Journal of Data Science 20 (3), 325–337
July 2022

DOI: 10.6339/21-JDS1032
Statistical Data Science

Bayesian Inference for Spatial Count Data that May be Over-Dispersed or Under-Dispersed with Application to the 2016 US Presidential Election

Hou-Cheng Yang[1,*] and Jonathan R. Bradley[1,*]

[1]*Department of Statistics, Florida State University, Tallahassee, FL, USA*

Abstract

We propose a method of spatial prediction using count data that can be reasonably modeled assuming the Conway-Maxwell Poisson distribution (COM-Poisson). The COM-Poisson model is a two parameter generalization of the Poisson distribution that allows for the flexibility needed to model count data that are either over or under-dispersed. The computationally limiting factor of the COM-Poisson distribution is that the likelihood function contains multiple intractable normalizing constants and is not always feasible when using Markov Chain Monte Carlo (MCMC) techniques. Thus, we develop a prior distribution of the parameters associated with the COM-Poisson that avoids the intractable normalizing constant. Also, allowing for spatial random effects induces additional variability that makes it unclear if a spatially correlated Conway-Maxwell Poisson random variable is over or under-dispersed. We propose a computationally efficient hierarchical Bayesian model that addresses these issues. In particular, in our model, the parameters associated with the COM-Poisson do not include spatial random effects (leading to additional variability that changes the dispersion properties of the data), and are then spatially smoothed in subsequent levels of the Bayesian hierarchical model. Furthermore, the spatially smoothed parameters have a simple regression interpretation that facilitates computation. We demonstrate the applicability of our approach using simulated examples, and a motivating application using 2016 US presidential election voting data in the state of Florida obtained from the Florida Division of Elections.

Keywords *Bayesian inference; Conway-Maxwell; count data; dispersion; Poisson distribution; spatial statistics*

1 Introduction

In United States presidential elections there are many states that historically vote for the same political party. For example, from recent past electoral results, Republican candidates tend to win most of the mountain states and Great Plains, and most of the South. Similarly, Democratic candidates often win the Mid-Atlantic states along with New England and the West Coast states. There are also so-called swing states, which refers to any state that could reasonably be won by either the Democratic or Republican presidential candidates. For example, Texas is the key to outcome of the 1960 election, Florida and New Hampshire are key in deciding the 2000 election, and Ohio was important during the 2004 election (Duquette et al., 2017). Thus, there is reason to suggest under-dispersion for some regions and over-dispersion in others, where under (over)

*Corresponding author. Email: hy15e@my.fsu.edu or jrbradley@fsu.edu.

Received August 30, 2021; Accepted November 27, 2021

dispersion refers to when the mean is larger (smaller) than the variance of the data. That is, historically consistent voters suggest that variance within certain regions might be very small and vice versa.

In the context of modeling spatial count data, such as US voting data, the Poisson distribution has gained widespread popularity (Cressie, 1993; Banerjee et al., 2014). In general, spatial statistical models for spatial count data have become a common choice in a large number of scientific disciplines. For example, spatial count data have been modeled in areas as diverse as small-area samples from surveys, meteorological observations, epidemiological data, transportation data, and relative abundance of various species in ecological monitoring studies, among others (Guikema and Goffelt, 2008; Gupta et al., 2014; Sellers and Raim, 2016). Poisson regression is a standard framework for modeling covariate dependent count data through a log-linear link function. Nevertheless, the distributional assumption of having the mean equal to the variance over the entire spatial domain (i.e., equi-dispersion) is rarely satisfied for the types of processes that typically are observed in practice. Count data can exhibit over-dispersion and under-dispersion causing traditional count data regression models to violate this property of the data. Several approaches for addressing over-dispersion have been developed such as quasi-likelihood methods, Poisson regression with random effects, and models based on the negative binomial distribution (Manton et al., 1981; Cameron and Trivedi, 2013; Ver Hoef and Boveng, 2007; Hilbe, 2011). Models based on the negative binomial distribution fit reasonably well for over-dispersed data, but often under perform in under-dispersion settings (Lindén and Mäntyniemi, 2011). Hence, methods that allow for both under-dispersion and over-dispersion are important contributions to the analysis of count-valued data.

There are several recent proposed flexible models that use the Conway-Maxwell Poisson distribution (COM-Poisson) (Conway and Maxwell, 1962), and have the potential to overcome the limitations of traditional count models. The COM-Poisson distribution was introduced into the statistics literature by (Shmueli et al., 2005) with several follow-up papers (e.g., see Daly and Gaunt, 2016; Chakraborty and Imoto, 2016; Sellers et al., 2016) Shmueli et al. (2005) demonstrate the probabilistic and statistical properties of the COM-Poisson distribution and describe several methods of parameter estimation. In particular, the COM-Poisson distribution is a member of the exponential family and can be seen as an extension to the Poisson distribution, with an extra parameter that flexibly controls the level of dispersion. Additionally, this distribution has the Poisson and geometric distributions as special cases and the Bernoulli distribution as a limiting case. The COM-Poisson provides a promising and flexible approach for performing count data regression. This distribution has become widely utilized in a variety of applications and has great practical interest. However, the statistical performance of this model has not yet been fully characterized. A comprehensive overview regarding the COM-Poisson model is provided by Sellers et al. (2012). The COM-Poisson distribution has since been used in more sophisticated Bayesian hierarchical models formulated to dynamically accommodate varying levels of spatial dispersion (Wu et al., 2013).

The application of hierarchical Bayesian models for spatial and spatio-temporal count data has become increasingly popular over the past decades. For example, Wikle and Hooten (2006) and Hooten et al. (2007) proposed spatio-temporal Poisson models. Bradley et al. (2018) also develops multivariate spatio-temporal models for high-dimensional count data. Moreover, there are many examples in the epidemiological field (see, e.g. Waller et al., 1997; Carlin and Banerjee, 2003, and beyond). Wu et al. (2013) developed a space-time COM-Poisson model. However, the overdispersion from the spatial random effects obfuscate the dispersion of the data. In this article, we propose a computationally efficient hierarchical Bayesian spatial COM-Poisson model

that preserves the dispersion properties of the data and simultaneously include spatial random effects. To do this, we assume conditional independence between the data and the spatial process conditioned on a set of parameters. We refer to this model as Over or Underdispersed Regression for Spatial (OURS) count data.

