The Annals of Applied Statistics
2015, Vol. 9, No. 1, 1–26
DOI: 10.1214/14-AOAS801

BAYESIAN BINOMIAL MIXTURE MODELS FOR ESTIMATING ABUNDANCE IN ECOLOGICAL MONITORING STUDIES[1]

BY GUOHUI WU*, SCOTT H. HOLAN†, CHARLES H. NILON†
AND CHRISTOPHER K. WIKLE†

SAS Institute Inc. and University of Missouri†*

Investigation of species abundance has become a vital component of many ecological monitoring studies. The primary objective of these studies is to understand how specific species are distributed across the study domain, as well as quantification of the sampling efficiency for detecting these species. To achieve these goals, preselected locations are sampled during scheduled visits, in which the number of species observed at each location is recorded. This results in spatially referenced replicated count data that are often unbalanced in structure and exhibit overdispersion. Motivated by the Baltimore Ecosystem Study, we propose Bayesian hierarchical binomial mixture models, including Binomial Conway–Maxwell Poisson (Bin-CMP) mixture models, that formally account for varying levels of spatial dispersion. Our proposed models also allow for variable selection of model covariates and grouping of dispersion parameters through the implementation of reversible jump Markov chain Monte Carlo methodology. Finally, using demographic covariates from the American Community Survey, we demonstrate the effectiveness of our approach through estimation of abundance for the American Robin (*Turdus migratorius*) in the Baltimore Ecosystem Study.

1. Introduction. Investigation of species abundance is a topic of widespread interest in ecology. To estimate and model variation in species abundance, predetermined survey points are visited at each sampling occasion and the number of animals detected are recorded. This results in spatially referenced point count data. Such a sampling protocol is easier to implement than the traditional capture–recapture experiment [e.g., see Williams, Nichols and Conroy (2002) and the references therein], since each animal encountered does not have to be distinctly tagged. Nevertheless, these spatially referenced data can be utilized to estimate the abundance of animals, for which individual tagging might be difficult or even infeasible due to the amount of effort involved, for example, in some avian ecology surveys. Therefore, to estimate abundance, the development of binomial mixture models

Received April 2014; revised December 2014.

[1] Supported in part by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF Grant SES-1132031, funded through the NSF-Census Research Network (NCRN) program. Additionally, this research was supported in part by NSF DMS-10-49093 and the NSF Long-Term Ecological Research Program (DEB-1027188).

Key words and phrases. American Community Survey, American Robin, Conway–Maxwell Poisson, negative binomial, overdispersion, parallel computing, unbalanced data, underdispersion.

has drawn significant attention over the past few decades [e.g., Carroll and Lombard (1985), Kéry (2008), Kéry, Royle and Schmid (2005), Royle (2004), Webster, Pollock and Simons (2008)].

In developing statistical models for count data, the choice of the distribution function frequently depends on the dispersion associated with the data. For equidispersed data (i.e., equal mean and variance), the Poisson distribution is frequently used due to its explicit assumption of equidispersion. However, to model overdsipersed data (i.e., the variance is greater than the mean), the choice of distribution functions needs to be made [e.g., see Ver Hoef and Boveng (2007)]. Often, the negative binomial (NB) distribution [Cameron and Trivedi (1998)] is employed, due to a dispersion parameter that conveniently controls the level of overdispersion. Alternatively, the Poisson distribution can also be used with a random effect included to relax the restrictive assumption of equidispersion. Although the Poisson and NB distributions have become the de facto options for count data, neither of them accounts for underdispersion (i.e., the variance is less than the mean). Admittedly, overdispersion is more common for data arising from ecological monitoring studies, while underdispersion is often present for rare event data [e.g., Herbers (1989), Oh, Washington and Nam (2006), Ridout and Besbeas (2004)]. Nevertheless, cases can arise in ecological monitoring studies where the species of interest is less prevalent (due to being rare occurrences). In principle, these situations would manifest themselves as underdispersion.

The Conway–Maxwell Poisson (CMP) distribution [Conway and Maxwell (1962)] is an ideal candidate for modeling count data with different types of dispersion, as it has an extra dispersion parameter that flexibly allows for equi-, over-, and underdispersion. Moreover, the CMP distribution is closely related to many other discrete distributions. For example, the CMP distribution contains the Poisson distribution as a special case and generalizes Bernoulli and geometric distributions in the limiting cases [Shmueli et al. (2005)]. Owing to its versatility, the CMP distribution has become increasingly popular among many subject-matter disciplines. For example, in the context of breeding bird surveys, Wu, Holan and Wikle (2013) develop a Bayesian hierarchical spatio-temporal CMP model for complex and high-dimensional count data. A unique aspect of this research is that it allows for dynamic spatial dispersion (i.e., the dispersion over the spatial domain evolves over time). A comprehensive overview regarding the CMP model is provided by Sellers, Borle and Shmueli (2012) and the references therein.

Binomial mixture models have become increasingly popular for analyzing spatial point referenced count data in the context of estimating and modeling variation in species abundance. As a result, various models have been developed with this application in mind. For example, Carroll and Lombard (1985) consider a Binomial-Beta mixture model to study the problem of estimating an unknown population, N, that follows a discrete uniform distribution, in which efficient estimators were obtained through the use of an integrated likelihood method. To improve the estimator proposed by Carroll and Lombard (1985), Royle (2004) develops a Binomial–Poisson (Bin–Pois) mixture model, in which N is considered to be an independent random variable from a Poisson distribution. Subsequently, Royle and Dorazio (2006) propose a more general hierarchical modeling framework with the goal of addressing animal abundance in the case of imperfect detection, wherein the variation associated with the observed data was partitioned into that of abundance and that of detectability. In the context of avian ecology studies, Kéry, Royle and Schmid (2005) and Kéry (2008) apply the Bin–Pois models to the estimation of bird abundance. Webster, Pollock and Simons (2008) propose a Bin–Pois model, in which a conditional autoregressive (CAR) model was used to address spatial dependence found in the bird density. Wenger and Freeman (2008) develop zero-inflated Bin–Pois and zero-inflated Binomial–negative binomial (Bin–NB) models for the estimation of species abundance. Kéry and Royle (2010) develop a Bin–Pois model with a site-specific random effect to allow for overdispersion and, thus, the equidispersion assumption of the Poisson distribution is relaxed. Graves et al. (2011) apply the Bin–Pois model to estimate abundance for a grizzly bear population using multiple detection methods, in which covariates are introduced to explain variation in both the detection and intensity process. Under the frequentist framework, Dail and Madsen (2011) propose a general Bin–Pois model to allow for a formal statistical test regarding the assumption of population closure. However, none of the aforementioned models simultaneously allows for data with different levels of dispersion (over- and underdispersion) and Bayesian model selection (e.g., using the Conway–Maxwell Poisson distribution and reversible jump Markov chain Monte Carlo).

Some experiments in ecological studies can be viewed as a robust design [e.g., see Pollock (1982)], that is, there are secondary, and possibly subsequent, sampling periods nested within each primary sampling occasion. For example, the American Robin (*Turdus migratorius*) data we consider from the Baltimore Ecosystem Study (BES) falls into this category. This nested sampling design contains the design with one primary sampling occasion as a special case. Motivated by American Robin data from BES (Section 6), we develop a Binomial Conway–Maxwell Poisson (Bin-CMP) mixture model that accommodates both overdispersed and underdispersed data under a nested/unbalanced data structure. The Bin-CMP models we propose are cast in a general Bayesian hierarchical binomial mixture model framework that can accommodate mixtures using distributions other than the CMP.

Compared with the existing models in the literature, our contribution can be seen as follows. First, we develop a flexible class of binomial mixture models to account for replicated count data with different types of dispersion, which is achieved by choosing a suitable model for the abundance parameter (e.g., using the CMP distribution). In the case of overdispersed data, our methodology is advantageous from an estimation perspective when compared to the general modeling strategy that includes a random effect to account for extra dispersion [e.g., see Kéry and Royle (2010)], as our model has a fewer number of parameters to be estimated. Although each parameter may be more computationally expensive, compared to the strategy of including a random effect, this computational burden can be alleviated through the use of a lower level programing language and parallel computation. More importantly, our model provides an explicit quantification of dispersion and can also be used in the context of underdispersed data. Additionally, the models we consider can flexibly account for spatial dependence in species abundance by adding a low-rank spatial component to the model for the intensity process. In contrast to the CAR models used by Webster, Pollock and Simons (2008), our methodology does not require us to define a neighborhood structure for the point count data, which can be difficult in many cases. In the setting of our motivating example, where the bird counts themselves are modeled at the point level rather than on areal units, a geostatistical approach may be more appropriate. Further, through reversible jump Markov chain Monte Carlo (RJMCMC), we introduce automated variable selection for covariates and grouping of dispersion parameters into the binomial mixture modeling framework and, to the best of our knowledge, our approach constitutes the first successful RJMCMC implemented on the CMP dispersion parameters. Last, the variable selection allows us to identify important predictors related to high detectability and abundance for a given species of interest.

This paper is organized as follows. Section 2 introduces our motivating data from the BES and provides preliminary background information on the CMP distribution. Section 3 describes our proposed Bayesian hierarchical binomial mixture models, including the Bin-CMP model. Section 4 provides relevant information on Bayesian variable selection and grouping using RJMCMC. Simulated examples are presented in Section 5, illustrating the effectiveness of our modeling approach. Section 6 contains an analysis of our motivating data, estimating abundance of the American Robin from the BES, and demonstrates the utility of our methodology. Discussion is provided in Section 7. For convenience of exposition, specific details surrounding our Markov chain Monte Carlo (MCMC) algorithm and full conditional distributions are left to a supplemental article [Wu et al. (2015)].

