Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–31

A Unified Bayesian Framework for Modeling Measurement Error in Multinomial Data

Matthew D. Koslovsky*, Andee Kaplan†,
Victoria A. Terranova‡, and Mevin B. Hooten§

Abstract. Measurement error in multinomial data is a well-known and well-studied inferential problem that is encountered in many fields, including engineering, biomedical and omics research, ecology, finance, official statistics, and social sciences. Methods developed to accommodate measurement error in multinomial data are typically equipped to handle false negatives or false positives, but not both. We provide a unified framework for accommodating both forms of measurement error using a Bayesian hierarchical approach. We demonstrate the proposed method's performance on simulated data and apply it to acoustic bat monitoring and official crime data.

Keywords: categorical data, criminology, ecology, misclassification, record linkage, zero-inflation.

1 Introduction

Measurement error in multinomial data is a well-known and well-studied inferential problem that is encountered in many fields, including engineering, biomedical and omics research, ecology, finance, official statistics, and social sciences (Swartz et al., 2004; Pérez et al., 2007; Molinari, 2008; Datta et al., 2021; Mulick et al., 2022). In this work, we define measurement error as the discrepancy between an observed or measured variable and its true value and consider two types of measurement error defined as (1) false negatives that occur when a particular category or class is present in the population but it is not observed in the sample and (2) false positives that occur when a sampled observation is misclassified into the wrong category. While both types of measurement error may be present in the data, existing methods are not designed to accommodate them simultaneously which may bias inference. To fill this gap, we propose a unified framework for accommodating both forms of measurement error when modeling multinomial data. Our approach differs from existing methods in that it models the probability of misclassification for each individual observation explicitly; is scalable to high-dimensional classification problems; and accommodates individual-level covariates associated with

*Department of Statistics, Colorado State University, Fort Collins, CO, USA, matt.koslovsky@colostate.edu
†Department of Statistics, Colorado State University, Fort Collins, CO, USA, andee.kaplan@colostate.edu
‡Department of Criminology & Criminal Justice, University of Northern Colorado, Greeley, CO, USA, victoria.terranova@unco.edu
§Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, TX, USA, mevin.hooten@austin.utexas.edu

 https://doi.org/10.1214/24-BA1477

the probability of being a true/false negative, the true classification probabilities, and the probability of misclassification.

Modeling false negatives in multinomial data is closely related to the notion of handling zero-inflation in univariate and multivariate count data. Count data are considered zero-inflated when the number of zeroes observed in the data set is larger than expected under the assumptions of the sampling distribution. Zero-inflated count models are typically constructed as a two-component mixture of a point mass at zero and a sampling distribution for the count data (e.g., Poisson or negative binomial distributions in the univariate setting; Xu et al., 2015; Zhang and Yi, 2020; Jiang et al., 2021; Shuler et al., 2021). To achieve this, an at-risk indicator is introduced into the model to differentiate between at-risk zeros (i.e., a zero count is observed, but there is a positive probability of occurrence) and structural zeros (i.e., a zero count is observed because there is zero probability of occurrence; Neelon, 2019), or equivalently between false negatives and true negatives, respectively. In multivariate settings, researchers link zero-inflated univariate count models via latent parameters that control the dependence structure between counts (Aitchison and Ho, 1989; Chiquet et al., 2021). This approach models the multivariate counts unconditionally on the total count for the sample and is therefore not designed for multinomial classification problems or multivariate compositional count data where the total number of counts is fixed. Koslovsky (2023) introduced a zero-inflated Dirichlet-multinomial (DM) distribution for handling excess zeros in multivariate compositional count data, which differs from traditional approaches for modeling zero-inflation in count data by assuming a mixture distribution on the count probabilities as opposed to the sampling distribution. Using a combination of data augmentation strategies, their approach is scalable to large compositional spaces, can accommodate covariates associated with zero-inflation and relative abundances, and has shown promising estimation performance in simulation.

Existing methods designed to model false positives in multinomial data typically assume the observed classifications follow a multinomial distribution given the true (latent) classification (Swartz et al., 2004; Pérez et al., 2007; Frénay and Verleysen, 2013; Wang et al., 2020). With this approach, the number of rows for the resulting matrix of classification probabilities equals the number of true categories, and the number of columns equals the number of observed categories with each row summing to one. The task of modeling false positives in multinomial data draws parallels to a popular approach for entity resolution. Entity resolution is the process of resolving duplicates in many overlapping data sets without the benefit of a unique identifying attribute. In the hit-miss approach to entity resolution, observed records are assumed to either represent the true records associated with an entity (hit) or a distorted version of this truth (miss) (Tancredi and Liseo, 2011; Copas and Hilton, 1990). These potentially noisy records are then directly modeled using a mixture model in which two records that are associated with the same latent truth refer to the same entity and can be deduplicated (Steorts et al., 2016). This direct approach of modeling measurement error in the likelihood is an analog to the misclassification problem we are interested in addressing in that the observed classifications can either be the true value (hit) or a distorted version of that truth (miss). Potential benefits of directly modeling the distortion process in a hit-miss framework include the ability to choose an appropriate distribution for the

misclassification, incorporation of expert knowledge into the model via the priors on misclassification probabilities, and inference on the probability of misclassification after obtaining data.

A fundamental issue shared by any model designed to accommodate misclassification is that the model is not identifiable without additional information about the true classification, as well as zero-inflation, process beyond the raw data (Swartz et al., 2004). Existing methods designed to accommodate false positives deal with identifiability using informative priors in Bayesian settings, auxiliary/calibration data to estimate the matrix of classification probabilities (typically referred to as a confusion matrix) separately from the count model, or validating the true classification of a subset of the data (Chambert et al., 2015; Guillera-Arroita et al., 2017; Wright et al., 2020). Stratton et al. (2022) explore these strategies rigorously in simulation. Swartz et al. (2004) provide an extensive discussion of identifiability issues in multinomial classification models and propose using constraints to break the symmetry of the model, similar to what is done to accommodate label switching in Bayesian mixture models (Jasra et al., 2005).

In this work, we propose a novel method for simultaneously modeling false positives and false negatives in multinomial data. Specifically, we assume a zero-inflated DM distribution to accommodate potential false negatives in the underlying true classifications as well as potential overdispersion. We then introduce a latent hit-miss indicator to model misclassification that allows our approach to differentiate between true detections and detection by chance. The proposed model belongs to the class of semi-supervised learning methods because it can incorporate any amount of individual-level validation data, including no validation data in unsupervised settings. We use a combination of data augmentation techniques to scale the model to high-dimensional settings found in practice. In a variety of simulation settings, we demonstrate the improved estimation performance of the proposed method compared to alternative approaches for handling false positives or false negatives in multivariate count data. We then apply the proposed method to two data sets collected in ecological and criminal justice research. Applied to bat monitoring data, we show how the proposed method can serve as an alternative approach for accommodating imperfect detection in multispecies occupancy-detection modeling. Our approach differs from existing multispecies occupancy-detection models because it does not require specification of the ecological process and instead models the true latent classifications explicitly. In a second application study, we show how the proposed method can accommodate potential biases attributed to zero-inflation and misclassification in official crime data. By taking a fully-Bayesian approach, our method propagates the uncertainty of potential false positives and false negatives in the estimation of parameters of interest.