We consider a modified version of the conjugate prior distribution for the COM-Poisson in Kadane et al. (2006). This modification completely avoids the need to compute the normalizing constant in the COM-Poisson. Furthermore, we show that our modified prior distribution leads to a well defined posterior distribution for the parameters in a COM-Poisson.

Another contribution of our method is that it is matrix free, which is a growing area in computational statistics (Dai et al., 2020; Yang and Bradley, 2021). To achieve a computationally efficient approach in the Bayesian setting we introduce latent random variables that allow one to model the dependency completely in the mean term. The first step of our procedure is using slice sampling (Neal, 2003) to sample from the aforementioned modified conjugate posterior from Kadane et al. (2006). The next step uses a posterior predictive distribution to a model with functional spatial dependencies defined in the mean. This two step procedure shows that we do not need to store a large matrix and nor do we need to compute the intractable normalizing constant in a COM-Poisson, which aids in computational efficiency.

The remaining sections of this article are organized as follows. We first review the COM-Poisson distribution then propose the methodology and our model framework in Section 2. In Section 3, we present simulated examples for over/under-dispersed counts. In Section 4, we implement our model using 2016 US presidential voting data in the state of Florida. Finally, Section 5 contains a discussion.

2 Methodology

2.1 Review of Bayesian Analysis of COM-Poisson Distributed Data

In this section, we review the COM-Poisson distribution. The COM-Poisson distribution generalizes the Poisson distribution to model over-dispersion or under-dispersion. Define an n-dimensional data vector $\mathbf{Z} = (z_1 \dots z_n)'$, where $z_i \in \{0, 1, 2, \dots\}$ is the count-valued outcome associated with the i-th region for every $i = 1 \dots n$. For example, z_i could represent the number of votes for a Republican candidate in county i. The probability mass function (p.m.f) is

$$f(z_i \mid \lambda_i, \nu_i) = \frac{\lambda_i^{z_i}}{(z_i!)^{\nu_i}} \frac{1}{Q(\lambda_i, \nu_i)}, \tag{1}$$

$$Q(\lambda_i, \nu_i) = \sum_{j=0}^{\infty} \frac{\lambda_i^j}{(j!)^{\nu_i}}; \quad \lambda_i > 0, \nu_i \geqslant 0, i = 1 \dots n, \tag{2}$$

where $Q(\lambda_i, \nu_i)$ serves as a normalization constant and ν_i is called the dispersion parameter.

The level of dispersion can be conveniently characterized with $\nu_i = 1$, $\nu_i < 1$, and $\nu_i > 1$ corresponding to equal-dispersion, over-dispersion, and under-dispersion, respectively. COM-Poisson is a member of the exponential family and it has the Poisson distribution (when $\nu_i = 1$) and geometric distribution (when $\nu_i = 0$ and $\lambda_i < 1$) as special cases and the Bernoulli distribution (when $\nu_i \to \infty$) as a limiting case. When $\nu_i = 0$ and $\lambda_i \geqslant 1$, $Q(\lambda_i, \nu_i)$ does not converge, and the distribution is undefined.

Kadane et al. (2006) used the exponential family structure of the COM-Poisson to establish a conjugate prior density of the form

$$f(\lambda_i, \nu_i) = \lambda_i^{a-1} \exp(-\nu_i b) Q(\lambda_i, \nu_i)^{-c} \kappa(a, b, c), \tag{3}$$

for $\lambda_i > 0$ and $\nu_i \geqslant 0$, where $\kappa(a, b, c)$ is the integration constant. The posterior has the same form as (3), with $a' = a + z_i$, $b' = b + \log(z_i!)$ and $c' = c + 1$. The values of a, b, and c are restricted so that $\kappa^{-1}(a, b, c) < \infty$. Using Jensen's inequality and the convexity of the log-gamma function, a necessary and sufficient condition for a finite $\kappa^{-1}(a, b, c)$ is

$$\frac{b}{c} > \log\left(\left\lfloor \frac{a}{c} \right\rfloor !\right) + \left(\frac{a}{c} - \left\lfloor \frac{a}{c} \right\rfloor\right) \log\left(\left\lfloor \frac{a}{c} \right\rfloor + 1\right), \tag{4}$$

where $\lfloor m \rfloor$ denotes the floor of m (see, Kadane et al., 2006, for a complete proof).

A spatial(-time) alternative to the Kadane et al. (2006)'s prior was introduced in Wu et al. (2013). Here, they assume the vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ has a basis function representation: $\log(\boldsymbol{\lambda}) = \boldsymbol{\mu} + \boldsymbol{\Psi}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, where Gaussian priors are given to the n-dimensional vector $\boldsymbol{\mu}$, r-dimensional vector $\boldsymbol{\alpha}$, and n-dimensional vector $\boldsymbol{\epsilon}$. Here $\boldsymbol{\Psi}$ represents a $n \times r$ matrix of spatial basis functions. Specification of basis functions is a key tool used for spatial and spatio-temporal models (Wikle, 2010). Basis functions imply dependence, since $cov\,(\log(\boldsymbol{\lambda})) = \boldsymbol{\Psi}\, cov(\boldsymbol{\alpha})\boldsymbol{\Psi}'$, which is not equal to a zero matrix.

2.2 A Modified Kadane et al. (2006) Prior Distribution

One can avoid the normalizing constant in the posterior distribution by setting $c = -1$. However, the implied posterior distribution is not proper (i.e., does not integrate to one). We gain flexibility in specifying $c = -1$ by truncating the support of λ_i. Namely consider the following prior distribution for (λ_i, ν_i):

$$h(\lambda_i, \nu_i) = \lambda_i^{a-1} \exp(-\nu_i b) Q(\lambda_i, \nu_i) I(0 < \lambda_i < w), \tag{5}$$

where $I(\cdot)$ is an indicator function, $\nu_i \geqslant 0$, $0 < w < \infty$. Then the posterior distribution for (λ_i, ν_i) is

$$h(\lambda_i, \nu_i | Z_i, a, b, w) = \frac{(Z_i + a)\,\{b + \log(Z_i!)\}\,\lambda_i^{Z_i + a - 1}}{w^{Z_i + a}} \exp\left[-\{b + \log(Z_i!)\}\nu_i\right] I(0 < \lambda_i < w), \tag{6}$$

where $\nu_i \geqslant 0$. The posterior means of λ_i and ν_i are given by $(\frac{w}{Z_i+a+1})Z_i + a(\frac{w}{Z_i+a+1})$ and $1/\left[b + \log(Z_i!)\right]$, respectively. Thus, the choice of w, a, and b are important from the perspective of point estimation. For example, if $w \approx Z_i + a + 1$ and $a \approx 0$ we see that the posterior mean is roughly an unbiased estimator for λ_i. When $b < 1$ and Z_i is equal to zero or one, we obtain a poster mean of ν_i greater than one suggesting under-dispersion. We note that in later stages in our Bayesian hierarchical models we effectively smooth values from this posterior predictive distribution (see Section 2.4), and hence, these smoothed parameter estimates have different properties.