2. Data and preliminary background.

2.1. *Baltimore Ecosystem Study survey data.* As a long-term ecological monitoring study, the BES considers the City of Baltimore, Maryland as a study area, with the objective of understanding how the City of Baltimore evolves as an ecosystem over time [Pickett et al. (2011)]. Collected as a part of the BES, the American Robin (*Turdus migratorius*) data we consider constitutes spatially replicated point count data on 132 bird census points in the City of Baltimore, which are randomly selected from a set of urban forest effect (UFORE or I-Tree Eco) model points (Section 6). Considered as the most widespread North American thrush, the American Robin has become common in many North American cities [Sallabanks and James (1999)]. Despite its abundance, conservation measures, which are enforced by the Migratory Bird Treaty Act of 2004, have been taken to protect the American Robin throughout its geographical range in the United States. Although BES data have been collected across bird survey points since 2005, as an illustration, we consider a subset of data over five years from 2005 to 2009, due to incomplete data in later years. In each year, three surveys were scheduled for each of the survey points throughout May and August, each of which consisted of a five minute survey conducted between 5 am and 10 am on days without rain. During each survey, the recorded count represents the combination of birds that were seen, heard, or flew over each survey point. In the current context, the secondary sampling period consists of the five minute daily survey, while the primary sampling periods are the time frames determined by the dates on which three daily surveys are conducted. As a result, the nested sampling design provides a maximum of 15 spatially referenced counts for each bird census point. Despite the fact that several species are available in the BES, as an illustration, we consider American Robin counts in our analysis, due to their higher abundance relative to other species. Among the 132 bird census points, 131 of them have American Robin detections (Figure 1).

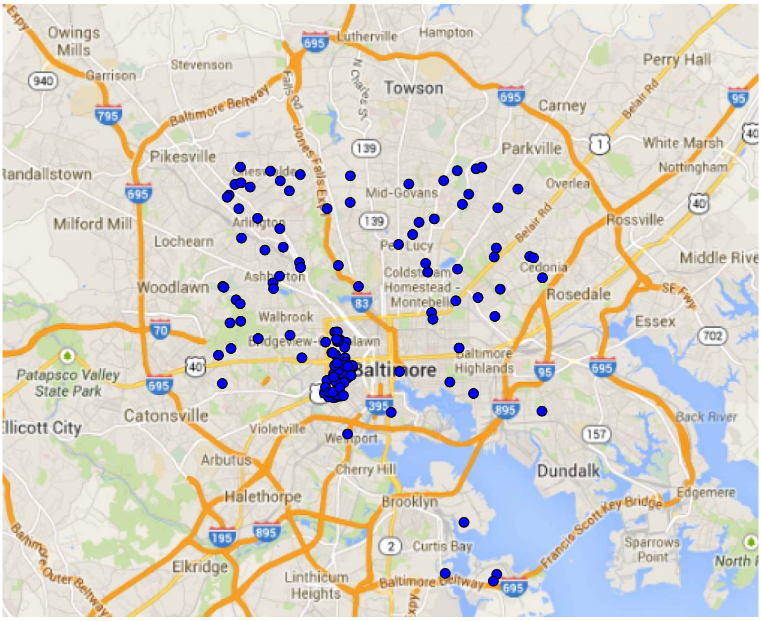


FIG. 1. *Plot of* 131 *bird census points for American Robin in the City of Baltimore*, *Maryland* (*using R package "RgoogleMaps"*). *The solid circles are bird census points*.

2.2. *The Conway–Maxwell Poisson distribution.* Let X denote a CMP distributed random variable, that is, $X \sim \text{CMP}(\lambda, \nu)$, where $\lambda > 0$ and $\nu \geq 0$ are the CMP intensity and dispersion parameters, respectively. The probability mass function (pmf) of X is given by

$$P(X = x) = \frac{\lambda^x}{(x!)^\nu} \frac{1}{Z(\lambda, \nu)}, \qquad x = 0, 1, 2, \ldots, \tag{1}$$

where

$$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu} \tag{2}$$

is a normalizing constant (often referred to as the "Z-function"). With the additional parameter ν, the CMP distribution conveniently accommodates equidispersion, overdispersion, and underdispersion. Specifically, $\nu = 1$ corresponds to the Poisson distribution, whereas $\nu < 1$ and $\nu > 1$ represent overdispersion and underdispersion, respectively. In addition, the CMP distribution generalizes to the geometric and Bernoulli distributions in the limiting cases [Shmueli et al. (2005)].

For the calculation of (1), the Z-function needs to be computed numerically due to the summation of an infinite series. For certain combinations of λ and ν, many terms will be needed in order to truncate the infinite summation with sufficient accuracy, which leads to intensive computation. For these cases, Minka et al. (2003) derived an asymptotic approximation to the Z-function, which is accurate when $\lambda > 10^\nu$. Wu, Holan and Wikle (2013) discuss further improvements on computation by taking advantage of parallel computing through Open Multiprocessing (OpenMP) and Compute Unified Device Architecture (CUDA), that is, graphics processing unit (GPU).

3. Hierarchical Binomial mixture models.

3.1. *Model development.* Let $\{\mathbf{s}_i\}_{i=1}^G, \mathbf{s}_i \in D \subset \mathbb{R}^2$ denote a set of sampling locations. We consider an experimental design in which animals are surveyed at each sampling location $\mathbf{s}_i$ for a total of J primary sampling occasions, in which there are potentially K nested secondary sampling periods. In principle, the primary sampling occasions can be over any arbitrary time interval, for example, in weeks or months. In addition, we assume a closed population within each primary sampling occasion so that the species abundance at each location varies across primary sampling occasions but not within. Relative to the primary sampling occasion, the secondary sampling period might be over a shorter time interval, for example, daily surveys within the three-month long primary sampling occasions. To allow for an unbalanced data structure, due to missing observations, we assume $n_{ij} \leq K$ successful visits to site $\mathbf{s}_i$ during the jth primary sampling period with the number of animals detected recorded. Therefore, it follows that $0 \leq n_{ij} \leq K$,

$i = 1, 2, \dots, G$; $j = 1, 2, \dots, J$. We note that "missing" values are not uncommon and can occur for many reasons. For example, some scheduled visits might not be made due to illness of the observer, and as a result no data will be recorded. In the current context, we assume that any missing data are missing completely at random (MCAR) [Little and Rubin (2002)].

For $i = 1, 2, \dots, G$, $j = 1, 2, \dots, J$, and $k = 1, 2, \dots, n_{ij}$, let y_{ijk} be the number of animals observed at location $\mathbf{s}_i$ during the kth secondary sampling within the jth primary sampling occasion. The observed data can be denoted by $\mathbf{Y} = \{\mathbf{y}_{ij} : i = 1, 2, \dots, G; j = 1, 2, \dots, J\}$, where $\mathbf{y}_{ij} = (y_{ij1}, y_{ij2}, \dots, y_{ijn_{ij}})'$ and $1 \leq n_{ij} \leq K$. Note that $n_{ij} = 0$ corresponds to the case that no successful visits are made to site i and, thus the vector $\mathbf{y}_{ij}$ does not have any elements. Further, let p_{ijk} be the probability of detecting an animal during the kth ($k = 1, 2, \dots, n_{ij}$) secondary sampling within the jth primary sampling occasion ($j = 1, 2, \dots, J$) at location $\mathbf{s}_i$ and denote N_{ij} as the unknown animal abundance at location $\mathbf{s}_i$ during the jth primary sampling occasion. In other words, N_{ij} represents the total number of animals available for sampling during the jth primary sampling occasion at location $\mathbf{s}_i$. Due to the closed population assumption, N_{ij} does not vary among secondary sampling periods within each primary sampling occasion.

The nested design we consider is more general than many of the designs previously investigated [e.g., Kéry (2008), Kéry, Royle and Schmid (2005), Royle (2004), Royle and Dorazio (2006), Royle and Link (2005), Webster, Pollock and Simons (2008)], all of which can be seen as a special case of ours by setting $K = 1$. In contrast, our study design is more similar to those found in Chandler, Royle and King (2011) and Dail and Madsen (2011). Additionally, for the sake of flexibility, it is not necessary that $n_{ij} \equiv K$ (for all $i = 1, 2, \dots, G$ and $j = 1, 2, \dots, J$). Importantly, the replicated data collected in the secondary sampling provides additional information that could alleviate potential issues caused by missing values as well as improve the accuracy of parameter estimation over the nonnested design. The primary objective of our analysis is to estimate abundance and draw inference about detectability. To achieve these goals, we propose a class of hierarchical binomial mixture models, that includes the Bin-CMP model.

The class of binomial mixture models naturally fits into the hierarchical framework [e.g., Cressie and Wikle (2011), Royle and Dorazio (2008)]. In this framework, we define the *observation model* as

$$y_{ijk} | N_{ij}, p_{ijk} \sim \mathrm{Bin}(N_{ij}, p_{ijk}), \tag{3}$$

for $i = 1, 2, \dots, G$; $j = 1, 2, \dots, J$; $k = 1, 2, \dots, n_{ij}$, where the probability p_{ijk} corresponds to the kth secondary sampling within the jth primary sampling occasion at location $\mathbf{s}_i$. For the design we consider, (3) allows us to estimate abundance parameters N_{ij}, which are both location- and time-specific. Also, since the abundance N_{ij} at each site $\mathbf{s}_i$ varies over time, we are able to describe the temporal changes in species abundance for all spatial locations, which is often vital in the context of long-term ecological monitoring studies. Another benefit of the design we consider is the potentially sharper estimates of the detection probability. Using a single probabilistically coherent model, we are able to provide spatial maps that illustrate the changes in abundance over time as well as the spatial variation [e.g., see Figures 2 and 3 in the supplementary article, Wu et al. (2015)]. More importantly, (3) also suggests how over- and underdispersion can be explicitly accounted for in the subsequent model development through the choice of an appropriate count model for abundance parameter, N_{ij}. Specifically, under the assumption of independence between N_{ij} and p_{ijk}, it follows that

$$\begin{aligned} E(y_{ijk}) &= E(p_{ijk})E(N_{ij}), \\ \operatorname{Var}(y_{ijk}) &= E(p_{ijk})E(N_{ij}) + E(p_{ijk}^2)\{\operatorname{Var}(N_{ij}) - E(N_{ij})\}. \end{aligned}$$

Hence, the mean and variance relationship in the data can be addressed through that of N_{ij}. For example, for data with over- and underdispersion, we can choose a model for N_{ij} such that $\operatorname{Var}(N_{ij}) > E(N_{ij})$ or $\operatorname{Var}(N_{ij}) < E(N_{ij})$, respectively. As such, our approach addresses over- and underdispersed count data through the choice of an appropriate model for abundance parameter, N_{ij}.