2 Methods

In this section, we present a general formulation for modeling measurement error in multinomial data, making connections to relevant occupancy-detection and official crime data modeling aspects as necessary. Let the C-dimensional vector $\boldsymbol{y}_{ijl}$ represent the observed classification for the ith ($i = 1, \ldots, N$) observation (or site/location/jurisdiction)

at the jth ($j = 1, \ldots, n_i$) measurement (or visit/survey) for the lth ($l = 1, \ldots, L_{ij}$) individual (or organism/incident), where $y_{ijlc} = 1$ indicates the observed individual was classified into the cth category (or species/crime) and 0 otherwise. In general, the model does not require more than one measurement of each site (i.e., $n_i > 1$). However in various fields, including ecological monitoring, each site is typically visited multiple times to improve inference (MacKenzie et al., 2002; Lele et al., 2012). We let the T-dimensional vector $\boldsymbol{z}_{ijl}$ represent the individual's true classification (or species/crime), where $z_{ijlt} = 1$ indicates the individual truly belongs to the tth category and 0 otherwise. For ease of presentation, we assume $T = C$ and that the ordering of the elements is the same in $\boldsymbol{y}_{ijl}$ and $\boldsymbol{z}_{ijl}$ (i.e., $y_{ijlt} = z_{ijlt} = 1$ indicates that the latent and observed categories are the same).

For each individual, we introduce a latent hit-miss or misclassification indicator $\tau_{ijl} \in \{0, 1\}$, where 0 indicates that $\boldsymbol{y}_{ijl} = \boldsymbol{z}_{ijl}$. To model the observed classifications, we assume

$$\boldsymbol{y}_{ijl}|\boldsymbol{\theta}_t, \tau_{ijl}, \boldsymbol{z}_{ijl} \sim \tau_{ijl}\,\text{Multinomial}(1, \boldsymbol{\theta}_t) + (1 - \tau_{ijl})\delta_{\boldsymbol{z}_{ijl}}(\boldsymbol{y}_{ijl}), \tag{1}$$

where $\boldsymbol{\theta}_t$ is a C-dimensional vector of observed classification probabilities for the tth true classification and $\delta_w(\cdot)$ is a Dirac delta function at w. With this formulation, we assume that if there is no misclassification (i.e., $\tau_{ijl} = 0$), then $\boldsymbol{y}_{ijl} = \boldsymbol{z}_{ijl}$, otherwise, the individual is considered misclassified with $\tau_{ijl} = 1$. Note that this approach allows for $\boldsymbol{y}_{ijl}$ to be misclassified into the correct category (i.e., $\boldsymbol{y}_{ijl} = \boldsymbol{z}_{ijl}$, but $\tau_{ijl} = 1$). As such, our modeling approach places a positive probability of a "lucky guess" to occur. In Section 3, we discuss how to restrict the model to prevent this from occurring if desired.

We assume the classification probabilities depend on the true classification of each individual with $\boldsymbol{\theta}_t \sim \text{Dirichlet}(\boldsymbol{\nu}_t)$, where $\boldsymbol{\nu}_t$ is a C-dimensional vector of concentration hyperparameters. To allow the classification probabilities to depend on an observed set of covariates, ν_{tc} can be replaced with a log-linear regression model similar to Wadsworth et al. (2017). Next, we let the latent misclassification indicators,

$$\tau_{ijl}|\boldsymbol{z}_{ijl}, \boldsymbol{\beta}_{\psi_t}, \boldsymbol{x}_{ijl} \sim \text{Bernoulli}(\psi_{ijl}), \tag{2}$$

where $\text{logit}(\psi_{ijl}) = \boldsymbol{x}'_{ijl}\boldsymbol{\beta}_{\psi_t}$, $\boldsymbol{x}_{ijl}$ is a P_ψ-dimensional vector of observed covariates that are observation-, measurement-, and/or individual-specific (including an intercept term), and $\boldsymbol{\beta}_{\psi_t}$ are the corresponding regression coefficients. We assume $\beta_{\psi_{tp}} \sim \text{Normal}(\mu_\psi, \sigma^2_\psi)$. Note that the covariate effects on misclassification are allowed to vary based on the true classification of the individual.

We model the true classification of each individual

$$\boldsymbol{z}_{ijl}|\boldsymbol{\Theta}_{ij} \sim \text{Multinomial}(1, \boldsymbol{\Theta}_{ij}), \tag{3}$$

where $\boldsymbol{\Theta}_{ij}$ is a T-dimensional vector of true classification probabilities (or relative abundances), which we assume follows a Dirichlet($\boldsymbol{\gamma}_{ij}$) with $\boldsymbol{\gamma}_{ij}$ a T-dimensional vector of concentration hyperparameters. With the availability of validation data, some of the $\boldsymbol{z}_{ijl}$ will be known and fixed in the model to inform the estimation of the classification matrix $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_T)'$, similar to Wright et al. (2020) and Spiers et al. (2022). In

practice, our model can accommodate any amount of validation data available, which, if error-free, can improve the estimation performance. By modeling the validation data at the individual level, the number of true categories does not have to equal the number of observed categories, which allows some categories to go unobserved, the potential for new categories to be observed, and the incorporation of individual-level covariates which can improve estimation, similar to Spiers et al. (2022). We refer to $\boldsymbol{\theta}$ as a classification matrix to differentiate it from methods that use a "confusion matrix," denoted as $\boldsymbol{\theta}^*$, to model false positives in multinomial data which, unlike our model, do not explicitly model misclassification.

We can equivalently model $\boldsymbol{\Theta}_{ij}$ as a set of independent gamma random variables normalized by their sum (i.e., $\boldsymbol{z}_{ijl}|\boldsymbol{\alpha}_{ij} \sim \text{Multinomial}(1, \frac{\alpha_{ijt}}{\bar{\alpha}_{ij}})$, where $\alpha_{ijt} \sim \text{Gamma}(\gamma_{ijt}, 1)$ and $\bar{\alpha}_{ij} = \sum_{t=1}^{T} \alpha_{ijt}$). This reparameterization enables us to differentiate between at-risk or structural zeros (i.e., account for potential zero-inflation, false negatives, or non-detection) by introducing a latent at-risk (or occupancy) indicator ζ_{ijt} for the tth category (or species/crime) at the ith observation (or site/jurisdiction) and jth measurement (or visit/survey). Specifically, we instead let

$$\alpha_{ijt}|\zeta_{ijt}, \gamma_{ijt} \sim \zeta_{ijt}\,\text{Gamma}(\gamma_{ijt}, 1) + (1 - \zeta_{ijt})\delta_0(\alpha_{ijt}), \tag{4}$$

similar to Koslovsky (2023). To model the relation between a set of observed covariates and the latent at-risk (or occupancy) indicators, we assume $\zeta_{ijt}|\boldsymbol{\beta}_{\eta_t}, \boldsymbol{x}_i \sim \text{Bernoulli}(\eta_{it})$, where $\text{logit}(\eta_{it}) = \boldsymbol{x}_i'\boldsymbol{\beta}_{\eta_t}$, $\boldsymbol{x}_i$ is a P_η-dimensional set of observation-specific covariates (including an intercept term), and $\boldsymbol{\beta}_{\eta_t}$ represent the corresponding true classification-specific regression coefficients. We then let $\beta_{\eta_{tp}} \sim \text{Normal}(\mu_\eta, \sigma_\eta^2)$. To allow the true classification probabilities (or relative abundances) to depend on a set of covariates, we set $\log(\gamma_{ijt}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}_{\gamma_t}$ with $\beta_{\gamma_{tp}} \sim \text{Normal}(\mu_\gamma, \sigma_\gamma^2)$ and $\boldsymbol{x}_{ij}$ an observation- and/or measurement-specific set of covariates. In general, the model does not require covariate information to inform the probability of an at-risk observation, the true classification probabilities, or the misclassification probabilities. In these settings, $\beta_{\eta_{t1}}$, $\beta_{\gamma_{t1}}$, and/or $\beta_{\psi_{t1}}$ can be treated as hyperparameters that reflect prior beliefs for the corresponding probabilities. Likewise, it is straightforward to adjust the model to accommodate observation-, measurement-, and/or, individual-level covariate information for the classification matrix concentration parameters, if available.