2.3 Over or Underdispersed Regression for Spatial Data

In order to incorporate dependence in an efficient manner, we introduce latent random effects into a hierarchical model. The joint distribution is given by

$$\prod_{i=1}^{n} f(z_i | \lambda_i, \nu_i) h(\lambda_i, \nu_i | a, b, w)\, f(\boldsymbol{\beta}_\lambda, \boldsymbol{\beta}_\nu | \boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\theta}), \tag{7}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)'$, $f(z_i|\lambda_i, \nu_i)$ is the p.m.f of the Conway-Maxwell Poisson distribution in Equation (1), $h(\lambda_i, \nu_i|a, b, w)$ is the prior distribution in Equation (5), and we specify $f(\boldsymbol{\beta}_\lambda, \boldsymbol{\beta}_\nu|\boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\theta})$ to be a multivariate normal distribution. The parameter λ_i and ν_i are the location and dispersion parameters in a COM-Poisson distribution, and β_λ and β_ν will be defined as smoothed version of these parameters that take into accounts covariates and spatial dependence. We let

$$f(\boldsymbol{\beta}_\lambda, \boldsymbol{\beta}_\nu|\boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\theta}) = f(\boldsymbol{\beta}_\lambda|\boldsymbol{\lambda}, \sigma_\lambda^2) f(\boldsymbol{\beta}_\nu|\boldsymbol{\nu}, \sigma_\nu^2), \tag{8}$$

so that $\boldsymbol{\theta} = (\sigma_\lambda^2, \sigma_\nu^2)'$, where $\sigma_\lambda^2 > 0$, $\sigma_\nu^2 > 0$, and the n-dimensional random vectors

$$\begin{aligned}
&\boldsymbol{\beta}_\lambda|\boldsymbol{\lambda}, \sigma_\lambda^2 \sim \text{Normal}(\mu_\lambda, \sigma_\lambda^2 \mathbf{I}_n),\\
&\boldsymbol{\mu}_\lambda = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\log(\boldsymbol{\lambda}) + \boldsymbol{\Psi}(\boldsymbol{\Psi}'\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'(\log(\boldsymbol{\lambda}) - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\log(\boldsymbol{\lambda})),\\
&\boldsymbol{\beta}_\nu|\boldsymbol{\nu}, \sigma_\nu^2 \sim \text{Normal}(\mu_\nu, \sigma_\nu^2 \mathbf{I}_n),\\
&\boldsymbol{\mu}_\nu = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\log(\boldsymbol{\nu}) + \boldsymbol{\Psi}(\boldsymbol{\Psi}'\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'(\log(\boldsymbol{\nu}) - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\log(\boldsymbol{\nu})),
\end{aligned} \tag{9}$$

where the n-dimensional vectors $\boldsymbol{\lambda} = (\lambda_1, \ldots \lambda_n)'$ and $\boldsymbol{\nu} = (\nu_1, \ldots \nu_n)'$ and "Normal($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$)" is a shorthand for a multivariate normal distribution with mean $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{x}_i$ be a known p-dimensional vector of covariates, $\mathbf{X} = \{\mathbf{x}_1 \ldots \mathbf{x}_n\}'$, $\boldsymbol{\Psi} \in \mathbb{R}^n \times \mathbb{R}^r$ is defined to be a matrix of basis functions which is of dimension $n \times r$ ($r \leqslant n$). We also assume $\sigma_\lambda^2 \sim \text{IG}(1, 1)$ and $\sigma_\nu^2 \sim \text{IG}(1, 1)$ where "IG(α, κ)" is a shorthand for the inverse gamma distribution with shape $\alpha > 0$ and scale $\kappa > 0$. Notice that the means of $\boldsymbol{\beta}_\lambda$ and $\boldsymbol{\beta}_\nu$ are a series of projections of $\log(\boldsymbol{\lambda})$ and $\log(\boldsymbol{\nu})$ onto the column space of $\mathbf{X}$ and $\boldsymbol{\Psi}$, respectively. In Equation (9), we see that the mean of $\boldsymbol{\beta}_\lambda$ is a smoothed (i.e., projection) version of the location parameter $\boldsymbol{\lambda}$. Consequently, we interpret $\boldsymbol{\beta}_\lambda$ as a location parameter that accounts for covariate (i.e., $\mathbf{X}$) and spatial behavior (i.e., $\boldsymbol{\Psi}$). Similarly, one can interpret $\boldsymbol{\beta}_\nu$ as a dispersion parameter that incorporates covariates and spatial behavior. Consequently, we are interested in inference on $\boldsymbol{\beta}_\lambda$ and $\boldsymbol{\beta}_\nu$ to adjust for covariate and spatial effects when learning about location and dispersion of the data.

Sellers and Shmueli (2013) and Wu et al. (2013) introduced extensions of COM-Poisson regression that allows for different group and spatial levels of dispersion by modeling $\boldsymbol{\nu}$ with additional spatial random effects. However the random effects in these models compete with the over-dispersion parameter $\boldsymbol{\nu}$. We remove this competition by assuming conditional independence between $\{\boldsymbol{\lambda}, \boldsymbol{\nu}\}$ and $\{\boldsymbol{\beta}_\lambda, \boldsymbol{\beta}_\nu\}$, where $\boldsymbol{\beta}_\lambda$ and $\boldsymbol{\beta}_\nu$ include spatial dependence.

2.4 Posterior Predictive Distribution

The posterior predictive distribution for $\{\boldsymbol{\beta}_\lambda, \boldsymbol{\beta}_\nu\}$ is given by

$$f(\boldsymbol{\beta}_\lambda, \boldsymbol{\beta}_\nu|\mathbf{Z}) = \int\int f(\boldsymbol{\beta}_\lambda, \boldsymbol{\beta}_\nu|\boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{Z}) f(\boldsymbol{\lambda}, \boldsymbol{\nu}|\mathbf{Z}) d\boldsymbol{\lambda} d\boldsymbol{\nu} = \int\int f(\boldsymbol{\beta}_\lambda, \boldsymbol{\beta}_\nu|\boldsymbol{\lambda}, \boldsymbol{\nu}) f(\boldsymbol{\lambda}, \boldsymbol{\nu}|\mathbf{Z}) d\boldsymbol{\lambda} d\boldsymbol{\nu}, \tag{10}$$

Although this integral does not have a closed form, Equation (10) leads to straightforward implementation via a Gibbs sampler. See Algorithm 1 for an outline.