For $i = 1, 2, \dots, G$ and $j = 1, 2, \dots, J$, the *process model* we consider for the abundance, N_{ij}, is given by

$$N_{ij}|\lambda_{ij}, \nu_j \sim f(\lambda_{ij}, \nu_j), \tag{4}$$

where $f(\cdot)$ is used to generically denote an appropriate count distribution with intensity parameter λ_{ij} and primary sampling period-varying dispersion parameters ν_j. There are many possible choices for the distribution function $f(\cdot)$ in the process model (4), including the Pois, NB, and CMP, among others. We focus on the case where $f(\cdot)$ is chosen to be the CMP distribution, resulting in a flexible Bin-CMP mixture model that allows for equi-, over-, and/or underdispersion. Alternatively, if $f(\cdot)$ is chosen to be the NB distribution, the resulting Bin–NB mixture model provides a suitable candidate for modeling overdispersed data. Finally, it is important to note that, although we focus on the CMP distribution, in our framework, $f(\cdot)$ can be chosen to be any valid count distribution.

Specification of the *parameter model* is usually problem-specific and often depends on the research questions under consideration. In long-term ecological monitoring studies, it is often of interest to understand which factors might be important constituents in the probability of detection, so that an efficient sampling protocol can be designed. To achieve this goal, we relate the detection probability, p_{ijk}, to the covariates $x_{ijk,1}, \dots, x_{ijk,P}$ through a logistic link function, that is,

$$\operatorname{logit}(p_{ijk}) = \beta_1 x_{ijk,1} + \cdots + \beta_P x_{ijk,P}, \tag{5}$$

where $\operatorname{logit}(r) = \log\{r/(1 - r)\}$, $i = 1, 2, \dots, G$, $j = 1, 2, \dots, J$, and $k = 1, 2, \dots, n_{ij}$. Note that (5) allows for an intercept, by setting $x_{ijk,1} \equiv 1$ for all i, j, and k. By incorporating covariates into the model, the objective is to identify and draw statistical inference on important factors governing the probability of detection. Another interest in long-term ecological studies is to gain deeper understanding surrounding the intensity λ_{ij}, which influences species abundance. The second part of the *parameter model* defines a model for the intensity, λ_{ij}, as

$$\log \lambda_{ij} = \mathbf{w}'_{ij}\boldsymbol{\gamma} = w_{ij,1}\gamma_1 + \cdots + w_{ij,M}\gamma_M, \qquad i = 1, \ldots, G;\ j = 1, \ldots, J. \tag{6}$$

Here, $\mathbf{w}_{ij} = (w_{ij1}, \ldots, w_{ij,M})'$ are a set of covariates and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_M)'$ denotes the associated coefficients.

3.2. *Accounting for spatial dependence.* For spatially replicated count data, such as those typically encountered in monitoring studies, it is sometimes necessary to explicitly account for spatial dependence in the model for intensity. Under this scenario, we can extend (6) to explicitly incorporate spatial dependence by adding a spatial component in the model for the intensity, that is,

$$\log \lambda_{ij} = \mathbf{w}'_{ij}\boldsymbol{\gamma} + \boldsymbol{\phi}'_i\boldsymbol{\alpha}_j, \qquad i = 1, \ldots, G;\ j = 1, \ldots, J, \tag{7}$$

or

$$\log \boldsymbol{\lambda} = \mathbf{w}'\boldsymbol{\gamma} + (\boldsymbol{\Phi} \otimes \boldsymbol{\alpha}')\,\mathrm{vec}(\mathbf{I}_{\tau\times\tau}),$$

where $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{j\tau})'$; $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_J)$; $\boldsymbol{\lambda} = (\lambda_{11}, \ldots, \lambda_{1J}, \ldots, \lambda_{G1}, \ldots, \lambda_{GJ})'$; $\mathbf{w} = (\mathbf{w}_{11}, \ldots, \mathbf{w}_{1J}, \ldots, \mathbf{w}_{G1}, \ldots, \mathbf{w}_{GJ})$; $\boldsymbol{\Phi}$ denotes a $G \times \tau$ matrix of spatial basis functions $\boldsymbol{\Phi} = [\phi'_1; \ldots; \phi'_G]$; $\boldsymbol{\phi}'_i = (\phi_{i1}, \ldots, \phi_{i\tau})$ is a row vector denoting the ith row of $\boldsymbol{\Phi}$; $\mathbf{I}_{\tau\times\tau}$ is a $\tau \times \tau$ identity matrix; τ is the number of basis functions and $\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\alpha)$. There are several advantages to incorporating spatial effects when modeling the intensity function. Most importantly, capturing spatial dependence in the intensity function among neighboring locations will allow us to borrow strength from correlated observations, potentially improving parameter estimation, statistical inference, and prediction.

The choice of basis functions is typically problem specific, with advantages arising from specific choices. Popular choices include empirical orthogonal functions (EOFs), Fourier basis function, splines, wavelets, bi-square and predictive process basis [e.g., see Cressie and Johannesson (2008), Cressie and Wikle (2011), Royle and Wikle (2005) and the references therein]. In spatial statistical modeling, low-rank representations are often considered [Wikle (2010)]. Following Ruppert, Wand and Carroll (2003), we use the thin plate spline basis functions, where

$$\boldsymbol{\Phi} = \bigl[C(\mathbf{s}_i - \boldsymbol{\kappa}_l)\bigr]_{\substack{1\le i\le I\\ 1\le l\le \tau}} \quad \text{and} \quad C(\mathbf{r}) = \|\mathbf{r}\|^{2v-2}\log\|\mathbf{r}\|, \qquad v > 1,$$

where $\boldsymbol{\kappa}_l$ $(l = 1, 2, \ldots, \tau)$ denote fixed knot points in $\mathbb{R}^2$ and v is a smoothness parameter [see Holan et al. (2008) for further discussion]. Here, we choose $v = 2$

[cf. Ruppert, Wand and Carroll (2003), page 257] and assume $\operatorname{cov}(\boldsymbol{\alpha}_j) = \sigma^2_{\alpha_j}\boldsymbol{\Omega}$, where

$$\boldsymbol{\Omega} = \bigl[C(\boldsymbol{\kappa}_l - \boldsymbol{\kappa}_{l'})\bigr]_{1\le l,l'\le\tau}.$$

The selection of knot points can be facilitated through space-filling designs, as implemented in the `fields` package [Furrer, Nychka and Sain (2012)] in R [R Development Core Team (2013)]. The number of knots τ can be chosen based on computational considerations followed by sensitivity analysis. Alternatively, the number of knots can be chosen according to $\tau = \max\{20, \min(G/4, 150)\}$ [Ruppert, Wand and Carroll (2003), page 257]. Following Ruppert, Wand and Carroll (2003), we define $\boldsymbol{\Phi}^* = \boldsymbol{\Phi}\boldsymbol{\Omega}^{-1/2}$ and $\boldsymbol{\alpha}^* = \boldsymbol{\Omega}^{1/2}\boldsymbol{\alpha}$. Then, for $i = 1, 2, \dots, G$ and $j = 1, 2, \dots, J$, we can rewrite (7) as

$$\log\lambda_{ij} = \mathbf{w}'_{ij}\boldsymbol{\gamma} + \boldsymbol{\phi}_i^{*\prime}\boldsymbol{\alpha}_j^* = \mathbf{g}'_{ij}\tilde{\boldsymbol{\gamma}}_j, \tag{8}$$

where $\boldsymbol{\phi}_i^{*\prime}$ is the ith row of the matrix $\boldsymbol{\Phi}^*$ and $\operatorname{cov}(\boldsymbol{\alpha}_j^*) = \sigma^2_{\alpha_j}\mathbf{I}_{\tau\times\tau}$. Further, $\mathbf{g}'_{ij} = (\mathbf{w}'_{ij}\boldsymbol{\phi}_i^{*\prime})$ and $\tilde{\boldsymbol{\gamma}}_j = (\gamma_1, \dots, \gamma_M, \alpha^*_{j1}, \dots, \alpha^*_{j\tau})'$.

3.3. *The likelihood.* To account for spatial dependence, we require that $\boldsymbol{\alpha}_j^*$, $j = 1, 2, \dots, J$ in (8) are in the model with probability one. Since (6) and (8) are essentially of the same form, we will use the former in the subsequent discussion. We now derive the likelihood function for the model defined by (3), (4), (5), and (6). Let $\mathcal{M} = \{\mathcal{M}_{\boldsymbol{\beta}}, \mathcal{M}_{\boldsymbol{\gamma}}, \mathcal{M}_{\boldsymbol{\nu}}\}$, and $\mathcal{M}_{\boldsymbol{\beta}}, \mathcal{M}_{\boldsymbol{\gamma}}, \mathcal{M}_{\boldsymbol{\nu}}$ denote the model structures for the set of covariates $\mathbf{x}$ and $\mathbf{w}$ and the dispersion parameters $\boldsymbol{\nu} = \{\nu_1, \dots, \nu_J\}$, respectively. For example, in the case of $P = 6$, $M = 6$, $J = 5$, $\mathcal{M}_{\boldsymbol{\beta}} = \{x_1, x_3\}$ indicates that only x_1 and x_3 are included in the model for detection probability or, equivalently, $\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$; $\mathcal{M}_{\boldsymbol{\gamma}} = \{w_1, w_2\}$ indicates that only w_1 and w_2 are included in the model for intensity; $\mathcal{M}_{\boldsymbol{\nu}} = \{1, 2, \dots, J\}$ indicates that there is only one grouping for dispersion parameters, meaning $\nu_j \equiv \nu$ for $j = 1, 2, \dots, J$. Under the assumption of conditional independence, the likelihood function for the binomial mixture models we propose is given by

$$\mathcal{L}(\mathbf{Y}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}) = \prod_{i=1}^{G}\prod_{j=1}^{J}\prod_{k=1}^{n_{ij}}[y_{ijk}|N_{ij}, \boldsymbol{\beta}, \mathcal{M}_{\boldsymbol{\beta}}][N_{ij}|\mathcal{M}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \mathcal{M}_{\boldsymbol{\nu}}, \nu_j], \tag{9}$$

where, generically, $[\xi|\boldsymbol{\theta}]$ denotes the conditional distribution of ξ given the parameters $\boldsymbol{\theta}$. Integrating out N_{ij} in (9) yields the marginal distribution of observing $\mathbf{y}_{ij}$ as

$$\begin{aligned} &P(\mathbf{y}_{ij}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \nu_j) \\ &\quad = \sum_{N_{ij}\ge y_{ij}^{\max}}^{\infty}\left\{\prod_{k=1}^{n_{ij}}\frac{N_{ij}!}{y_{ijk}!(N_{ij}-y_{ijk})!}p_{ijk}^{y_{ijk}}(1-p_{ijk})^{N_{ij}-y_{ijk}}\right\} \\ &\qquad \times f(N_{ij}|\mathcal{M}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \mathcal{M}_{\boldsymbol{\nu}}, \upsilon_j), \end{aligned} \tag{10}$$

where $y_{ij}^{\max} = \max\{\mathbf{y}_{ij}\}$. Consequently, we can derive the joint posterior distribution function $\pi(\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}|\mathbf{Y})$ based on (10) as

$$\begin{aligned}
&\pi(\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}|\mathbf{Y}) \\
&\quad\propto \left\{\prod_{i=1}^{G}\prod_{j=1}^{J} P(\mathbf{y}_{ij}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \nu_j)\right\} \\
&\qquad \times [\boldsymbol{\beta}|\mathcal{M}_{\boldsymbol{\beta}}][\boldsymbol{\gamma}|\mathcal{M}_{\boldsymbol{\gamma}}][\boldsymbol{\nu}|\mathcal{M}_{\boldsymbol{\nu}}][\mathcal{M}_{\boldsymbol{\beta}}][\mathcal{M}_{\boldsymbol{\gamma}}][\mathcal{M}_{\boldsymbol{\nu}}].
\end{aligned} \tag{11}$$

Here $[\boldsymbol{\theta}]$ denotes the joint prior distribution function of the parameters $\boldsymbol{\theta}$.