3 Posterior Sampling and Inference

For posterior inference, we construct a Metropolis-Hastings within Gibbs sampler. The full joint distribution is defined as

$$\prod_{i=1}^{N}\prod_{j=1}^{n_i}\prod_{l=1}^{L_{ij}} p(\boldsymbol{y}_{ijl}|\boldsymbol{\theta}_t, \tau_{ijl}, \boldsymbol{z}_{ijl})p(\boldsymbol{z}_{ijl}|\boldsymbol{\Theta}_{ij})p(\tau_{ijl}|\boldsymbol{z}_{ijl}, \boldsymbol{\beta}_{\psi_t}, \boldsymbol{x}_{ijl})p(\omega_{\tau_{ijl}})$$
$$\times \prod_{i=1}^{N}\prod_{j=1}^{n_i}\prod_{t=1}^{T} p(\alpha_{ijt}|\zeta_{ijt}, \boldsymbol{\beta}_{\gamma_t}, \boldsymbol{x}_{ij})p(\zeta_{ijt}|\boldsymbol{\beta}_{\eta_t}, \boldsymbol{x}_i)p(\omega_{\zeta_{tij}})$$

$$\times \prod_{i=1}^{N}\prod_{j=1}^{n_i} p(\mu_{ij}|\bar{\alpha}_{ij}) \prod_{t=1}^{T}\left[p(\boldsymbol{\beta}_{\eta_t})p(\boldsymbol{\beta}_{\psi_t})p(\boldsymbol{\beta}_{\gamma_t})p(u_t|\bar{a}_t)\prod_{c=1}^{C} p(a_{tc})\right], \tag{5}$$

where we introduce an auxiliary parameter $\mu_{ij}|\bar{\alpha}_{ij} \sim \text{Gamma}(1, \bar{\alpha}_{ij})$ for efficient sampling of α_{ijt}. The reparameterization of $\boldsymbol{\Theta}_{ij}$ with $\boldsymbol{\alpha}_{ij}$, coupled with the inclusion of μ_{ij}, reduces the computational demand of updating $\boldsymbol{\beta}_{\gamma_t}$ and provides closed-form Gibbs updates for μ_{ij} and $\alpha_{ijt}|\zeta_{ijt} = 1$, which greatly improves the overall scalability of the model to large T settings. See Koslovsky et al. (2020) and Koslovsky (2023) for more technical details of this data augmentation technique, its performance in high-dimensional settings, and parameter identifiability in this portion of the model. Similarly, we reparameterize $\theta_{tc} = a_{tc}/\bar{a}_t$ and assume $a_{tc} \sim \text{Gamma}(\nu_{tc}, 1)$ with auxiliary parameter $u_t \sim \text{Gamma}(1, \bar{a}_t)$ and $\bar{a}_t = \sum_{c=1}^{C} a_{tc}$. In addition to enabling efficient sampling of $\boldsymbol{\theta}_t$, this step provides the opportunity to easily incorporate covariates and restrict the model to disallow correct classifications by chance by fixing $a_{tt} = 0$. Further, we exploit a Pólya-Gamma (PG) augmentation scheme that provides closed-form Gibbs updates for the regression coefficients associated with the latent at-risk (or occupancy) indicators, ζ_{ijt}, and the misclassification indicators, τ_{ijl}, without sacrificing their interpretability as log odds ratios following Polson et al. (2013). Specifically by introducing a latent set of auxiliary parameters $\omega_{\zeta_{tij}} \sim \text{PG}(1, 0)$ and $\omega_{\tau_{ijl}} \sim \text{PG}(1, 0)$, the full-conditional distributions of $\boldsymbol{\beta}_{\eta_t}$ and $\boldsymbol{\beta}_{\psi_t}$ are multivariate normal. A graphical representation of the proposed approach for modeling misclassification in zero-inflated Dirichlet-multinomial models, missZIDM, is presented in Figure 1. More details of the Markov chain Monte Carlo (MCMC) sampler used to implement our model are provided in the Supplementary Material (Koslovsky et al., 2024a).

Under the assumptions of the proposed model, $\beta_{\eta_{tp}}$ is interpreted as the expected change in log odds ratio of the tth category being at-risk at the ijth measurement, and $\exp(\beta_{\gamma_{tp}})$ is interpreted as the multiplicative change in the concentration parameter γ_t for a one unit increase in the pth covariate holding all else constant. While the latter provides inference about how concentrated the true classification probabilities (or relative abundances) are around the mean, we also are interested in inferring the relation between a covariate and the true classification probabilities directly. This inference is more complicated because each covariate is potentially associated with each category, as described in Dai et al. (2019). The multiplicative effect on the tth category for a one unit increase in the pth covariate for the ith observation at the jth measurement is defined as

$$\pi_{ijtp} = \frac{\Theta_{ijt}(\boldsymbol{x}_{ij}^{(p)})}{\Theta_{ijt}(\boldsymbol{x}_{ij})} = \exp(\beta_{\gamma_{tp}}) \frac{\sum_{s=1}^{T}\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_{\gamma_s})}{\sum_{s=1}^{T}\exp(\boldsymbol{x}_{ij}^{(p)\prime}\boldsymbol{\beta}_{\gamma_s})}, \tag{6}$$

where $\boldsymbol{x}_{ij}^{(p)} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp}+1, \ldots, x_{ijP})'$. Because the effect of the pth covariate on the tth category depends on its corresponding regression coefficient, $\beta_{\gamma_{tp}}$, in addition to its effect on the other categories and their corresponding concentration parameters, we may observe a decrease (increase) in the probability of the tth category with an increase in x_{ijp}, even if $\beta_{\gamma_{tp}} > 0$ ($\beta_{\gamma_{tp}} < 0$). Estimates of the true classification probabilities for each observation can be obtained by normalizing the vector $\boldsymbol{\alpha}_i$ over its

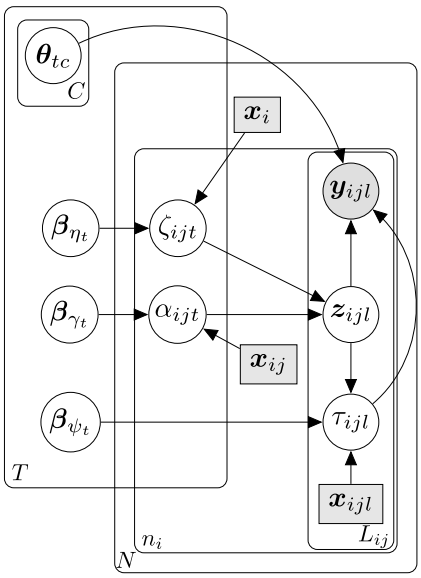

Figure 1: Graphical representation of missZIDM with covariate dependence. Note that auxiliary parameters and hyperparameters have been suppressed for clarity. $\boldsymbol{N}$ – total observations; $\boldsymbol{n_i}$ – total measurements per observation; $\boldsymbol{L}_{ij}$ – number of individuals at each measurement; $\boldsymbol{T}$ – true number of categories; $\boldsymbol{C}$ – observed number of categories.