Algorithm 1 Implementation.

1: Initialize all the parameters.
2: For b=1... B, sample $\boldsymbol{\lambda}^{[b]}$, $\boldsymbol{\nu}^{[b]}$ from $f(\boldsymbol{\lambda}, \boldsymbol{\nu}|\mathbf{Z})$ (using a slice sampler (Neal, 2003))
3: Set $b = b + 1$.
4: Sample ${\boldsymbol{\beta}_\lambda}^{[b]}$ from Normal$(\boldsymbol{\mu}_\lambda, {\sigma_\lambda^2}^{[b-1]}\mathbf{I}_n)$, where recall $\boldsymbol{\mu}_\lambda$ is defined in (9).
5: Sample ${\boldsymbol{\beta}_\nu}^{[b]}$ from Normal$(\boldsymbol{\mu}_\nu, {\sigma_\nu^2}^{[b-1]}\mathbf{I}_n)$, where recall $\boldsymbol{\mu}_\nu$ is defined in (9).
6: Simulate ${\sigma_\lambda^2}^{[b]}$ from $f(\sigma_\lambda^2|{\boldsymbol{\beta}_\lambda}^{[b]})$, which is the probability density function for a IG $\left(1 + \frac{n}{2}, 1 + \frac{({\boldsymbol{\beta}_\lambda}^{[b]} - \boldsymbol{\mu}_\lambda)'({\boldsymbol{\beta}_\lambda}^{[b]} - \boldsymbol{\mu}_\lambda)}{2}\right)$.
7: Simulate ${\sigma_\nu^2}^{[b]}$ from $f(\sigma_\nu^2|{\boldsymbol{\beta}_\nu}^{[b]})$, which is the probability density function for a IG $\left(1 + \frac{n}{2}, 1 + \frac{({\boldsymbol{\beta}_\nu}^{[b]} - \boldsymbol{\mu}_\nu)'({\boldsymbol{\beta}_\nu}^{[b]} - \boldsymbol{\mu}_\nu)}{2}\right)$.
8: Generate a new replicate of $\mathbf{Z}^{[b]}$ from $\prod_{i=1}^{n} f\left(z_i \mid \lambda_i = \exp\left(\beta_{\lambda_i}^{[b]}\right), \nu_i = \exp\left(\beta_{\nu_i}^{[b]}\right)\right)$, where f is defined in Equation (1) and $\boldsymbol{\beta}_\lambda^{[b]} = \left(\beta_{\lambda_1}^{[b]}, \ldots, \beta_{\lambda_n}^{[b]}\right)'$ and $\boldsymbol{\beta}_\nu^{[b]} = \left(\beta_{\nu_1}^{[b]}, \ldots, \beta_{\nu_n}^{[b]}\right)'$
9: Repeat steps 3 through 8 until b=B.

Thus, our implementation involves generating values from two easy to sample from conditional distributions. That is, first generate $\boldsymbol{\lambda}^{[b]}$ and $\boldsymbol{\nu}^{[b]}$ from $f(\boldsymbol{\lambda}, \boldsymbol{\nu}|\mathbf{Z})$ and then generate the value from $f(\boldsymbol{\beta}_\lambda, \boldsymbol{\beta}_\nu|\boldsymbol{\lambda}^{[b]}, \boldsymbol{\nu}^{[b]})$. It is important to emphasize that Step 3-8 in Algorithm 1 does not require computationally difficult covariance matrices to store and invert. However, this does not imply that we do not model spatial dependence. That is, let $\mathbf{P}_x = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{P}_\Psi = \boldsymbol{\Psi}(\boldsymbol{\Psi}'\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'$ so that

$$
\begin{aligned}
E\,(\mu_\lambda) &= \mathbf{P}_x E\{\log(\boldsymbol{\lambda})\} + \mathbf{P}_\Psi(\mathbf{I} - \mathbf{P}_x)E\{\log(\boldsymbol{\lambda})\},\\
Cov(\mu_\lambda) &= \mathbf{P}_x\, Var\{\log(\boldsymbol{\lambda})\}\mathbf{P}_x + \mathbf{P}_\Psi(\mathbf{I} - \mathbf{P}_x)\, Var\{\log(\boldsymbol{\lambda})\}(\mathbf{I} - \mathbf{P}_x)\mathbf{P}_\Psi,\\
E(\mu_\nu) &= \mathbf{P}_x E\{\log(\boldsymbol{\nu})\} + \mathbf{P}_\Psi(\mathbf{I} - \mathbf{P}_x)E\{\log(\boldsymbol{\nu})\},\\
Cov(\mu_\nu) &= \mathbf{P}_x\, Var\{\log(\boldsymbol{\nu})\}\mathbf{P}_x + \mathbf{P}_\Psi(\mathbf{I} - \mathbf{P}_x)\, Var\{\log(\boldsymbol{\nu})\}(\mathbf{I} - \mathbf{P}_x)\mathbf{P}_\Psi,
\end{aligned}
$$

imply non-zero means and non-diagonal convariance matrices, where the expectations are taken with respect to (7). Thus, predictions from our incorporate spatial dependence, and are computed from Algorithm 1 with

$$\hat{\boldsymbol{Z}} = \frac{1}{B - b_0}\sum_{b=b_0}^{B} \boldsymbol{Z}^{[b]},$$

where b_0 is a burn-in, and in Step 8 we have generated from posterior predictive distribution based $\boldsymbol{\beta}_\lambda$ and $\boldsymbol{\beta}_\nu$ instead of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$. The use of $\boldsymbol{\beta}_\lambda$ and $\boldsymbol{\beta}_\nu$ is preferable because, again, these parameters incorporate spatial and covariate information into the prediction of the mean of $\boldsymbol{Z}$.