Examination of (11) raises several computational concerns. First, the calculation of $P(\mathbf{y}_{ij}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \nu_j)$ can be computationally prohibitive, since a multiple integral is involved. This computational issue becomes exacerbated when the domain of N_{ij} covers a wide range of values and/or if G and J are large. In addition to calculating a multiple integral, in the case where $f(\cdot)$ denotes the CMP distribution, evaluating (10) requires computing the Z-function, which involves the summation of infinite series. Specifically, for the Bin-CMP model, it is worth pointing out that within each MCMC iteration, sampling elements in $\boldsymbol{\gamma}$ or $\boldsymbol{\nu}$ from their full conditionals requires both the computation of the multiple integral and the approximation of the Z-function. Therefore, implementation of our proposed model can be computationally intensive in some cases. We resolve these computational issues through the use of low level programming in C and parallel computing with OpenMP.

Finally, we assume the following prior distributions for the model parameters: $\boldsymbol{\beta} \sim \text{Gau}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$; $\boldsymbol{\gamma} \sim \text{Gau}(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma)$. For the dispersion parameters, we assume $\nu_j \sim \text{Unif}(a_j, b_j)$, $j = 1, 2, \dots, J$, where a_j and b_j are chosen appropriately to allow for different levels of dispersion in the data (e.g., for overdispersed data, one may set $a_j \equiv 0.02$ and $b_j \equiv 1.0$). In our case, we assign vague prior distributions that are noninformative relative to scale of the data.

4. Automated Bayesian model selection.

For the binomial mixture models we propose, there are several ecological objectives. First, there is a clear need to identify important covariates among a set of candidate covariates in order to gain an understanding of the factors affecting the detectability for a given species of interest. In addition, the selection of influential covariates is vital for studying which factors influence species abundance. Last, the grouping of dispersion parameters will provide us with further information about the level of dispersion associated with the data across different years in the study. In such cases, grouping is desired since some years may exhibit a similar level of dispersion due to environmental changes or other exogenous factors. For example, in our setting, specific neighborhoods may experience slow growth in terms of the number of buildings established and/or certain climate conditions may be more (or less) similar from year to year.

Thus, it is conceivable that some years may experience a similar dispersion parameter. As such, we allow for data-driven grouping of the dispersion parameters. To achieve these goals, we first discuss variable selection and grouping in the context of the models we propose.

4.1. *Bayesian variable selection and grouping.* The literature on Bayesian variable selection is fairly extensive [e.g., see Hooten and Hobbs (2015), O'Hara and Sillanpää (2009) for a comprehensive review]. Among the many available choices, the two most commonly used techniques are stochastic search variable selection [George and McCulloch (1993, 1997)] and reversible jump MCMC (RJMCMC) [Green (1995)]. For grouping, however, RJMCMC is typically considered more appropriate and, thus, we utilize it for both model selection and grouping. Although one could consider model selection through various model selection criteria [e.g., Deviance Information Criterion—Spiegelhalter et al. (2002)], this would be less advantageous when the goal is both simultaneous variable selection and grouping.

For convenience of exposition, we explain our algorithm in the context of the Bin-CMP model and note that the migration to other binomial mixture models is analogous. The implementation of variable selection for $\mathbf{x}$ and $\mathbf{w}$ involves two types of moves: BIRTH (B) and DEATH (D) defined as follows:

B: propose to add a covariate (x_m or w_m) to the current model with probability p_m^b,

D: propose to remove a covariate (x_m or w_m) from the current model with probability p_m^d.

As an example, we consider a D move for $\mathbf{x}$. In general, only a subset of covariates are subject to variable selection, while others are forced to remain in the model with probability one. For notational simplification, let $\mathbf{A}_x$ denote the set of indices corresponding to covariates $\mathbf{x}$ that are available for variable selection. For example, if there are three covariates x_1, x_2, and x_3 available and only x_1 and x_3 are subject to variable selection (i.e., x_2 is in the model with the probability 1), then we have $\mathbf{A}_x = \{1, 3\}$. Moreover, let $|\mathbf{A}_x|$ denote the cardinality of the set $\mathbf{A}_x$. For each covariate in $\mathbf{A}_x$, we assume an equal probability of a B or D move, that is,

$$p_m^b = p_m^d = 1/2 \qquad \text{for } m \in \mathbf{A}_x.$$

Suppose at the current iteration t, the model structure is given by $\mathcal{M}^t = \{\mathcal{M}_{\boldsymbol{\beta}}^t, \mathcal{M}_{\boldsymbol{\gamma}}^t, \mathcal{M}_{\boldsymbol{\nu}}^t\}$. The RJMCMC algorithm for variable selection on $\mathbf{x}$ can be outlined as follows:

Step 1: Start with the model structure $\mathcal{M}^t = \{\mathcal{M}_{\boldsymbol{\beta}}^t, \mathcal{M}_{\boldsymbol{\gamma}}^t, \mathcal{M}_{\boldsymbol{\nu}}^t\}$, where $\mathcal{M}_{\boldsymbol{\beta}}^t = \{x_{i_1}, \dots, x_{i_m}\}$ with $\boldsymbol{\beta}^t = \{\beta_{i_1}, \dots, \beta_{i_m}\}$.

Step 2: Randomly draw an index from $\mathbf{A}_x$ with an equal probability $1/|\mathbf{A}_x|$. Assume $i_s \in \mathbf{A}_x$ is chosen:

- if $i_s \in \mathcal{M}^t_{\boldsymbol{\beta}}$, then propose a D move and obtain $\mathcal{M}'_{\boldsymbol{\beta}} = \mathcal{M}^t_{\boldsymbol{\beta}} \setminus \{x_{i_s}\}$ and $\mathcal{M}' = \{\mathcal{M}'_{\boldsymbol{\beta}}, \mathcal{M}^t_{\boldsymbol{\gamma}}, \mathcal{M}^t_{\boldsymbol{\nu}}\}$ and $\boldsymbol{\beta}' = \{\beta_{i_1}, \dots, \beta_{i_s} = 0, \dots, \beta_{i_m}\}$;
- otherwise propose a B move and obtain $\mathcal{M}'_{\boldsymbol{\beta}} = \mathcal{M}^t_{\boldsymbol{\beta}} \cup \{x_{i_s}\}$ and $\mathcal{M}' = \{\mathcal{M}'_{\boldsymbol{\beta}}, \mathcal{M}^t_{\boldsymbol{\gamma}}, \mathcal{M}^t_{\boldsymbol{\nu}}\}$ and $\boldsymbol{\beta}' = \{\beta_{i_1}, \dots, \beta_{i_m}, \beta_{i_s}\}$.

Step 3: Adjust the coefficient β_{is} corresponding to the covariate x_{i_s}:

- if a D move, set $\beta_{is} = 0$;
- otherwise generate $\beta_{is} \sim q(\cdot)$.

Step 4: Generate $u \sim \text{Unif}(0, 1)$:

- if $u < \min\{1, \text{BF}(\mathcal{M}'_{\boldsymbol{\beta}}, \mathcal{M}^t_{\boldsymbol{\beta}}) \times R\}$, then set $\mathcal{M}^{t+1}_{\boldsymbol{\beta}} = \mathcal{M}'_{\boldsymbol{\beta}}$ and $\mathcal{M}^{t+1} = \mathcal{M}'$;
- otherwise $\mathcal{M}^{t+1}_{\boldsymbol{\beta}} = \mathcal{M}^t_{\boldsymbol{\beta}}$ and $\mathcal{M}^{t+1} = \mathcal{M}^t$.

Step 5: Repeat.

In terms of the proposal distribution $q(\cdot)$, we used a Gau(0, ζ) distribution with ζ being a user-defined tuning parameter. Moreover,

$$R = \begin{cases} \dfrac{p^b_{i_s}}{p^d_{i_s}} \times q(\beta_{is}), & \text{if D move,} \\[2ex] \dfrac{p^d_{i_s}}{p^b_{i_s}} \times \dfrac{1}{q(\beta_{is})}, & \text{if B move,} \end{cases}$$

and

$$\text{BF}(\mathcal{M}'_{\boldsymbol{\beta}}, \mathcal{M}^t_{\boldsymbol{\beta}}) = \frac{P(\mathcal{M}'_{\boldsymbol{\beta}}, \boldsymbol{\beta}' | \mathbf{Y}, \mathcal{M}^t_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \mathcal{M}^t_{\boldsymbol{\nu}}, \boldsymbol{\nu})}{P(\mathcal{M}^t_{\boldsymbol{\beta}}, \boldsymbol{\beta}^t | \mathbf{Y}, \mathcal{M}^t_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \mathcal{M}^t_{\boldsymbol{\nu}}, \boldsymbol{\nu})}.$$

We now discuss the grouping algorithm for the dispersion parameters $\boldsymbol{\nu}$. Assume there are n_t different arrangements $T_1, T_2, \dots, T_{n_t}$ for $\boldsymbol{\nu}$ at the tth iteration of the MCMC, that is, $\mathcal{M}^t_{\boldsymbol{\nu}} = \{T_1, T_2, \dots, T_m, \dots, T_{n_t}\}$. For each grouping T_m, $m = 1, 2, \dots, n_t$, the corresponding elements are subscripts for the dispersion parameter group membership. For example, if $n_t = 1$, we have $T_1 = \{1, 2, \dots, J\}$, that is, $\nu_j \equiv \nu$, for $j = 1, 2, \dots, J$. Similar to the variable selection previously described, we allow for two types of moves as follows:

C: propose to combine two different arrangements into one arrangement with p_c,
S: propose to split the arrangement into two arrangements with probability p_s.