sum for each MCMC iteration and then averaging over the samples. Estimates of the classification matrix $\boldsymbol{\theta}$ can be obtained similarly given $\boldsymbol{a}_t$. While we model the probability of misclassification separately from the classification matrix, we can generate inference on the confusion matrix, $\boldsymbol{\theta}^*$, typically estimated by other approaches. Assuming an intercept-term only model for the probability of misclassification for simplicity, each element of $\boldsymbol{\theta}^*$, $\theta^*_{tc} = \theta_{tc}\frac{\exp(\beta_{\psi_{t1}})}{1+\exp(\beta_{\psi_{t1}})} + [1 - \frac{\exp(\beta_{\psi_{t1}})}{1+\exp(\beta_{\psi_{t1}})}]I(t = c)$, where $I(\cdot)$ is an indicator function. For inference on these quantities, the posterior means of the MCMC samples are calculated and 95% credible intervals are constructed using the empirical quantiles.

4 Empirical Studies

In this section, we first compare the proposed model to alternative methods for handling false positives or false negatives in multinomial data in simulated settings without covariate information, repeated measurements, or validation data to inform the model, similar to the official crime data application. In a second scenario, we compare the proposed model to a multispecies occupancy model in settings designed to mimic the bat monitoring study.

The first scenario examines the estimation performance of missZIDM with respect to the at-risk probability, $\boldsymbol{\eta}$, the probability of misclassification, $\boldsymbol{\psi}$, the true probability of each category, $\boldsymbol{\Theta}$, and the confusion matrix, $\boldsymbol{\theta}^*$, at varying percentages of at-risk observations and misclassification. In the first scenario, we compare missZIDM to a similar approach that does not accommodate false negatives (missDM), an approach that assumes the true and observed classifications follow Dirichlet-multinomial models which does not explicitly model misclassification or handle false negatives (DMDM, similar to Swartz et al., 2004), and the recently developed zero-inflated Dirichlet-multinomial model (ZIDM; Koslovsky, 2023), which is designed to accommodate false negatives but not false positives. In this scenario, the models ignore any potential covariates and therefore only estimate intercept terms for $\boldsymbol{\eta}$, $\boldsymbol{\psi}$, and $\boldsymbol{\Theta}$. For the DMDM model, we set $\nu_{tt} = (\text{logit}^{-1}(\mu_\psi)/T + (1 - \text{logit}^{-1}(\mu_\psi))) \times T/\text{logit}^{-1}(\mu_\psi)$, which places a similar prior probability for correct classification as the methods with misclassification indicators. All methods were implemented in `R` using `Rcpp` (Eddelbuettel and François, 2011).

In this scenario, we generated $N = 50$ observations of $L_{ij} = 100$ individuals to cluster into $C = 10$ categories. We assumed that the true number of categories, T, matched the potentially observed number of categories, C. We evaluated the model in four settings with varying percentages of at-risk observations (1 - % true negatives) and misclassification (false positives). In these settings, we set $n_i = 1$ (i.e., no repeated measurements). Observation-specific at-risk indicators were sampled from a Bernoulli distribution with the probability of an at-risk observation set to either 0.25 or 0.75. The true classification of each individual was generated from a Dirichlet-multinomial distribution with concentration parameters set to one and overdispersion parameter set to 0.01, so that the model assumptions did not match the true data generation process. Misclassification occurred with 0.25 or 0.75 probability. The observed classifications were generated from a Dirichlet-multinomial model with a similar overdispersion parameter as above. We set concentration parameters $\boldsymbol{\nu}_t$ equal to their index (e.g., $\nu_{tc} = c$) with the tth element also equal to one, placing the least probability on a correct classification by chance. No validation data were used in these settings.

In the second scenario, we investigate how the proposed method performs when used for inference in multispecies occupancy-detection settings with data generated to mimic the bat monitoring data set. We evaluate the performance of the method with varying percentages of overdispersion in the true counts. We compare the proposed model to a similar version of the multispecies occupancy-detection model presented in Wright et al. (2020), which we refer to as DMZIP. This approach differs from the proposed model in that it makes distributional assumptions for the ecological process, does not explicitly model the latent classifications, and uses a confusion matrix to accommodate potential misclassification. Using the notation of our proposed model, DMZIP assumes the detection counts of each species $\sum_{l=1}^{L_{ij}} I(z_{ijlt} = 1) \sim \text{Poisson}(\lambda_{ijt}\zeta_{ijt})$, where $\zeta_{ijt} \sim \text{Bernoulli}(\eta_{it})$. Given the true latent cluster, the observed individual classifications $\boldsymbol{y}_{ijl} \sim \text{Multinomial}(\boldsymbol{\theta}_t^*)$, where $\boldsymbol{\theta}_t^* \sim \text{Dirichlet}(\boldsymbol{\gamma}_{ij})$. Code to implement DMZIP was adapted from Stratton (2022).

Specifically, we generated $N = 50$ sites with $n_i = 5$ visits per site and $C = 10$ possible species to observe. No covariates were used in the baseline simulation setting. Parameter

values for the occupancy probabilities, encounter rates, and confusion matrix used to simulate the data were obtained from the posterior mean estimates obtained with the count detection model fit to acoustic data in Stratton et al. (2022). The true occurrence probabilities ranged from 23% to 90%. Instead of assuming the total number of counts was fixed, as in scenario 1, the relative activity or encounter rates for each species at each site visit was generated from a negative binomial distribution with mean $\zeta_{it} * \lambda_t$ and variance $(\zeta_{it} * \lambda_t)^2/\sigma$, where ζ_{it} is the site-level occupancy indicator for a given species, λ_t is the expected number of detections or encounter rate of species t obtained from the bat monitoring data which ranged from 2.0 to 28.2, and σ is an overdispersion parameter. Note that as σ increases, the variance of the sampling distribution approaches the mean, and the data are more Poisson-like. The models were evaluated with $\sigma \in \{0.1, 1, 100\}$ and 25% validation data.

We then evaluated model performance on data generated similar to scenario 2 but with covariates informing the occupancy probability and the true encounter rates. We simulated 5 continuous covariates from a standard normal distribution in both levels of the model. In this setting, we set the overdispersion parameter for the negative binomial distribution $\sigma = 1$. The intercept terms $\beta_{\eta t0}$ ($\beta_{\gamma t0}$) were randomly sampled uniformly from logit(0.25) to logit(0.95) (0 to log(10)) with covariate effects set to ± 1 (± 0.2) with equal probability. The off-diagonal elements of the classification matrix $\boldsymbol{\theta}$ were sampled uniformly from 0.01 to 0.2 with diagonal elements uniformly sampled from 0.5 to 0.95. Thereafter, the rows of $\boldsymbol{\theta}$ were scaled to sum to one, and the individual classifications were sampled from a Multinomial$(1, \boldsymbol{\theta}_t)$. In this setting, we assumed 25% of the data were validated. Additionally, we evaluated the models in various other data generation settings including those with different sample sizes, sampling efforts, and percent validated data.