3 Illustrations Using Simulated Examples

In this section, we present two different simulation scenarios. The first specifies ν_i so that z_i is under-dispersed and the other specifies z_i to be over-dispersed. We simulate z_i from a COM-Poisson distribution, and our main goal is to assess whether or not $\boldsymbol{\beta}_\lambda$ can be used to accurately estimate the true location parameter. We fix the dispersion parameter ν_i as constant (e.g., $\nu_i = 2$ for all i) in both simulation settings. Values of $\nu > 1$ indicate under-dispersion relative to the

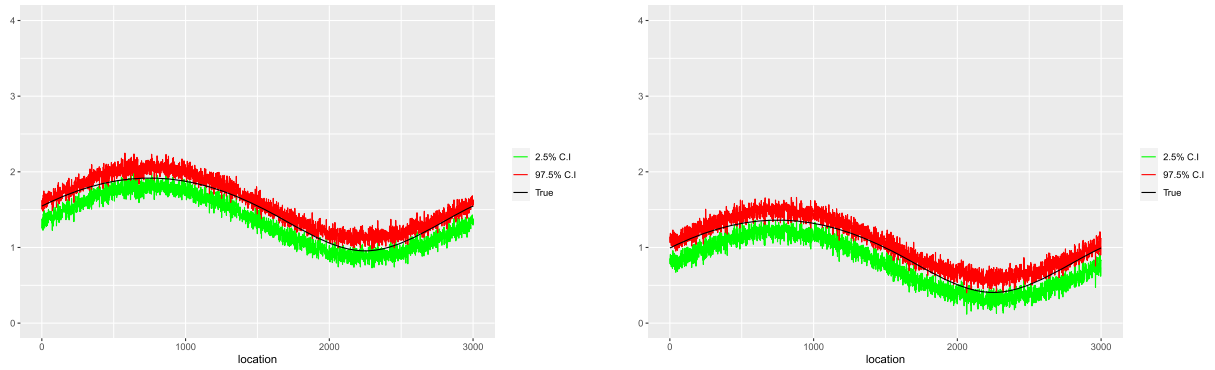

Figure 1: Credible intervals for $\boldsymbol{\beta}_\lambda$ and the true value of $\log(\boldsymbol{\lambda})$ for the one-dimensional region with under-dispersion setting. In the left panel data is generated with $\boldsymbol{\nu} = (2, \ldots, 2)'$ and in the right panel the data is generated with $\boldsymbol{\nu} = (0.95, \ldots, 0.95)'$.

Poisson, while $\nu < 1$ indicates over-dispersion. Also, there are many possible choices for basis functions (see Section 2.1). In this paper, we use the thin plates basis function (Wahba, 1990). The thin plates basis function are defined as

$$\begin{aligned} \boldsymbol{\psi}(\boldsymbol{s}) &\equiv \left(\varphi\left(\left\|\boldsymbol{s} - \mathbf{c}_{1j}\right\|\right), \ldots, \varphi\left(\left\|\boldsymbol{s} - \mathbf{c}_{ij}\right\|\right)\right)', \ j = 1, 2, 3, \\ \varphi\left(\left\|\boldsymbol{s} - \mathbf{c}_{ij}\right\|\right) &= \left(\frac{\|\boldsymbol{s} - \mathbf{c}_{ij}\|}{\alpha_j}\right)^2 \log\left(\frac{\|\boldsymbol{s} - \mathbf{c}_{ij}\|}{\alpha_j}\right), \ j = 1, 2, 3, \end{aligned} \tag{11}$$

where $\boldsymbol{s}$ is a centroid of one of the n study regions (e.g., counties in Florida) and $\{\mathbf{c}_{ij} : i = 1, \ldots r, \ j = 1, 2, 3\}$ represent the knot locations defined over resolutions $j = 1, 2, 3$. Specifically, for $j = 1, 2, 3$ let $\{\mathbf{c}_{ij} : i = 1, \ldots, \omega_j\}$ be equally spaced centroids over the n regions. This implies $r = \omega_1 + \omega_2 + \omega_3$ and $\alpha_i > 0$ is called a bandwidth so that $\boldsymbol{\Psi} = \{\boldsymbol{\psi}(\mathbf{c}_{11}), \ldots, \boldsymbol{\psi}(\mathbf{c}_{\omega_3 3})\}'$. Specifying $\{\mathbf{c}_{ij}\}$ in this way is referred to as a multi-resolutional choice of knots, and is a common choice in spatial statistics (e.g., see Cressie and Johannesson, 2008). Let $\|\cdot\|$ denote the usual Euclidean norm. Also, the choice of hyperparameters (i.e., a, b, w, and r) are set constant across simulated replicates. For each simulation, we implement with Algorithm 1 with $B = 5000$ and treated the first 2000 iterations as a burn-in.

Consider generating data in the following way. Generate 3,000 observations over a one-dimensional spatial domain (i.e., $n = 3{,}000$), such that z_i is distributed according to a COM-Poisson with location λ_i and dispersion parameter ν_i. Define the true $\lambda_i = 2.1\sin(2\pi s_i) + 4.7$ and where the centroid of region i is denoted as $s_i \in \{s_1 \ldots s_{3000}\} \subset [0, 1]$ and are equally spaced. We fix the true dispersion parameter as a constant. When generating under-dispersed data we set $\boldsymbol{\nu} = (2, \ldots, 2)'$, and when generating over-dispersed data we set $\boldsymbol{\nu} = (0.95, \ldots, 0.95)'$. Then to generate from a COM-Poisson the infinite sum in (2) is truncated, and we truncate at 100-th term. This is done using the R-package `CompGLM`. In (4), we set $a = 2$, $b = 2$, and $w = 30$. We implement the model in (7) with $\mathbf{x}_i = (\mathbf{1}, \sin(2\pi s_i))'$ and set $\omega_1 = 4$, $\omega_2 = 5$, and $\omega_3 = 6$ so that we have a total of $r = 15$ thin plate splines over the three resolutions $\alpha_1 = 3.01$, $\alpha_2 = 3.97$, and $\alpha_3 = 5.02$. In Figure 1, we plot the true $\log(\boldsymbol{\lambda})$ and 95% pointwise credible intervals associated with $\boldsymbol{\beta}_\lambda$ using replicates from Algorithm 1. Here we see that the credible intervals display the pattern of the truth in both dispersion settings, and tend to contain $\log(\boldsymbol{\lambda})$. This pattern is consistent when simulating 50 replications. We provide the average performance of parameter

Table 1: Performance of parameter estimates under 50 replicates.