Without loss of generality, assume an equal probability of proposing a C or S move, that is, $p_c = p_s = 1/2$. As an illustration, we describe only the S move. Suppose there are n^s_t out of n_t arrangements in $\mathcal{M}^t_{\boldsymbol{\nu}}$ that have more than one single element. We randomly choose each of these n^s_t arrangements with an equal probability.

Assume that group T_m is chosen, where $m \in \{1, \dots, n_t^s\}$ and $|T_m| > 1$. Assuming we split T_m into two nonempty sets T_{m_1} and T_{m_2}, we denote the resulting model structure as $\mathcal{M}'_{\boldsymbol{\nu}} = \{T_1, T_2, \dots, T_{m_1}, T_{m_2}, \dots, T_{n_t}\}$. The RJMCMC algorithm for grouping of $\boldsymbol{\nu}$ can be outlined as follows:

Step 1: Calculate the probability $P(\mathcal{M}'_{\boldsymbol{\nu}}|\mathcal{M}_{\boldsymbol{\nu}})$ and $P(\mathcal{M}_{\boldsymbol{\nu}}|\mathcal{M}'_{\boldsymbol{\nu}})$ as

$$P(\mathcal{M}'_{\boldsymbol{\nu}}|\mathcal{M}_{\boldsymbol{\nu}}) = \frac{1}{2}\frac{1}{n_t^s}\frac{1}{2^{(|T_m|-1)} - 1},$$

$$P(\mathcal{M}_{\boldsymbol{\nu}}|\mathcal{M}'_{\boldsymbol{\nu}}) = \frac{1}{2}\frac{1}{\binom{n_t^s+1}{2}}$$

[King and Brooks (2002)].

Step 2: Let ν_m denote the value common to all dispersion parameters in T_m and ν_{m_1} and ν_{m_2} be the values of dispersion parameters in T_{m_1} and T_{m_2}, respectively. Define a bijective mapping between ν_m and ν_{m_1}, ν_{m_2} as

$$\nu_{m_1} = \nu_m + \varepsilon \quad \text{and} \quad \nu_{m_2} = \nu_m - \varepsilon,$$

where $\varepsilon \sim h(\cdot)$.

Step 3: Generate $\xi \sim \text{Unif}(0, 1)$:

- if $\xi < \min\{1, \text{BF}(\mathcal{M}'_{\boldsymbol{\nu}}, \mathcal{M}^t_{\boldsymbol{\nu}}) \times R_s\}$, then set $\mathcal{M}^{t+1}_{\boldsymbol{\nu}} = \mathcal{M}'_{\boldsymbol{\nu}}$ and $\mathcal{M}^{t+1} = \mathcal{M}'$;
- otherwise $\mathcal{M}^{t+1}_{\boldsymbol{\nu}} = \mathcal{M}^t_{\boldsymbol{\nu}}$ and $\mathcal{M}^{t+1} = \mathcal{M}^t$.

In terms of the proposal distribution $h(\cdot)$, we used $h(\eta) = \text{Unif}(-\eta, \eta)$ where η is chosen through pilot tuning. Moreover,

$$\text{BF}(\mathcal{M}'_{\boldsymbol{\nu}}, \mathcal{M}^t_{\boldsymbol{\nu}}) = \frac{P(\mathcal{M}'_{\boldsymbol{\nu}}, \nu_{m_1}, \nu_{m_2}|\mathbf{Y}, \mathcal{M}^t_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \mathcal{M}^t_{\boldsymbol{\beta}}, \boldsymbol{\beta}^t)}{P(\mathcal{M}^t_{\boldsymbol{\nu}}, \nu_m|\mathbf{Y}, \mathcal{M}^t_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \mathcal{M}^t_{\boldsymbol{\beta}}, \boldsymbol{\beta}^t)},$$

$$R_s = \frac{P(\mathcal{M}_{\boldsymbol{\nu}}|\mathcal{M}'_{\boldsymbol{\nu}})}{P(\mathcal{M}'_{\boldsymbol{\nu}}|\mathcal{M}_{\boldsymbol{\nu}})} \times \frac{1}{h(\varepsilon)} \times \left|\frac{\partial(\nu_{m_1}, \nu_{m_2})}{\partial(\nu_m, \varepsilon)}\right|.$$

5. Simulated examples.

To evaluate the performance of the binomial mixture models we propose, we considered two simulated examples using the Bin-CMP model, the difference of which only resides in whether or not a spatial component is included in the intensity model. For both simulations, we choose $G = 131$, $J = 5$, and $K = 3$ to be the same as the American Robin data presented in Section 6. For both examples, we simulate data as $y_{ijk}|N_{ij}, p_{ijk} \sim \text{Bin}(N_{ij}, p_{ijk})$. For the probability of detection, we consider

$$\text{logit}(p_{ijk}) = \beta_1 x_{ijk,1} + \beta_2 x_{ijk,2} + \cdots + \beta_P x_{ijk,P},$$

where the values for the covariates $\mathbf{x}$ are set to be the same as in the American Robin data for $i = 1, 2, \dots, G$, $j = 1, 2, \dots, J$, $k = 1, 2, \dots, K$, $l = 1, 2, \dots,$

$P = 4$. In addition, we set $\boldsymbol{\beta} = (-2.31, -0.4, 0.0, -0.4)'$ with $\{x_1, x_2, x_4\}$ being important covariates. For the true abundance parameters N_{ij}, we simulated from $N_{ij} \sim \text{CMP}(\lambda_{ij}, \nu_j)$, with $\nu_1 = \nu_3 = \nu_5 = 0.15$, $\nu_2 = \nu_4 = 0.06$ and $\boldsymbol{\gamma}_0 = (0.31, 0.13, 0.44, 0.16, 0.35)'$, as estimated from the American Robin data presented in Section 6. For $i = 1, 2, \dots, G$ and $j = 1, 2, \dots, J$, the intensity λ_{ij} is simulated according to

$$\textbf{S1}: \quad \log \lambda_{ij} = \mathbf{w}_i' \boldsymbol{\gamma} + \gamma_{0j},$$

$$\textbf{S2}: \quad \log \lambda_{ij} = \mathbf{w}_i' \boldsymbol{\gamma} + \boldsymbol{\phi}_i^{*\prime} \boldsymbol{\alpha} + \gamma_{0j},$$

where $\boldsymbol{\phi}_i^{*\prime}$ for $i = 1, 2, \dots, G$ and $\boldsymbol{\gamma}_0 = (\gamma_{01}, \dots, \gamma_{05})'$ are determined according to the American Robin data with $\tau = 10$. In each of the two models, $\mathbf{w}_i$ are set to be the same as in the American Robin data presented in Section 6. Further, we set $M = 11$ and $\boldsymbol{\gamma} = (0.0, 0.0, 0.0, 0.0, 0.0, 0.06, 0.0, 0.0, 0.0, 0.03, 0.0)'$, that is, with $\{w_6, w_{10}\}$ being important covariates. Particularly, for **S2**, the coefficients of spatial components, $\boldsymbol{\alpha}$, are randomly sampled from Unif(0, 1) to avoid y_{ijk} being too large. For the two simulations, we apply RJMCMC to perform variable selection and grouping. Similar to the analysis presented in Section 6, we require $\boldsymbol{\alpha}$ to be included in the model with probability one for **S2** and set $a_j \equiv 0.02$ and $b_j \equiv 2.0$ to allow for both over- and underdispersion. In addition, we set $\boldsymbol{\mu}_\beta = \boldsymbol{\mu}_\gamma \equiv \mathbf{0}$, $\boldsymbol{\Sigma}_\beta = 10^2 \mathbf{I}_P$, and $\boldsymbol{\Sigma}_\gamma = 10^2 \mathbf{I}_M$.

Table 1 provides the posterior marginal probabilities for the most probable model for $\mathbf{x}$, $\mathbf{w}$, and $\boldsymbol{\nu}$ in the Bin-CMP models **S1** and **S2**. For model **S1**, the most frequent detection probability model was given by $\{x_1, x_2, x_4\}$ and appeared with a frequency of 99.73%. The most frequent intensity model was defined by $\{w_6, w_{10}\}$ and had a frequency of 89.92%. In addition, the most frequent grouping for dispersion parameters is $\mathcal{M}_{\boldsymbol{\nu}} = \{\{2, 4\}, \{1, 3, 5\}\}$, which appeared with a frequency of 72.51%. In all cases, the RJMCMC correctly identified the set of important covariates as well as grouping for dispersion parameters with the posterior marginal probability greater than or equal to 72.51%. In terms of parameter estimation, in most cases the 95% credible intervals (CIs), averaged over the different models, contain the true values—providing further indication that the correct model is selected with high probability. For model **S2**, the most frequent set of covariates for the detection probability model was given by $\{x_1, x_2, x_4\}$ and appeared with a frequency of 99.57%. The most frequent set of covariates $\{w_6, w_{10}\}$ for the intensity model had a frequency of 93.57%. In addition, the most frequent grouping for the dispersion parameters is $\mathcal{M}_{\boldsymbol{\nu}} = \{\{2, 4\}, \{1, 3, 5\}\}$, which appeared with a frequency of 76.00%.

In summary, the two simulations suggest that we are able to correctly identify important covariates and grouping for dispersion parameters with high posterior probability. Finally, for the estimation of abundance in the two simulations, our approach performs satisfactorily, as measured by coverage of the 95% CIs. In the presence of spatial components, however, we note that the model averaged estimates of dispersion parameters can be adversely affected by missing data.