In both scenarios, each of the MCMC algorithms were run for 5,000 iterations treating the first 2,500 as burn-in and thinning to every other iteration, providing 1,250 iterations for inference. We assumed non- or weakly-informative priors $\gamma_{tc} = \nu_{tc} = \sigma^2_\eta = \sigma^2_\psi = \sigma^2_\gamma = 1$. In settings with no covariates in the model, we set μ_η and μ_ψ following the data generation process for all models. In settings with covariates in the model, the prior mean for the regression coefficients was set to 0. In the sensitivity analysis presented in the Supplementary Material, we explore the impact of prior misspecification of these hyperparameters on inference. To initialize each model, we set the true classifications $\boldsymbol{Z}_i$ to the observed classifications $\boldsymbol{Y}_i$, with $\boldsymbol{\tau}_i$ set accordingly. Auxiliary parameters, $\omega_{\tau_{tijl}}$ and $\omega_{\zeta_{tij}}$, and at-risk indicators, ζ_{tij} and α_{tc}, were initialized at one. The auxiliary parameters u_t and μ_{ij} were randomly initialized from a Gamma(1,1) and regression coefficients were set to 0.

We evaluated the models in terms of the average absolute value of the difference between the estimated and true probabilities (ABS) and Frobenius norm (FROB), which is the square root of the sum of the squared difference of the estimated and true probabilities for $\boldsymbol{\eta}$, $\boldsymbol{\psi}$, $\boldsymbol{\theta}^*$, and $\boldsymbol{\Theta}$. Note that the proposed method is the only method that provides estimates for all parameters simultaneously. In settings where covariates were incorporated into the data generation process (results presented in the Supplementary Material), the models were compared with respect to the estimation of the occupancy

probability regression coefficients, $\boldsymbol{\beta}_\eta$, and the confusion matrix, $\boldsymbol{\theta}^*$, because these maintained similar interpretation among all models. Results we report below were obtained by averaging over 50 replicated data sets for each setting.

4.1 Results

In the first scenario, the estimation performance for the probability of an at-risk observation, $\boldsymbol{\eta}$, improved as the true percentage of at-risk observations increased for missZIDM and ZIDM, with missZIDM demonstrating better performance than ZIDM in settings with more structural zeros (i.e., 25% at-risk observations) regardless of the amount of misclassification (Table 1). We observed that the proposed missZIDM always outperformed missDM, which ignores potential zero-inflation, when estimating the probability of misclassification, $\boldsymbol{\psi}$. All methods obtained relatively similar estimation performance for the true classification probabilities, $\boldsymbol{\Theta}$, with the proposed method demonstrating a slight advantage in the setting with 25% at-risk observations and misclassification. Estimation accuracy for the confusion matrix, $\boldsymbol{\theta}^*$, reduced as the percentage of misclassification increased for all methods. The proposed method and DMDM both outperformed missDM with respect to estimating the confusion matrix, $\boldsymbol{\theta}^*$. However, DMDM obtained a two-fold reduction in the absolute value of the bias compared to missZIDM when data were generated with greater misclassification percentages (0.04 and 0.08, respectively). Recall that missZIDM, unlike DMDM, does not directly provide estimates for $\boldsymbol{\theta}^*$, but they can be obtained using the estimated $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ values.

In scenario 2, we found the DMZIP model obtained the best estimation performance for $\boldsymbol{\beta}_\eta$ and $\boldsymbol{\theta}^*$ when $\sigma = 100$ (Table 2). However in settings with more overdispersion ($\sigma = 1$ and 0.1), the proposed method outperformed DMZIP with respect to both parameters. Notably, estimation performance was worse for both methods when $\sigma = 0.1$. Similar trends were observed with 75% validated data (Supplementary Table S1). These results demonstrate how the proposed method is preferred in the presence of overdispersion. Additionally, we found that missZIDM obtained the best estimation performance for $\boldsymbol{\beta}_\eta$ and $\boldsymbol{\theta}^*$ with covariates in the model, more species types, visits, and sites (Supplementary Table S2).

One of the major challenges of modeling measurement error in multinomial data is non-identifiability of the parameters, because there is no information contained in the raw data to inform zero-inflation or misclassification probabilities. Our approach is designed to account for non-identifiability through informative prior specifications and/or incorporating a subset of validation data to inform parameter estimates. As such, inferential results obtained by our method, and any method designed to model measurement error in multinomial data, will be sensitive to the amount of validation data used to inform the model as well as the specification of the hyperparameters. In the Supplementary Material, we present an extensive sensitivity analysis of the proposed model with varying percentages of validated data (Supplementary Tables S3 and S4) and hyperparameter specification (Supplementary Tables S5 and S6). Based on these results, we found that validation data are useful when available. With only 10% validation data, the proposed method was able to obtain less than 0.05% bias for all probability estimates on average. In the absence of validation data, the model performed better

	25% at-risk observations and 25% misclassification							
	η		ψ		Θ		θ^*	
	ABS	FROB	ABS	FROB	ABS	FROB	ABS	FROB
missZIDM	**0.05 (0.02)**	**0.25 (0.05)**	**0.02 (0.01)**	**0.07 (0.02)**	**0.02 (0.00)**	**0.71 (0.06)**	**0.01 (0.00)**	**0.14 (0.02)**
missDM	–	–	0.18 (0.01)	0.56 (0.02)	0.05 (0.00)	1.49 (0.07)	0.04 (0.00)	0.62 (0.02)
DMDM	–	–	–	–	0.05 (0.00)	1.66 (0.08)	0.01 (0.00)	0.15 (0.00)
ZIDM	0.13 (0.01)	0.43 (0.02)	–	–	0.05 (0.00)	1.63 (0.08)	–	–
	25% at-risk observations and 75% misclassification							
	η		ψ		Θ		θ^*	
	ABS	FROB	ABS	FROB	ABS	FROB	ABS	FROB
missZIDM	**0.09 (0.01)**	**0.33 (0.03)**	**0.36 (0.01)**	**1.15 (0.04)**	**0.11 (0.00)**	3.76 (0.14)	0.08 (0.00)	1.29 (0.04)
missDM	–	–	0.47 (0.01)	1.49 (0.03)	0.12 (0.00)	**3.75 (0.16)**	0.09 (0.00)	1.61 (0.03)
DMDM	–	–	–	–	0.12 (0.00)	3.93 (0.18)	**0.04 (0.00)**	**0.45 (0.00)**
ZIDM	0.14 (0.01)	0.45 (0.01)	–	–	0.12 (0.00)	3.93 (0.18)	–	–
	75% at-risk observations and 25% misclassification							
	η		ψ		Θ		θ^*	
	ABS	FROB	ABS	FROB	ABS	FROB	ABS	FROB
missZIDM	0.03 (0.01)	0.12 (0.04)	**0.03 (0.01)**	**0.10 (0.04)**	0.03 (0.00)	0.84 (0.03)	0.01 (0.00)	0.17 (0.03)
missDM	–	–	0.09 (0.02)	0.28 (0.05)	0.03 (0.00)	0.82 (0.02)	0.02 (0.00)	0.35 (0.05)
DMDM	–	–	–	–	0.03 (0.00)	0.81 (0.02)	**0.01 (0.00)**	**0.15 (0.00)**
ZIDM	**0.02 (0.01)**	**0.07 (0.04)**	–	–	**0.03 (0.00)**	**0.81 (0.02)**	–	–
	75% at-risk observations and 75% misclassification							
	η		ψ		Θ		θ^*	
	ABS	FROB	ABS	FROB	ABS	FROB	ABS	FROB
missZIDM	0.03 (0.01)	0.11 (0.03)	**0.40 (0.01)**	**1.26 (0.03)**	0.07 (0.00)	1.99 (0.08)	0.08 (0.00)	1.40 (0.03)
missDM	–	–	0.44 (0.01)	1.40 (0.02)	0.06 (0.00)	1.70 (0.07)	0.09 (0.00)	1.53 (0.02)
DMDM	–	–	–	–	0.06 (0.00)	1.53 (0.07)	**0.04 (0.00)**	**0.45 (0.00)**
ZIDM	**0.01 (0.00)**	**0.04 (0.01)**	–	–	**0.06 (0.00)**	**1.53 (0.07)**	–	–