	Dispersion Type	MMAE	MSD	MMSE
$\hat{\lambda}$	Under	0.2181	0.0821	0.0697
	Over	0.1171	0.0232	0.0201
$\hat{\nu}$	Under	0.0375	0.0039	0.0027
	Over	0.0183	0.0008	0.0006

estimates under 50 replicate data vectors in Table 1 by simulation set-up (i.e., over or under dispersion data). We use different metrics to evaluate the posterior performance. Those metrics including the mean (over 50 simulated data vectors) of mean (over elements of our data vector) absolute error (MMAE), the mean standard deviation (MSD), the mean of mean squared error (MMSE). The estimation performance is stable, in the sense that the metrics also small for both parameters. In general, we tend to perform better in the over-dispersion setting.

4 Real Data Application

4.1 Data Description

As an illustration, we analyze voting data from the 2016 United States presidential election. We focus on data in the state of Florida at the county-level. In this application, we ignore the third parties and only focus on two main political parties, the Republican and Democratic parties. The candidate for the Republican party was Donald J. Trump and the candidate for the Democratic party was Hillary Clinton. In 2016, there were 67 counties in Florida. A transformation of the voting results can be seen of Figure 2. This dataset is made publically available on the Florida Division of Elections (https://results.elections.myflorida.com).

4.2 Analysis

We present an analysis of the election dataset using our method. In this analysis, we run 30,000 iterations and treated the first 20,000 iterations as a burn-in. We informally check trace plots for convergence, and no lack of convergence was detected. The main goals of our analysis is to compare the predictive performance of OURS relative to competing methods, and to infer values of over and under-dispersion. Recall, this is not immediately possible to do for our competitors. The response in this analysis is a transformation of the difference in the number of votes between Trump and Clinton. We shift this difference so the smallest value is zero, we rescale (12) by $\frac{1}{30,000}$ for a numeric reason, and we take ceiling so that the response in integer-valued.

$$z_i = \left\lceil \frac{(V_{i,1} - V_{i,2}) - \min(V_{i,1} - V_{i,2})}{30,000} \right\rceil, i = 1 \ldots 67, \tag{12}$$

where $V_{i,1}$ is the number of votes for Trump over each county, $V_{i,2}$ is the number of votes for Clinton over each county, "min" is the minimum, and $\lceil \cdot \rceil$ represent the ceiling function. For illustration, we specify an intercept only model (i.e., $\mathbf{x}_i = 1$). In our model, we set $r = 67$ vector $\boldsymbol{\psi}(\mathbf{s})$ was chosen to consist of thin plates basis functions (see, Equation (11)) associated with the centroids of each county. The bandwidth is set to 0.75 which is chosen based on minimizing a criterion over a range of choices of the bandwidth.

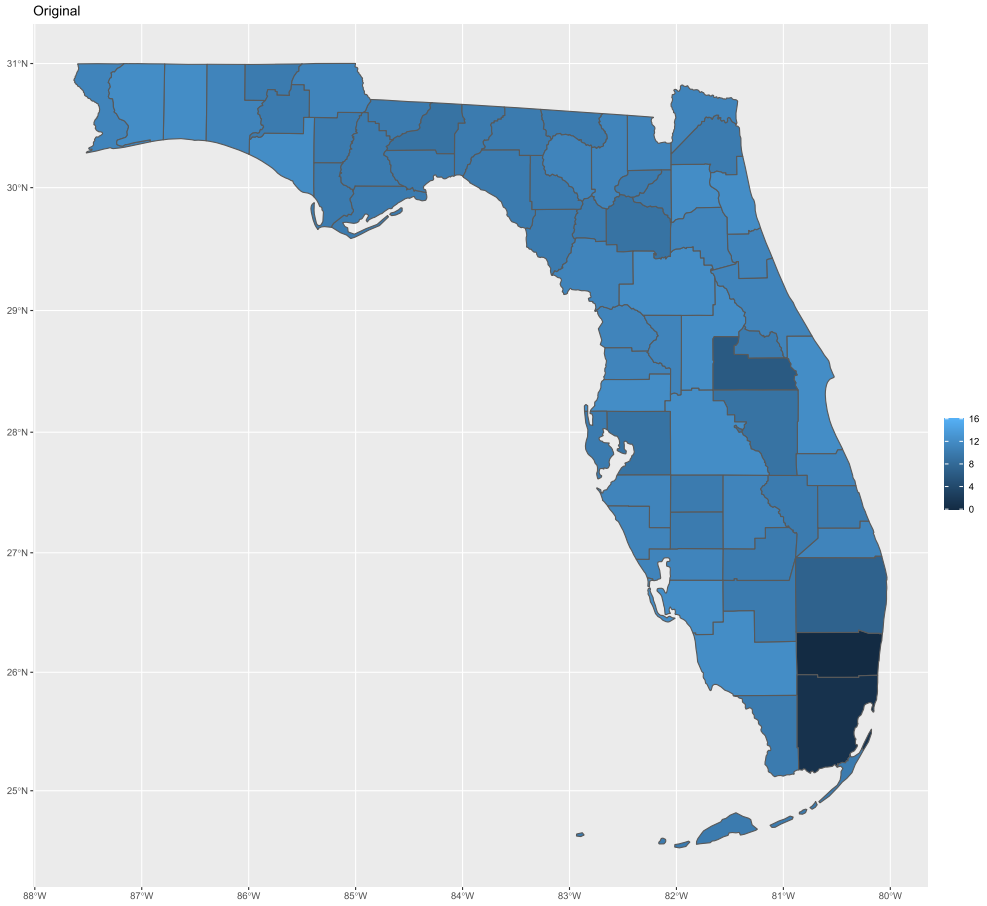

Figure 2: Transformed difference in the number of votes. See Equation (12).

Table 2: The pointwise 95% credible interval for eight counties.

