ba1132.
Neelon, B., Wen, C.-C., Benjamin-Neelon, S.E., 2022. A multivariate spatiotemporal model for tracking COVID-19 incidence and death rates in socially vulnerable populations. J. Appl. Stat. 1–24. http://dx.doi.org/10.1080/02664763.2022.2046713.

de Oliveira, G.L., Argiento, R., Loschi, R.H., Assunção, R.M., Ruggeri, F., Branco, M.D., 2022. Bias correction in clustered underreported data. Bayesian Anal. 17 (1), http://dx.doi.org/10.1214/20-Ba1244.

Pillow, J., Scott, J., 2012. Fully Bayesian inference for neural models with negative-binomial spiking. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems. Vol. 25. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2012/file/b55ec28c52d5f6205684a473a2193564-Paper.pdf.

Plummer, M., Best, N., Cowles, K., Vines, K., 2006. CODA: Convergence diagnosis and output analysis for MCMC. R News 6 (1), 7–11. URL https://journal.r-project.org/archive/.

Polson, N.G., Scott, J.G., Windle, J., 2013a. Bayesian inference for logistic models using pólya-Gamma latent variables. J. Amer. Statist. Assoc. 108 (504), 1339–1349. http://dx.doi.org/10.1080/01621459.2013.829001.

Polson, N., Scott, J., Windle, J., 2013b. Bayeslogit: Bayesian inference for logistic models using pólya-Gamma latent variables. Most recent version: Feb. 2013.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rushworth, A., Lee, D., Mitchell, R., 2014. A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in greater London. Spat. Spatio-Temp. Epidemiol. 10, 29–38. http://dx.doi.org/10.1016/j.sste.2014.05.001.

RWJ Foundation, 2020. County health rankings & roadmaps: Rankings data & documentation. URL https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation. (Accessed 05 April 2021).

Sahu, S.K., Böhning, D., 2021. Bayesian spatio-temporal joint disease mapping of Covid-19 cases and deaths in local authorities of England. Spatial Stat. 100519. http://dx.doi.org/10.1016/j.spasta.2021.100519.

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., Boatwright, P., 2005. A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. J. R. Stat. Soc. Ser. C. Appl. Stat. 54 (1), 127–142. http://dx.doi.org/10.1111/j.1467-9876.2005.00474.x.

Stewart, J.A., 1994. The poisson-lognormal model for bibliometric/scientometric distributions. Inform. Process. Manage. 30 (2), 239–251. http://dx.doi.org/10.1016/0306-4573(94)90067-1.

US Census Bureau, 2021. County population totals: 2010–2019. URL https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html. (Accessed 05 April 2021).

USA Facts, 2022. US COVID-19 cases and deaths by state. URL https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/. Retrieved April 24, 2022.

Wakefield, J., 2006. Disease mapping and spatial regression with count data. Biostatistics 8 (2), 158–183. http://dx.doi.org/10.1093/biostatistics/kx008.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11, 3571–3594.

Whaley, C.M., Cantor, J., Pera, M., Jena, A.B., 2021. Assessing the association between social gatherings and COVID-19 risk using birthdays. JAMA Intern. Med. 181 (8), 1090. http://dx.doi.org/10.1001/jamainternmed.2021.2915.

Zhang, Q., Yi, G.Y., 2022. Zero-inflated Poisson models with measurement error in the response. Biometrics http://dx.doi.org/10.1111/biom.13657.