Table 1: Simulation Results for Scenario 1: Estimation performance for $N = 50$ observations, $n_i = 1$ measurements, $L_{ij} = 100$ individuals, and $T = 10$ categories at varying percentages of at-risk observations and misclassification with 0% validation data. Bold indicates the best performing models. Standard deviations of performance metrics across the replicate data sets are provided in parentheses. ABS – absolute value of the difference between the estimated and true probabilities; FROB – Frobenius norm.

	$\sigma = 0.1$			
	$\boldsymbol{\eta}$		$\boldsymbol{\theta}^*$	
	ABS	FROB	ABS	FROB
missZIDM	**0.27 (0.07)**	**0.94 (0.20)**	**0.01 (0.00)**	**0.25 (0.01)**
DMZIP	0.46 (0.01)	1.54 (0.02)	0.04 (0.01)	1.24 (0.02)
	$\sigma = 1$			
	$\boldsymbol{\eta}$		$\boldsymbol{\theta}^*$	
	ABS	FROB	ABS	FROB
missZIDM	**0.15 (0.02)**	**0.60 (0.08)**	**0.01 (0.00)**	**0.22 (0.02)**
DMZIP	0.22 (0.01)	0.76 (0.04)	0.03 (0.00)	1.04 (0.05
	$\sigma = 100$			
	$\boldsymbol{\eta}$		$\boldsymbol{\theta}^*$	
	ABS	FROB	ABS	FROB
missZIDM	0.17 (0.01)	0.66 (0.06)	0.02 (0.00)	0.28 (0.03)
DMZIP	**0.03 (0.01)**	**0.12 (0.03)**	**0.01 (0.00)**	**0.15 (0.03)**

Table 2: Simulation Results for Scenario 2: Estimation performance for data generated similar to the bat monitoring data with 25% validation. Bold indicates the best performing models. Standard deviations of performance metrics across the replicate data sets are provided in parentheses. ABS – absolute value of the difference between the estimated and true probabilities; FROB – Frobenius norm.

with lower concentration parameters for the true classification probabilities, or relative abundances, and the observed classification probabilities. Additionally, we found that the model was relatively robust to misspecification of the misclassification prior for all other parameters. Similar results were observed for changes in the prior for the at-risk probability.

5 Real Data Applications

In this section, we apply the proposed method to two publicly available data sets. In Section 5.1, we show how the proposed method can serve as an alternative approach for accommodating imperfect detection in multispecies occupancy-detection modeling. For this analysis, we incorporate validation data to inform the true relative abundances, classification probabilities, and occupancy probabilities. In Section 5.2, we demonstrate how to accommodate potential zero-inflation and misclassification in official crime data in an unsupervised setting.

5.1 Application to Multispecies Bat Acoustic Monitoring Data

The goal of occupancy modeling in ecological research is to draw inference on species' true occurrence given a set of observations that are subject to imperfect detection due to observational error. Imperfect detection typically occurs in two different ways: (1) a species may go undetected and (2) an observed individual may be misclassified. Even with increased sampling effort, imperfect detection may still occur, resulting in biased inference if ignored when modeling (Kellner and Swihart, 2014).

Historically, statistical methods developed to handle imperfect detection have focused on non-detection (Hoeting et al., 2000; Bayley and Peterson, 2001; MacKenzie et al., 2002; Royle and Nichols, 2003; MacKenzie et al., 2003; Tyre et al., 2003; Broms et al., 2015; Dorazio et al., 2006, 2011; Devarajan et al., 2020). However, more recently researchers have proposed methods that account for both non-detection and misclassification, in part due to the emergence of automated species detection methods (e.g., unmanned aerial systems and automated recording units) and volunteer-based surveys (e.g., citizen science) for monitoring wildlife populations (McClintock et al., 2010). When developing single- or multispecies (community) occupancy models, researchers typically take a hierarchical approach, often referred to as occupancy-detection models, which jointly model the ecological and observation (or detection) process. This technique allows researchers to differentiate between latent species occupancy and observed species detection and effectively account for potential misclassification. Typically, this is achieved by introducing a site- or location-specific latent species indicator that models whether or not a species is present, reminiscent of the latent at-risk indicator for handling zero-inflation in count data. If the species is present (absent) at that site, there is a positive (zero) probability of detecting it. See Blasco-Moreno et al. (2019) for an in-depth discussion of zero counts in the context of ecological research studies, Scharf et al. (2022) for an overview of hierarchical models for occupancy data, and MacKenzie et al. (2017) for more background on methods for estimating and modeling occupancy.

Methods that handle potential species misclassification in occupancy modeling were initially developed for single species studies (Royle and Link, 2006; Miller et al., 2011; Chambert et al., 2015; Ruiz-Gutierrez et al., 2016; Chambert et al., 2018b). Chambert et al. (2018a) introduced a two-species occupancy model that accounts for both species misidentification and non-detection. Their approach is based on the premise that false detections for a given species occur due to the misidentification with a closely related species. Recently, Wright et al. (2020) developed a multispecies occupancy model that handles both forms of measurement error for two or more species at each site visit. By assuming (1) the true count of each species follows a Poisson distribution given it is present at the site visit, (2) the number of detections for each species follows a multinomial distribution given the true species counts, and (3) the detection counts are independent across species, the authors demonstrate how the observed/detected counts can be directly modeled without conditioning on the true count for each species. Spiers et al. (2022) developed a multispecies occupancy model similar to Wright et al. (2020) that accommodates individual-level validation data (as opposed to site-level) which allows for more flexibility when modeling heterogeneity with covariates and morphospecies.