County	95% Credible Interval	County	95% Credible Interval
Clay	(1.001, 1.134)	Indian River	(1.007, 1.142)
Collier	(1.007, 1.139)	Lee	(1.009, 1.146)
Franklin	(1.023, 1.159)	Putnam	(1.003, 1.137)
Holmes	(1.017, 1.158)	Wakulla	(1.001, 1.135)

We compare our model to a Poisson model with latent multivariate log gamma random effects (Bradley et al., 2018) and a Poisson model with latent Gaussian random effects (Hadfield et al., 2010). The multivariate log-gamma distribution is a type of conjugate multivariate (CM) distribution (Bradley et al., 2020). The Poisson model with latent Gaussian random effects is sometimes called a Latent Gaussian Process (LGP) (Gelfand and Schliep, 2016). The same covariates and basis functions were used in both of the competing models, and public-use code for both methods were used. We use the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) and the logarithm pseudo marginal likelihood (LPML) (Chen et al., 2008) as an overall model fit measure. These two criteria involve a tradeoff between the goodness of fit

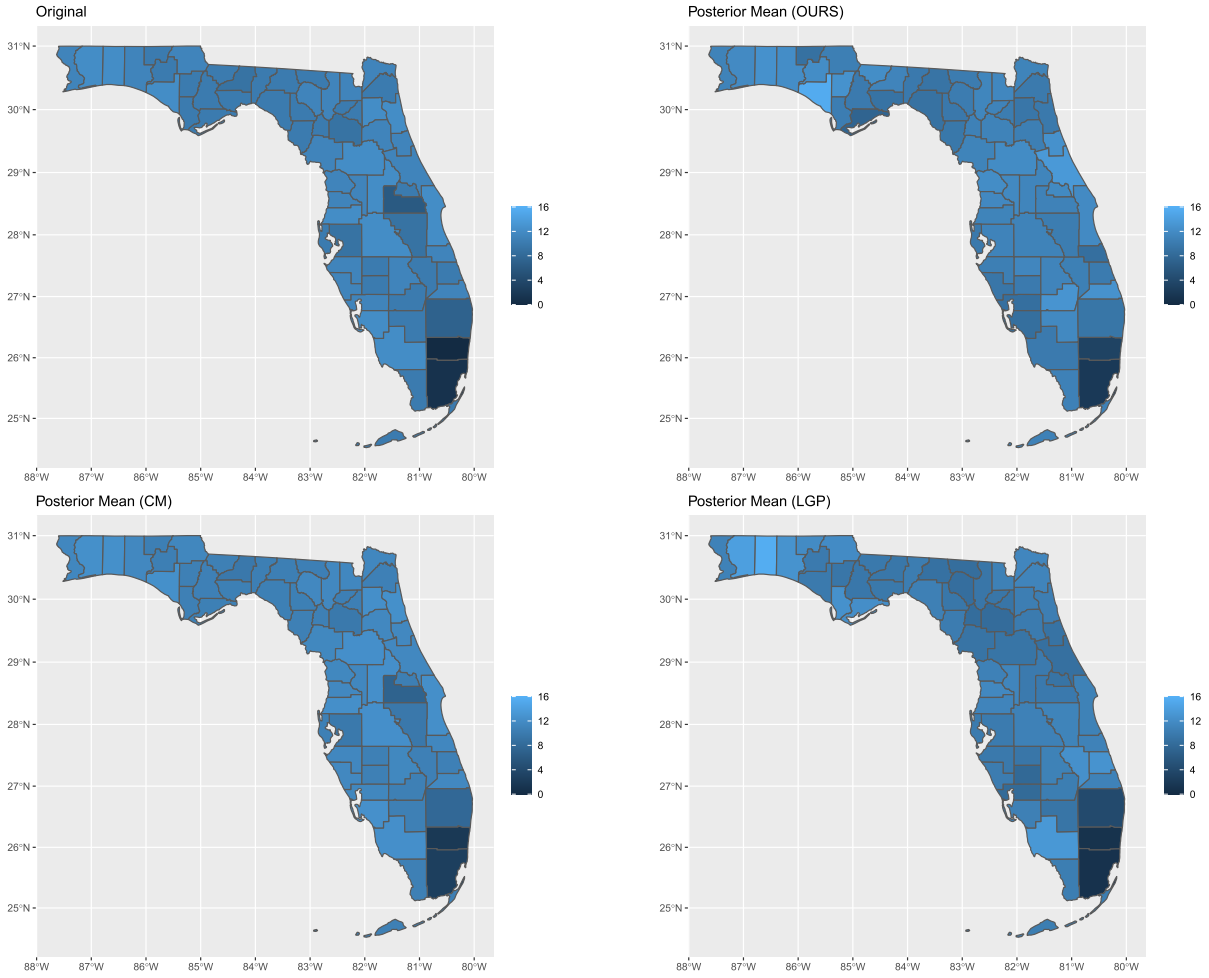

Figure 3: Left top: Original data; Right top: posterior mean of OURS; Left bottom: posterior mean of Poisson CM model; Right bottom: posterior mean of Poisson LGP model.

and model complexity. Both criteria can be easily obtained from a Markov chain Monte Carlo (MCMC) output. Here, the largest LPML value and the smallest DIC values are preferable.

In Figure 3, we plot predictions of $\mathbf{Z}$ using OUR (Algorithm 1), predictions from a Poisson CM model, and a Poisson LGP model. All predictions look similar, except the Poisson CM appears to overfit. This is verified by the value of DIC and LPML, see Table 3. Our model has the smallest DIC and largest LPML marginally. This gives motivation for our method because OURS not only gives competitive predictions, but also allows us to assess over-and-under dispersion more readily (i.e., through estimates of $\boldsymbol{\lambda}$).

For this dataset, we find that eight counties appear to be under-dispersed (i.e., the pointwise 95% credible interval of $\beta_{v,i}$ is greater than 1). The eight counties are Clay, Collier, Franklin, Holmes, Indian River, Lee, Putnam, and Wakulla counties. Considering the past four presidential elections from 2000 to 2012, these counties consistently favored the Republican party in each election. Four of these counties are rural areas and the remaining four are adjacent to rural areas.

Table 3: Comparison Results.

Model	DIC	LPML
OURS	319.58	−156.32
CM	395.95	−212.87
LGP	334.76	−161.11

5 Discussion

The COM-Poisson model has received a great deal of attention in recent years in many fields of application. This because COM-Poisson regression is a popular model for count data due to its ability to capture both under-dispersion and over-dispersion. In practice, modeling spatial count data using COM-Poisson is challenging, since spatial random effects compete with the dispersion parameter.

We have proposed a new "matrix free" Bayesian approach for modeling dependent count data. The first contribution is that we modify the prior distribution from Kadane et al. (2006) so that one can avoid computing/approximating the normalizing constant. Additionally, we impose a conditional independence assumption between the COM-Poisson parameters, and a spatially smoothed version of these parameters to avoid inflating variance of the data with spatial random effects changing the dispersion properties of the data. Thus, our approach can model both over-dispersed and under-dispersed count data (unlike negative binomial and many other count models). We refer to this model as Over or Under-dispersed Regression for Spatial (OURS) count data. Another contribution is that our approach is "matrix free" and computationally efficient. To achieve computationally efficient implementation of OURS, in the Bayesian setting, we model spatial dependency through latent random variables and two step procedure. We first sample from posterior distribution of the COM-Poisson parameters based on the modified prior distribution then use a posterior predictive distribution to model functional spatial dependence in the mean of latent processes. Consequently, there is no inversion or storage of a large matrix in our approach.