TABLE 1

Posterior marginal probabilities of the most probable model for $\mathbf{x}$, $\mathbf{w}$, *and* $\boldsymbol{v}$ *in the Bin-CMP mixture models* **S1** *and* **S2** *simulated examples* (*Section* 5) *using RJMCMC. Note that* **S1** *contains only the covariates in the intensity model, whereas* **S2** *contains both covariates and spatial components in the intensity model and that the posterior probability for* $\mathbf{x}$ *under both* **S1** *and* **S2** *are slightly less than* 1.00 *and become* 1.00 *as a result of rounding*

(a) Variable selection and grouping for **S1**

Para- meter	**Model**	**Frequency**	**Posterior probability**
$\mathbf{x}$	$\{x_1, x_2, x_4\}$	59,838	1.00
$\mathbf{w}$	$\{w_6, w_{10}\}$	53,951	0.90
	$\{w_2, w_6, w_{10}\}$	4386	0.07
$\boldsymbol{v}$	$T_1 = \{2, 4\}, T_2 = \{1, 3, 5\}$	43,507	0.73
	$T_1 = \{1, 3\}, T_2 = \{2, 4\}, T_3 = \{5\}$	7801	0.13
	$T_1 = \{1\}, T_2 = \{2, 4\}, T_3 = \{3, 5\}$	3918	0.07

(b) Variable selection and grouping for **S2**

Para- meter	**Model**	**Frequency**	**Posterior probability**
$\mathbf{x}$	$\{x_1, x_2, x_4\}$	59,741	1.00
$\mathbf{w}$	$\{w_6, w_{10}\}$	56,139	0.94
$\boldsymbol{v}$	$T_1 = \{2, 4\}, T_2 = \{1, 3, 5\}$	37,071	0.76
	$T_1 = \{3\}, T_2 = \{2, 4\}, T_2 = \{1, 5\}$	7573	0.13

6. Application: The Baltimore Ecosystem Study.

In the urban ecosystems literature, bird communities are often used as surrogates for studying urban biodiversity or species responses to urbanization [Aronson et al. (2014), Shochat, Lerman and Fernández-Juricic (2010)]. Within urban areas the bird community is shaped by local-scale features such as habitat features that vary among neighborhoods, landscape pattern, and socioeconomic characteristics of residents that may influence land management decisions [Pickett et al. (2012)]. The American Community Survey (ACS) is an ongoing survey that is able to provide timely economic, social, and demographic information on small geographies such as census tracts. Thus, to examine the effects of certain demographic characteristics on abundance, we consider several ACS variables. Additionally, environmental features of different neighborhoods can be described by many factors, such as vegetation diversity and are, therefore, also considered in our analysis.

Substantial research has been undertaken to investigate how socioeconomic status and environmental variables influence the abundance and diversity of various avian species [see Denison (2010), Loss, Ruiz and Brawn (2009), Smallbone, Luck and Wassens (2011) and the references therein]. Using socioeconomic variables from the decennial census in 2000 associated with each census tract block groups as covariates, Denison (2010) considered a simple NB regression with no spatial components under the frequentist paradigm to estimate the relative abundance for European starling in the City of Baltimore, Maryland using a portion of data collected from 2005 to 2007. In contrast, we consider American Robin data from the BES collected from 2005 to 2009 and apply various Bin-CMP models in order to select important covariates for estimating the detection probability and abundance of the American Robin, as well as to identify the grouping of dispersion parameters. Due to missing values, the data we consider has an unbalanced structure. In particular, the percentage of secondary sampling occasions with at least one missing observation for each of five primary sampling occasions is 6.87%, 6.87%, 3.05%, 77.1%, and 50.38%, respectively. Moreover, the overall percentage of missing observations in the American Robin data set is 9.62%.

For the American Robin data, a total of 131 bird survey points were visited during three secondary daily surveys within each of the five primary sampling occasions from 2005 to 2009. With three covariates available, we considered a full model for the detection probability as

$$\operatorname{logit}(p_{ijk}) = \beta_1 + \beta_2 \texttt{time}_{ijk} + \beta_3 \texttt{airtemp}_{ijk} + \beta_4 \texttt{cloudcover}_{ijk}, \tag{12}$$

for $i = 1, \ldots, 131$, $j = 1, \ldots, 5$, and $k = 1, \ldots, n_{ij} \le K = 3$. Regarding the covariates in (12), `time`, `airtemp`, and `cloudcover` correspond to the start time, air temperature, and cloud cover (i.e., the fraction of the sky obscured by clouds) recorded on each visit to the bird survey points, respectively.

In terms of full models for the intensity, we considered the following three models:

$$\begin{aligned}
&\textbf{M1}: \quad \log\lambda_{ij} = \mathbf{w}_i'\boldsymbol{\gamma} + \widetilde{\boldsymbol{\phi}}_i^{*\prime}\boldsymbol{\alpha} + \gamma_{0j},\\
&\textbf{M2}: \quad \log\lambda_{ij} = \mathbf{w}_i'\boldsymbol{\gamma} + \gamma_{0j},\\
&\textbf{M3}: \quad \log\lambda_{ij} = \boldsymbol{\phi}_i^{*\prime}\boldsymbol{\alpha} + \gamma_{0j},
\end{aligned}$$

where, for $j = 1, \ldots, J$, γ_{0j} is a year-specific intercept and $\boldsymbol{\phi}_i^{*\prime}$ is the ith row of the matrix $\boldsymbol{\Phi}^*$ as discussed in Section 3. Moreover, the covariates in the intensity model are given by $\mathbf{w}_i' = (\texttt{uftree}_i, \texttt{ufbldg}_i, \texttt{ufmgrass}_i, \texttt{bld200m}_i, \texttt{for200m}_i, \texttt{veg200m}_i, \texttt{African}_i, \texttt{bachelor}_i, \texttt{fmkds}_i, \texttt{pubassit}_i, \texttt{houseyr}_i)$. These covariates are specific to each survey location and do not vary with primary sampling occasions. Among these environmental variables, `uftree`, `ufbldg`, and `ufmgrass` are the UFORE plots variables that indicate tree cover, ground cover by buildings and maintained grass, respectively. Further, `bld200m`, `for200m`, and `veg200m` are variables that measure tree cover, other vegetation cover, and cover by buildings in the 200 meter radius plot, respectively [see Figure 1 in the supplemental article, Wu et al. (2015)]. For the ACS variables specific to each census tract block group, `African` is the percentage of African American residents; `bachelor` is the percentage of population with Bachelor's degree or higher; `fmkds` is the percentage of housing units occupied by female householder and children under 18 years; `pubassit` is the percentage of households on government public income assistance; `hourseyr` is the median year that a housing unit was built. We used the five-year period estimates from 2005 to 2009 for these ACS variables, which can be obtained at the U.S. Census Bureau website (http://www.census.gov/acs/www/). Our specific choice of ACS variables was facilitated by a social areas analysis approach [Denison (2010), Maloney and Auffrey (2013), Müller et al. (2013)]. Note that we standardize the covariates in (12) and in the intensity model for numerical stability. Further, based on exploratory analysis involving various collinearity diagnostics (e.g., condition number, etc.) of the site covariates (not shown) and subject matter knowledge, we expect any effects of collinearity between the site covariates to have a minimal affect on the variable selection algorithm. Finally, for model **M1**, we orthogonalize the matrix of spatial basis function with respect to covariates, to alleviate potential confounding with the covariate effects [Hodges and Reich (2010)]. As a result, $\widetilde{\boldsymbol{\phi}}_i^{*\prime}$ is the ith row of the matrix of $\widetilde{\boldsymbol{\Phi}}^*$ after the orthogonalization.

It is worth pointing out that the choice of models above depends on the goal of the ecological study. For example, **M3** can be used if no covariates are available for modeling the intensity. For other cases where covariates are available, but there is no spatial dependence (or the spatial dependence is negligible after accounting for covariates), model **M2** can be utilized. Given both covariates are available and spatial dependence is present, **M1** represents a potential model.

When implementing the RJMCMC algorithm, we require the "intercept" term β_1 in (12) and $\boldsymbol{\gamma}_0$, in the model for intensity, to be included with probability one. In addition, in the presence of spatial components, we require $\boldsymbol{\alpha}$ to be in the model for the intensity with probability one. For the choice of knot points, when using low-rank thin plate basis functions, we considered a sensitivity analysis to choose the number of knots and a space-filling design for placement. Specifically, for three different choices of the number of knot points, $\tau = 10$, 15, and 32 in **M1**, similar results are obtained in terms of abundance estimation, although parameter estimation becomes more difficult as τ gets large. Equally important, the results of a sensitivity analysis indicate that the variable selection and grouping for the dispersion parameters seem robust to a different number of knot points. Hence, we choose $\tau = 10$ for both **M1** and **M3**. We used a Metropolis–Hastings within Gibbs sampler consisting of a total of 120,000 MCMC iterations, with the first 60,000 discarded as burn-in. Our inference is based on every third sample after burn-in, which results in a total of 20,000 samples used.

In terms of posterior marginal probability, the model having `time` and `cloudcover` has the highest probability of being selected in the model for detectability. Similarly, for the intensity model, `ufbldg`, `veg200m`, and `pubassit` are selected with higher probability relative to other covariates. However, the grouping of dispersion parameters varies across models depending on whether spatial components are included. This is not unexpected, as there is a trade-off between the dispersion parameter and inclusion of spatial components. The three models we considered all produce similar results in terms of the selection of important covariates and abundance estimates (results not shown). However, since the goal of our analysis is to identify and draw inference on important covariates relating to detectability and abundance, we present results from the more parsimonious model **M2**. From Table 2, it can be seen that `time` and `cloudcover` are identi-

TABLE 2

Posterior probabilities of the most probable model for **M2** *and the posterior summary statistics in the Bin-CMP model assuming the posterior mode model for* **M2**. *Note that* **M2** *only contains covariates in the intensity model, and* $\widehat{R}$ *refers to the Gelman–Rubin diagnostics*

(a) Variable selection and grouping

Variable	**Model**	**Frequency**	**Posterior probability**
$\mathbf{x}$	{cloudcover}	31,992	0.53
	{time, cloudcover}	27,587	0.46
$\mathbf{w}$	{veg200m, pubassit}	51,343	0.86
	{ufbldg, veg200m, pubassit}	7234	0.12
$\boldsymbol{\nu}$	$T_1 = \{2, 4\}, T_2 = \{1, 3, 5\}$	38,973	0.65
	$T_1 = \{2\}, T_2 = \{1, 3, 4, 5\}$	7445	0.12
	$T_1 = \{2\}, T_2 = \{4\}, T_2 = \{1, 3, 5\}$	3745	0.06