In this analysis, we demonstrate the proposed method on data collected in a multispecies bat acoustic monitoring study conducted in British Columbia, Canada between 2016 and 2020. Details of the study design and data are found in Stratton et al. (2022) and Stratton (2022). Briefly, one to six stationary acoustic recording devices were placed in $N = 55$ sites following the North American Bat Monitoring Program guidelines and were typically activated for seven nights (Loeb et al., 2015). Similar to Stratton et al. (2022), we analyze detections from the first and last nights to minimize potential overlap and dependencies, leading to $n_i = 2$ to 12 measurements for each site over the five

Common Name	Scientific Name	Mean (Variance)	% Zero
Big brown bat	*Eptesicus fuscus* (EPFU)	3.5 (256.0)	66
Hoary bat	*Lasiurus cinereus* (LACI)	3.1 (396.8)	65
Silver-haired bat	*Lasionycteris noctivagans* (LANO)	12.4 (2703.4)	34
California myotis	*Myotis californicus* (MYCA)	4.5 (264.3)	54
Western small-footed myotis	*Myotis ciliolabrum* (MYCI)	2.9 (265.4)	83
Western long-eared myotis	*Myotis evotis* (MYEV)	1.7 (28.2)	63
Little brown myotis	*Myotis lucifugus* (MYLU)	25.9 (4339.6)	26
Long-legged myotis	*Myotis volans* (MYVO)	2.7 (114.3)	59
Yuma myotis	*Myotis yumanensis* (MYYU)	5.0 (625.5)	69
Other	–	0.98 (25.7)	77

Table 3: Observed mean, variance, and % zero observations for each species in the bat monitoring study.

year period. There were $T = C = 10$ total bat species categories available for analysis, including an *other* category for species that were difficult to detect acoustically or that were not widespread. Each acoustic recording was classified using Kaleidoscope Pro acoustic classification software for bats (`https://www.wildlifeacoustics.com`). A unique attribute of these data is that each acoustic recording was additionally validated by a bat expert. For this analysis, we let half of all revisits from every site include validation data, similar to Stratton et al. (2022). Additionally, we included site-specific covariates, $\boldsymbol{x}_i$, measuring year (categorical with 2016 as the reference), annual mean elevation (kilometers), precipitation (millimeters), and temperature (degrees Celsius) for the occupancy (or at-risk) portion of the model. We included nightly minimum air temperature (degrees Celsius), total precipitation (millimeters), and percentage of the moon illuminated by the sun (percent) measured from the centroid of the site at each visit to model the encounter rates or relative abundances for the DMZIP and missZIDM models, respectively. All continuous covariates were standardized prior to analysis. In this analysis, the average number of observed individuals per site visit was 62.6, ranging from 1 to 992. The means, variances, and percent zero counts for each species across site visits are presented in Table 3. We observed mean counts ranging from 1.7 to 25.9 with variances ranging from 25.7 to 4339.6 and percent zeros ranging from 26% to 83%.

For inference, missZIDM was run for 10,000 iterations, treating the first 5,000 as burn-in and thinning to every other iteration. The model was initialized similar to scenario 2 of the simulation study. We assumed relatively weak or non-informative priors with $\nu_{tc} = 1$, $\gamma_t = 1$, $\mu_{\beta_\eta} = \mu_{\beta_\psi} = 0$, and $\sigma^2_\eta = \sigma^2_\psi = \sigma^2_\gamma = 1$. We compared the results of the proposed model to DMZIP with similar prior assumptions. Convergence and mixing of the models was visually inspected using traceplots. See Supplementary Material for traceplots for a random subset of the parameters (Supplementary Figures S1–S7). To further assess the convergence of the model, we ran another chain initialized with $\omega_{\tau_{tijl}} = \omega_{\zeta_{tij}} = \zeta_{tij} = \alpha_{tc} = 0.5$, u_t, $\mu_{ij} \sim$ Gamma(1,1), and regression coefficients sampled from a standard normal. We then compared the two chains with the Gelman-Rubin statistic, which was less than 1.1 for each parameter (Brooks and Gelman, 1998).

Figure 2 presents the estimated confusion matrix probabilities for the proposed missZIDM model, as well as the differences with the DMZIP model. The models found

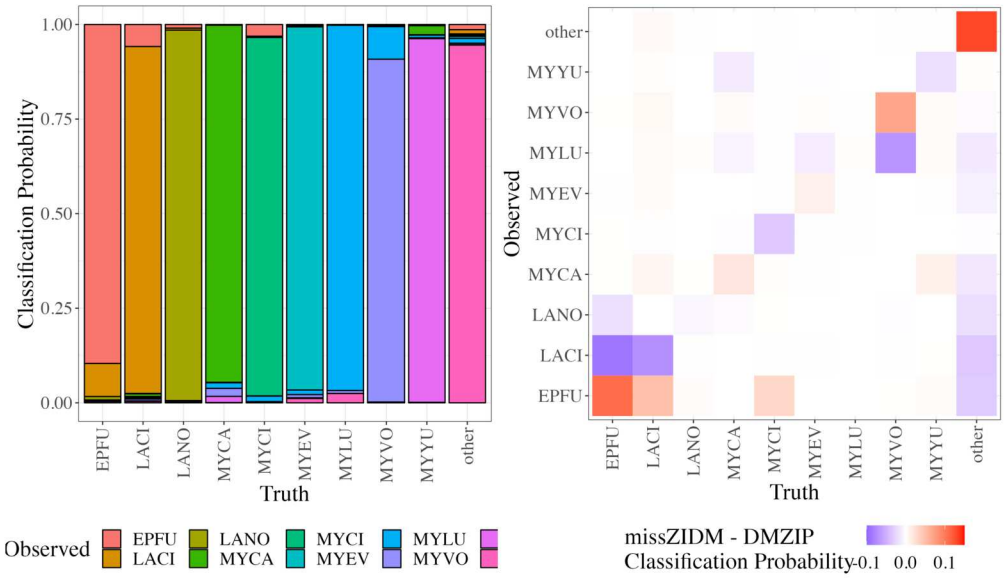

Figure 2: Bat Monitoring Application Results: The left subplot shows the estimated confusion matrix probabilities using the proposed missZIDM model. The right subplot presents a heatmap of the difference between the estimated confusion matrix probabilities of missZIDM and DMZIP.

relatively similar results, with the average (SD) absolute difference between all cells of the confusion matrix equal to 1% (2%). The largest differences in misclassification (1− diagonal elements of the confusion matrix) estimates were DMZIP estimating 10.1% more misclassification for EPFU, 7% more for MYVO, and 13% more for the *other* category. Additionally, the proposed method estimated 7% more misclassification for LACI and 3% more for MYYU compared to DMZIP. In the Supplementary Material, we provide tables for the estimated confusion matrix probabilities and corresponding 95% credible intervals for missZIDM and DMZIP (Supplementary Tables S7 and S8, respectively).

We investigated the estimated covariate associations for occupancy in both models. The estimates plotted in Figure 3 for $\boldsymbol{\beta}_\eta$ are interpreted as log odds ratios for occupancy at each site visit. Overall, the results were quite similar between the models. Neither method found a strong effect for time on occupancy. However, both models estimated a decrease in the odds of occupancy for EPFU in 2019 compared to 2016, and DMZIP estimated an increase in the odds of occupancy for LANO in 2020 compared to 2016. We found that an increase in elevation was associated with an increase in the odds of occupancy for most species, with the exception of MYVO. We observed mostly negative associations between precipitation and occupancy, although most of the 95% credible intervals contained 0 for both models. The strongest relation was for MYCI, where a millimeter increase in precipitation was associated with a 95% decrease in occupancy. EPFU, LACI, LANO, MYCA, MYCI, and MYYU were all found to