In Section 3, we present different simulation scenarios. It includes one-dimensional spatial locations with both under-dispersed and over-dispersed data and two-dimensional spatial locations for under-dispersed data. In each scenario, we find that we can accurately estimate the true location and dispersion parameters. We see that the credible interval displays the pattern of the truth in each scenario. In Section 4, we present a real data application of Florida voting data in the 2016 US presidential election. OURS produces better measures of out-of-sample error than Poisson CM model and Poisson LGP model. Furthermore, OURS allow us to assess over-and-under dispersion while the Poisson CM and Poisson LGP only allow for overdispersion.

There are several possibilities of interesting future work. For example, the COM-Poisson distribution's structure allows for a variety of generalizations such as zero-inflated data. Its appeal from a practical point of view is even stronger: it is easy to use, flexible for fitting over-dispersed and under-dispersed data, and the second step of Algorithm 1 could easily be adapted to other setting such as time series or the spatio-temporal settings.

Supplementary Material

The real data and R code needed to reproduce the results in this paper can be found on the supplementary materials.

Funding

Jonathan Bradley's research was partially supported by the US National Science Foundation (NSF) grant SES-1853099.

References

Banerjee S, Carlin BP, Gelfand AE (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

Bradley JR, Holan SH, Wikle CK (2018). Computationally efficient distribution theory for bayesian inference of high-dimensional dependent count-valued data (with discussion). *Bayesian Analysis*, 13: 253–302.

Bradley JR, Holan SH, Wikle CK (2020). Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family. *Journal of the American Statistical Association*, 115: 2037–2052.

Cameron AC, Trivedi PK (2013). *Regression Analysis of Count Data*, volume 53. Cambridge University Press.

Carlin BP, Banerjee S (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data. *Bayesian Statistics*, 7: 45–63.

Chakraborty S, Imoto T (2016). Extended Conway-Maxwell-Poisson distribution and its properties and applications. *Journal of Statistical Distributions and Applications*, 3: 5.

Chen MH, Huang L, Ibrahim JG, Kim S (2008). Bayesian variable selection and computation for generalized linear models with conjugate priors. *Bayesian Analysis*, 3: 585–614.

Conway RW, Maxwell WL (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12: 132–136.

Cressie N (1993). *Statistics for Spatial Data*. John Wiley & Sons, New York, NY.

Cressie N, Johannesson G (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 70: 209–226.

Dai F, Dutta S, Maitra R (2020). A matrix-free likelihood method for exploratory factor analysis of high-dimensional gaussian data. *Journal of Computational and Graphical Statistics*, 29: 675–680.

Daly F, Gaunt RE (2016). The Conway-Maxwell-Poisson distribution: distributional theory and approximation. arXiv preprint: https://arxiv.org/abs/1503.07012.

Duquette CM, Mixon FG, Cebula RJ (2017). Swing states, the winner-take-all electoral college, and fiscal federalism. *Atlantic Economic Journal*, 45: 45–57.

Gelfand AE, Schliep EM (2016). Spatial statistics and Gaussian processes: A beautiful marriage. *Spatial Statistics*, 18: 86–104.

Guikema SD, Goffelt JP (2008). A flexible count data regression model for risk analysis. *Risk Analysis*, 28: 213–223.

Gupta RC, Sim S, Ong S (2014). Analysis of discrete data by Conway–Maxwell Poisson distribution. *AStA Advances in Statistical Analysis*, 98: 327–343.

Hadfield JD, et al. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33: 1–22.

Hilbe JM (2011). *Negative Binomial Regression*. Cambridge University Press.

Hooten MB, Wikle CK, Dorazio RM, Royle JA (2007). Hierarchical spatio-temporal matrix models for characterizing invasions. *Biometrics*, 63: 558–567.

Kadane JB, Shmueli G, Minka TP, Borle S, Boatwright P (2006). Conjugate analysis of the Conway-Maxwell-Poisson distribution. *Bayesian Analysis*, 1: 363–374.

Lindén A, Mäntyniemi S (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, 92: 1414–1421.

Manton KG, Woodbury MA, Stallard E (1981). A variance components approach to categorical data models with heterogenous cell populations: Analysis of spatial gradients in lung cancer mortality rates in North Carolina counties. *Biometrics*, 37: 259–269.

Neal RM (2003). Slice sampling. *The Annals of Statistics*, 31: 705–767.

Sellers KF, Borle S, Shmueli G (2012). The COM-Poisson model for count data: A survey of methods and applications. *Applied Stochastic Models in Business and Industry*, 28: 104–116.

Sellers KF, Morris DS, Balakrishnan N (2016). Bivariate Conway–Maxwell–Poisson distribution: Formulation, properties, and inference. *Journal of Multivariate Analysis*, 150: 152–168.

Sellers KF, Raim A (2016). A flexible zero-inflated model to address data dispersion. *Computational Statistics & Data Analysis*, 99: 68–80.

Sellers KF, Shmueli G (2013). Data dispersion: now you see it... now you don't. *Communications in Statistics. Theory and Methods*, 42(17): 3134–3147.

Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P (2005). A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 54: 127–142.

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 64: 583–639.

Ver Hoef JM, Boveng PL (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88: 2766–2772.

Wahba G (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA.

Waller LA, Carlin BP, Xia H, Gelfand AE (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92: 607–617.

Wikle CK (2010). Low-rank representations for spatial processes. In: *Handbook of Spatial Statistics* (AE Gelfand, P Diggle, P Guttorp, M Fuentes, eds.), 114–125. CRC Press.

Wikle CK, Hooten MB (2006). Hierarchical Bayesian spatio-temporal models for population spread. In: *Hierarchical Modelling for the Environmental Sciences: Statistical Methods and Applications* (JS Clark, AE Gelfand, eds.). Ch. 8.

Wu G, Holan SH, Wikle CK (2013). Hierarchical Bayesian spatio-temporal Conway–Maxwell Poisson models with dynamic dispersion. *Journal of Agricultural, Biological, and Environmental Statistics*, 18: 335–356.

Yang HC, Bradley JR (2021). Bayesian inference for big spatial data using non-stationary spectral simulation. *Spatial Statistics*, 43: 100507.