(b) Parameter estimation

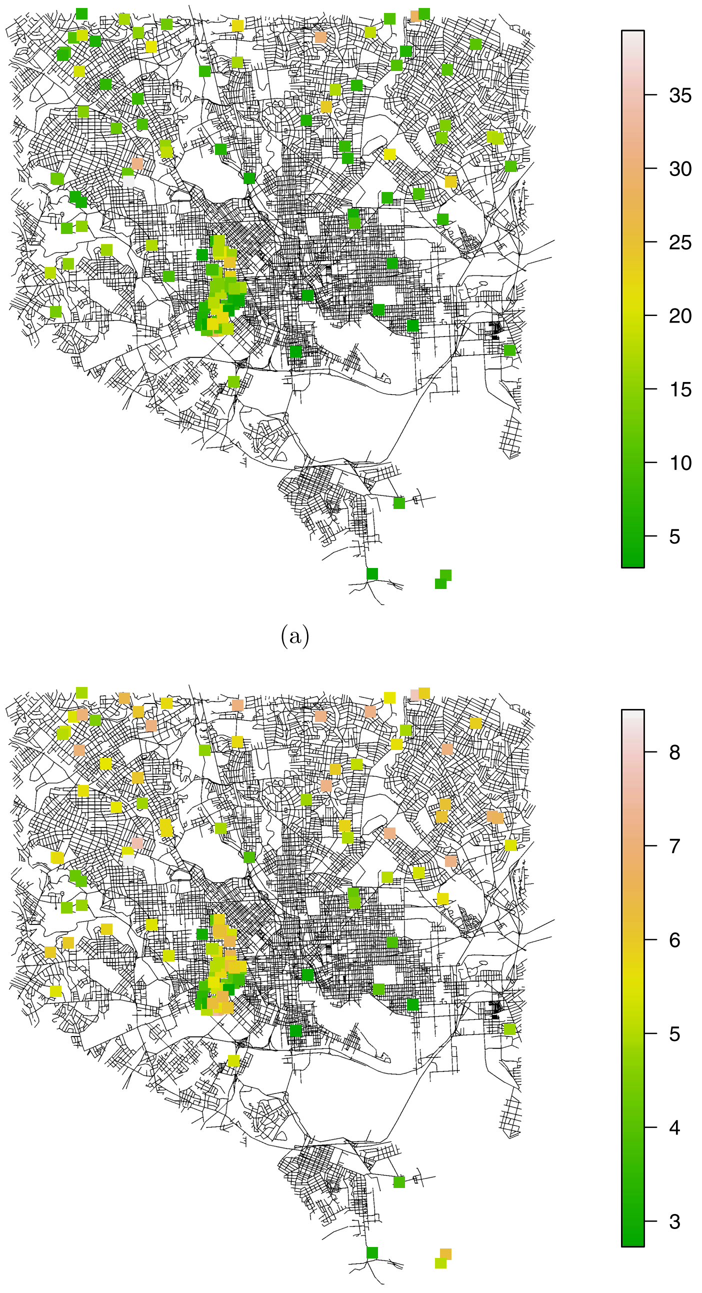
Parameter	$\boldsymbol{\mu_{\text{post}}}$	$\boldsymbol{\sigma_{\text{post}}}$	$\boldsymbol{Q_{0.025}}$	$\boldsymbol{Q_{0.975}}$	$\boldsymbol{\widehat{R}}$
intercept	−2.31	0.07	−2.45	−2.17	1.00
time	−0.10	0.03	−0.15	−0.04	1.00
cloudcover	−0.04	0.03	−0.09	0.01	1.00
ufbldg	−0.02	0.01	−0.03	−0.01	1.00
veg200m	0.06	0.01	0.05	0.09	1.00
pubassit	0.02	0.01	0.01	0.04	1.00
γ_{01}	0.35	0.07	0.23	0.51	1.01
γ_{02}	0.16	0.05	0.07	0.26	1.01
γ_{03}	0.48	0.08	0.33	0.67	1.01
γ_{04}	0.14	0.05	0.05	0.24	1.01
γ_{05}	0.38	0.07	0.25	0.55	1.01
ν_{24}	0.08	0.02	0.05	0.11	1.01
ν_{135}	0.17	0.03	0.12	0.23	1.01

fied as important predictors for detectability of American Robin. For the covariates in the intensity model, `ufbldg`, `veg200m`, and `pubassit` are selected as the important factors in all cases. For the dispersion parameters, the results suggest the most probable model has the grouping $T_1 = \{2, 4\}$, $T_2 = \{1, 3, 5\}$ (with posterior probability 0.6496), indicating that the data in 2005, 2007, and 2009 exhibit a similar amount of dispersion, whereas the data for 2006 and 2008 show similar amounts of dispersion.

Last, we consider the posterior mode model (i.e., the model with the highest posterior probability) for the Bin-CMP mixture model **M2** in order to draw inference about how the different covariates affect high detectability and abundance of the American Robin within the study domain. We conclude that an important covariate is a positively (or negatively) significant factor if the lower (or upper) end of 95% CIs is greater (or smaller) than 0, respectively. For the posterior mode model, we include only the intercept, `time`, and `cloudcover` in (12), whereas for the covariates in the intensity model, only `ufbldg`, `veg200m`, and `pubassit` are included. For the dispersion parameters, we consider the case where $\nu_2 = \nu_4 = \nu_{24}$ and $\nu_1 = \nu_3 = \nu_5 = \nu_{135}$. Table 2 presents the posterior summary statistics and Gelman–Rubin diagnostics [Brooks and Gelman (1998)] for model parameters. It is shown that in all cases $\widehat{R}$ is close to 1, indicating convergence has been reached. Moreover, `time` is negatively correlated with the detectability of the American Robin, that is, the earlier the survey is conducted, the more likely it is that we can detect American Robin. In terms of the intensity, `ufbldg` is negatively related to the abundance of American Robin, whereas `veg200m` and `pubassit` are positively related. As a result, for bird survey points nearby more buildings, the abundance of American Robin is lower; while for survey points with a higher percentage of vegetation and residents of lower socio-economic status, the abundance of American Robin is higher. As an example, Figure 2 provides a spatial map for the posterior mean and standard deviation of the abundance estimate (from **M2**) for 2009, whereas Figures 2 and 3 of the supplemental article [Wu et al. (2015)] illustrate how the abundance estimates and their standard errors change over the duration of the period studied (2005–2009). Last, our results suggest that the American Robin are overdispersed within the study domain over all of the years considered.

7. Discussion.

Motivated by the American Robin data from the BES, we developed a class of Bayesian hierarchical binomial mixture models that allow for automated variable selection and grouping in the presence of unbalanced nested design. In addition, we demonstrate that over- and underdispersion in the data can be accounted for by specifying an appropriate model for the abundance parameter, namely, a Bin-CMP model. More importantly, we allow for large-scale spatial dependence to be accounted for by adding a spatial component to the intensity model (i.e., through a spatial basis function expansion). Under the binomial mixture modeling framework, the use of a low-rank spatial representation proves to be a computationally advantageous approach to building in spatial dependence.



(a)

(b)

FIG. 2. *Plots of posterior mean and standard deviation of abundance estimates for* 2009 *in the Bin-CMP model assuming the posterior mode model for* **M2**. *Note that* **M2** *only contains covariates in the intensity model*. (a) *Posterior mean*, (b) *posterior sd*.

Although we have presented a model (**M2**) that accounts for covariate information, spatial maps that predict abundance at unobserved locations could be obtained using model **M3** and thereby take advantage of the spline formulation. In contrast, both models **M1** and **M2** would require imputation of covariates at unobserved locations (i.e., additional data models) to predict abundance at unobserved locations. Consequently, since our goal is primarily inferential, this direction has not been pursued here.

The class of binomial mixture models we consider assume population closure within each primary sampling period. Such an assumption is often justified based on biological and/or ecological considerations, when the primary sampling period covers a relatively short time frame. In our case, the justification of the closed population assumption is based on ecological considerations. However, it may also be possible to extend our model to verify the assumption of population closure following the framework of Dail and Madsen (2011) by decomposing the true abundance into the sum of two independent components, that is, the total number of survivors from the previous sampling period (by introducing a survival rate parameter in the model) and new additions prior to the current sampling period (by introducing a birth parameter in the model). This is a subject of future research.

Although the binomial mixture models we propose can accommodate unbalanced data structures, the amount of missing data can impact model selection and parameter estimation. As discussed in the second simulated example, the model averaged estimates for dispersion parameters are positively biased when the simulated data exhibit the same missing pattern as the American Robin data and spatial components are included to account for spatial dependence in the intensity model. Nevertheless, we note that grouping of dispersion parameters leads to a "borrowing of strength," since data collected over different years are pooled together if the corresponding dispersion parameters fall into the same group. In other words, this pooling of data helps mitigate the negative impacts of missing values. In general, a comprehensive assessment of the effect of missing data is problem specific and depends on both the pattern of missingness and the underlying spatial dependence (e.g., the effective sample size). In practice, we advocate evaluating these effects through simulated data examples, similar to those conducted here.

It is important to note that all of the models we considered for the American Robin data provide similar results regarding the identification of important covariates for detectability and intensity, as well as the grouping of dispersion parameters. First, `time,` and `cloudcover` are identified to be important covariates for high detectability of the American Robin, with the former being negatively related to observing the American Robin. However, one should be careful when interpretating `cloudcover` due to the difficulty in estimating it objectively [Vignola, Michalsky and Stoffel (2012)]. On the other hand, `ufbldg`, `veg200m,` and `pubassit` are found to be important predictors for abundance of the American Robin. In terms of dispersion, the American Robin data demonstrates overdisperion. Importantly, the class of binomial mixture models we propose is of independent interest and when coupled with the CMP distribution can be used in cases where the type of dispersion (i.e., over- and underdispersion) varies over time. In this sense, the Bin-CMP mixture model is extremely versatile, as it can be used for modeling equi-, over-, and underdispersed data (e.g., for modeling abundance of less prevalent species, such as the Eastern Wood Pewee or Wood Thrush in the BES).

Acknowledgments. The authors would like to thank the Editor Tilmann Gneiting, Associate Editor, and three anonymous referees for providing valuable comments that have helped strengthen this manuscript.

SUPPLEMENTARY MATERIAL

Supplement to "Bayesian binomial mixture models for estimating abundance in ecological monitoring studies" (DOI: 10.1214/14-AOAS801SUPP; .pdf). The supplementary material contains the MCMC sampling algorithm, details regarding computation times for the models implemented, and additional figures.

REFERENCES

ARONSON, M. F., LA SORTE, F. A., NILON, C. H., KATTI, M., GODDARD, M. A., LEPCZYK, C. A., WARREN, P. S., WILLIAMS, N. S., CILLIERS, S., CLARKSON, B. et al. (2014). A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers. *Proc. R. Soc. B Biol. Sci.* **281** 20133330. 1780.

BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. MR1665662

CAMERON, A. C. and TRIVEDI, P. K. (1998). *Regression Analysis of Count Data. Econometric Society Monographs* **30**. Cambridge Univ. Press, Cambridge. MR1648274

CARROLL, R. J. and LOMBARD, F. (1985). A note on N estimators for the binomial distribution. *J. Amer. Statist. Assoc.* **80** 423–426. MR0792743

CHANDLER, R. B., ROYLE, J. A. and KING, D. I. (2011). Inference about density and temporary emigration in unmarked populations. *Ecology* **92** 1429–1435.

CONWAY, R. and MAXWELL, W. (1962). A queuing model with state dependent service rates. *Int. J. Ind. Eng.* **12** 132–136.

CRESSIE, N. and JOHANNESSON, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 209–226. MR2412639

CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ. MR2848400

DAIL, D. and MADSEN, L. (2011). Models for estimating abundance from repeated counts of an open metapopulation. *Biometrics* **67** 577–587. MR2829026

DENISON, C. (2010). Effects of socioeconomics on European Starling (Sturnus Vulgaris) abundance in Baltimore, Maryland. Master's thesis, Univ. Missouri, Columbia, MO.

FURRER, R., NYCHKA, D. and SAIN, S. (2012). fields: Tools for spatial data. R package version 6.7.

GEORGE, E. and MCCULLOCH, R. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GEORGE, E. and MCCULLOCH, R. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–374.

GRAVES, T., KENDALL, K., ROYLE, J., STETZ, J. and MACLEOD, A. (2011). Linking landscape characteristics to local grizzly bear abundance using multiple detection methods in a hierarchical model. *Animal Conservation* **14** 652–664.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

HERBERS, J. M. (1989). Community structure in North temperate ants: Temporal and spatial variation. *Oecologia* **81** 201–211.

HODGES, J. S. and REICH, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *Amer. Statist.* **64** 325–334. MR2758564

HOLAN, S., WANG, S., ARAB, A., SADLER, E. J. and STONE, K. (2008). Semiparametric geographically weighted response curves with application to site-specific agriculture. *J. Agric. Biol. Environ. Stat.* **13** 424–439. MR2590938

HOOTEN, M. B. and HOBBS, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecol. Mono.* **85** 3–28.

KÉRY, M. (2008). Estimating abundance from bird counts: Binomial mixture models uncover complex covariate relationships. *The Auk* **125** 336–345.

KÉRY, M. and ROYLE, J. (2010). Hierarchical modelling and estimation of abundance and population trends in metapopulation designs. *J. Anim. Ecol.* **79** 453–461.

KÉRY, M., ROYLE, J. and SCHMID, H. (2005). Modeling avian abundance from replicated counts using binomial mixture models. *Ecol. Appl.* **15** 1450–1461.

KING, R. and BROOKS, S. P. (2002). Bayesian model discrimination for multiple strata capture–recapture data. *Biometrika* **89** 785–806. MR1946510

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, Hoboken, NJ. MR1925014

LOSS, S. R., RUIZ, M. O. and BRAWN, J. D. (2009). Relationships between avian diversity, neighborhood age, income, and environmental characteristics of an urban landscape. *Biol. Conserv.* **142** 2578–2585.

MALONEY, M. and AUFFREY, C. (2013). *The Social Areas of Cincinnati*, 5th ed. School of Planning, Univ. Cincinnati, Cincinnati, OH. Available at http://www.socialareasofcincinnati.org/files/FifthEdition/SASBook.pdf.

MINKA, T., SHMUELI, G., KADANE, J., BORLE, S. and BOATWRIGHT, P. (2003). Computing with the COM-Poisson distribution. Technical report, Statistics Dept., Carnegie Mellon Univ., Pittsburgh, PA.

MÜLLER, N., IGNATIEVA, M., NILON, C. H., WERNER, P. and ZIPPERER, W. C. (2013). Patterns and trends in urban biodiversity and landscape design. In *Urbanization, Biodiversity and Ecosystem Services: Challenges and Opportunities* (T. Elmqvist, S. Parnell, M. Fragkias, M. Schewenius, J. Goodness, M. Sendstad, B. Güneralp, K. C. Seto, P. J. Marcotullio, C. Wilkinson and R. I. McDonald, eds.) 123–174. Springer, Berlin.

O'HARA, R. B. and SILLANPÄÄ, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Anal.* **4** 85–117. MR2486240

OH, J., WASHINGTON, S. P. and NAM, D. (2006). Accident prediction model for railway-highway interfaces. *Accident Anal. Prev.* **38** 346–356.

PICKETT, S. T. A., CADENASSO, M. L., GROVE, J. M., BOONE, C. G., GROFFMAN, P. M., IRWIN, E., KAUSHAL, S. S., MARSHALL, V., MCGRATH, B. P., NILON, C. H., POUYAT, R. V., SZLAVECZ, K., TROY, A. and WARREN, P. (2011). Urban ecological systems: Scientific foundations and a decade of progress. *J. Environ. Econ. Manage.* **92** 331–362.

PICKETT, S., BRUSH, G., FELSON, A., MCGRATH, B., GROVE, J., NILON, C., SZLAVECZ, K., SWAN, C. and WARREN, P. (2012). Understanding and working with urban biodiversity: The Baltimore Ecosystem Study. *CityGreen* **4** 68–77.

POLLOCK, K. H. (1982). A capture–recapture design robust to unequal probability of capture. *J. Wildl. Manag.* **46** 752–757.

RIDOUT, M. S. and BESBEAS, P. (2004). An empirical model for underdispersed count data. *Stat. Model.* **4** 77–89. MR2037815

ROYLE, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics* **60** 108–115. MR2043625

ROYLE, J. and DORAZIO, R. (2006). Hierarchical models of animal abundance and occurrence. *J. Agric. Biol. Environ. Stat.* **11** 249–263.

ROYLE, J. A. and DORAZIO, R. M. (2008). *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press, San Diego, CA.

ROYLE, J. and LINK, W. (2005). A general class of multinomial mixture models for anuran calling survey data. *Ecology* **86** 2505–2512.

ROYLE, J. A. and WIKLE, C. K. (2005). Efficient statistical mapping of avian count data. *Environ. Ecol. Stat.* **12** 225–243. MR2144403

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press, Cambridge. MR1998720

SALLABANKS, R. and JAMES, F. C. (1999). American Robin (Turdus migratorius). In *The Birds of North America* **462** (A. Pool and F. Gill, eds.) 1–27. Academy of Natural Sciences, Philadelphia, PA.

SELLERS, K. F., BORLE, S. and SHMUELI, G. (2012). The COM-Poisson model for count data: A survey of methods and applications. *Appl. Stoch. Models Bus. Ind.* **28** 104–116. MR2911711

SHMUELI, G., MINKA, T. P., KADANE, J. B., BORLE, S. and BOATWRIGHT, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. *J. Roy. Statist. Soc. Ser. C* **54** 127–142. MR2134602

SHOCHAT, E., LERMAN, S. and FERNÁNDEZ-JURICIC, E. (2010). Birds in urban ecosystems: Population dynamics, community structure, biodiversity, and conservation. In *Urban Ecosystem Ecology* (J. A. Peterson and A. Volder, eds.) *Agronomy Monographs* **55** 75–86. ASA-CSSA-SSSA Madison, WI.

SMALLBONE, L. T., LUCK, G. W. and WASSENS, S. (2011). Anuran species in urban landscapes: Relationships with biophysical, built environment and socio-economic factors. *Landsc. Urb. Plan.* **101** 43–51.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380

R DEVELOPMENT CORE TEAM (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

VER HOEF, J. M. V. and BOVENG, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology* **88** 2766–2772.

VIGNOLA, F., MICHALSKY, J. and STOFFEL, T. (2012). *Solar and Infrared Radiation Measurements*. CRC Press, Boca Raton, FL.

WEBSTER, R. A., POLLOCK, K. H. and SIMONS, T. R. (2008). Bayesian spatial modeling of data from avian point count surveys. *J. Agric. Biol. Environ. Stat.* **13** 121–139. MR2432396

WENGER, S. J. and FREEMAN, M. C. (2008). Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology* **89** 2953–2959.

WIKLE, C. K. (2010). Low-rank representations for spatial processes. In *Handbook of Spatial Statistics* (A. E. Gelfand, P. J. Diggle, M. Fuentes and P. Guttorp, eds.) 107–118. Chapman & Hall, Boca Raton, FL. MR2730946

WILLIAMS, B., NICHOLS, J. and CONROY, M. (2002). *Analysis and Management of Animal Populations*. Academic Press, San Diego, CA.

WU, G., HOLAN, S. H. and WIKLE, C. K. (2013). Hierarchical Bayesian spatio-temporal Conway–Maxwell Poisson models with dynamic dispersion. *J. Agric. Biol. Environ. Stat.* **18** 335–356. MR3110897

WU, G., HOLAN, S. H., NILON, C. H. and WIKLE, C. K. (2015). Supplement to "Bayesian binomial mixture models for estimating abundance in ecological monitoring studies." DOI:10.1214/14-AOAS801SUPP.

G. WU
SAS INSTITUTE INC.
SAS CAMPUS DRIVE
CARY, NORTH CAROLINA 27513
USA

S. H. HOLAN
C. K. WIKLE
DEPARTMENT OF STATISTICS
UNIVERSITY OF MISSOURI
146 MIDDLEBUSH HALL
COLUMBIA, MISSOURI 65211-6100
USA
E-MAIL: holans@missouri.edu

C. H. NILON
DEPARTMENT OF FISHERIES AND WILDLIFE SCIENCES
UNIVERSITY OF MISSOURI
ANHEUSER-BUSCH NATURAL RESOURCES BUILDING
COLUMBIA, MISSOURI 65211-7240
USA