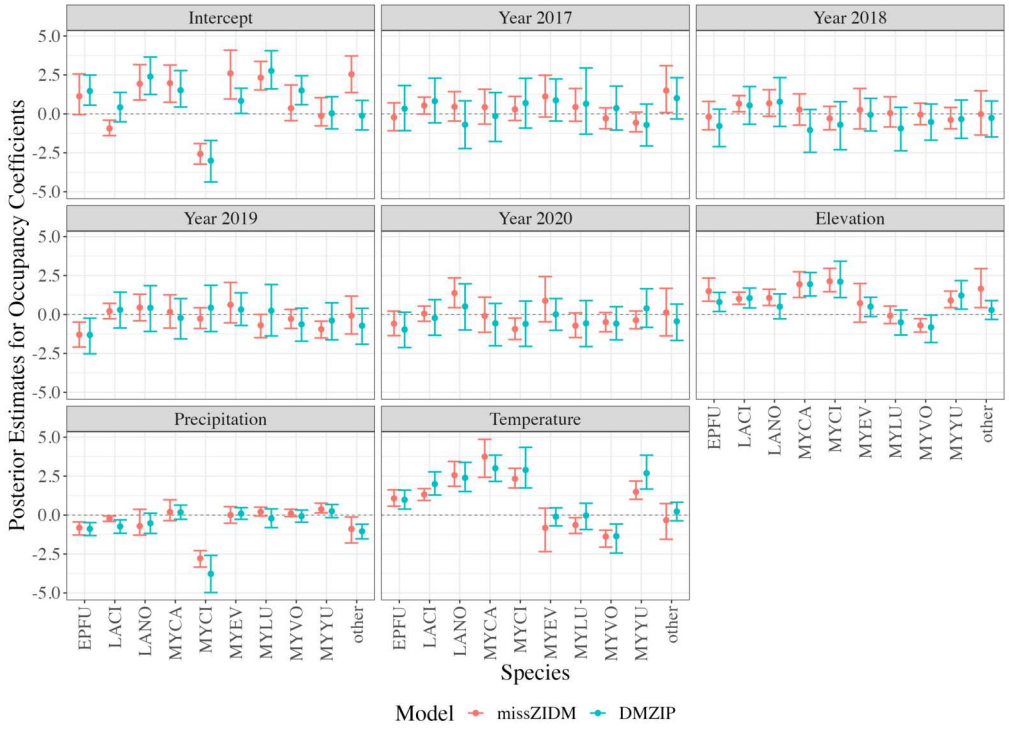

Figure 3: Bat Monitoring Application Results: Posterior estimates of regression coefficients for occupancy for the proposed missZIDM and DMZIP models. Dot represents the posterior mean with error bars representing the 95% credible intervals.

have positive associations between temperature and occupancy. However, temperature was negatively associated with occupancy for MYVO. Temperature has previously been found to be associated with occupancy for EPFU, LACI, and LANO in a study conducted in North Carolina during the winter season (Parker Jr et al., 2020). Typically the proposed method was more conservative than DMZIP with respect to parameter uncertainty for all covariate effects.

Figure 4 presents the estimated multiplicative effect for a one-unit increase in each covariate (i.e., temperature, precipitation, and illumination) on the relative abundance of each species using the proposed model given the sample average of the other covariates, π_{tp}. As described previously, the effect of a covariate on each species' relative abundance depends on its association with the other species' relative abundances. As such, a positive (negative) association between a covariate and a species' relative *activity* does not imply a positive (negative) association with the species' relative *abundance*. For example, precipitation was found to be negatively associated with a majority of the bat species' relative activity using the DMZIP model (Figure 5) as well as the concentration parameters of the proposed model (Supplementary Figure S8). However, precipitation was positively associated with 7/10 species' relative abundances, which

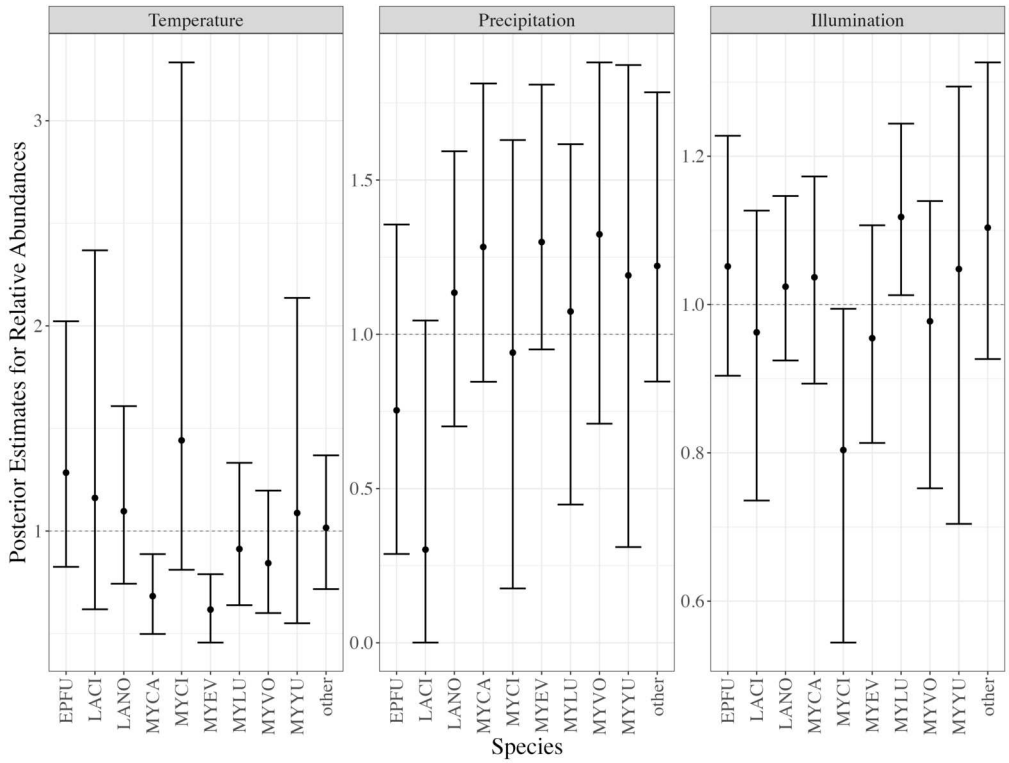

Figure 4: Bat Monitoring Application Results: Posterior estimates of the multiplicative effect for a one-unit increase in each covariate on the relative abundance of each species using the proposed model. Dot represents the posterior mean with error bars representing the 95% credible intervals.

reflects the differences in the effects of precipitation on bat activity across species. The overall negative relations observed between bat activity and precipitation in this analysis corroborate previous findings that attribute the reduction to various factors, including increased flight metabolism of wet bats, interference with echolocation, and availability of prey (Griffin, 1971; Burles et al., 2009; Voigt et al., 2011). Previous studies have shown that activity by insectivorous bats is sensitive to environmental conditions, with different effects observed for different species (Thies et al., 2006; Vásquez et al., 2020; Rodríguez-San Pedro et al., 2024). For example, bat species' response to moonlight intensity and temperature has been found to be species-specific (Saldaña-Vázquez and Munguía-Rosas, 2013; Klüg-Baerwald et al., 2016; Appel et al., 2017; Vásquez et al., 2020). We found mixed results regarding the associations between nightly minimum air temperature and moon illumination and relative abundances. We observed a negative association between temperature and the relative abundances of MYCA and MYEV. We also observed a negative association between moon illumination and the relative abundance of MYCI with missZIDM, in addition to its relative activity with DMZIP.

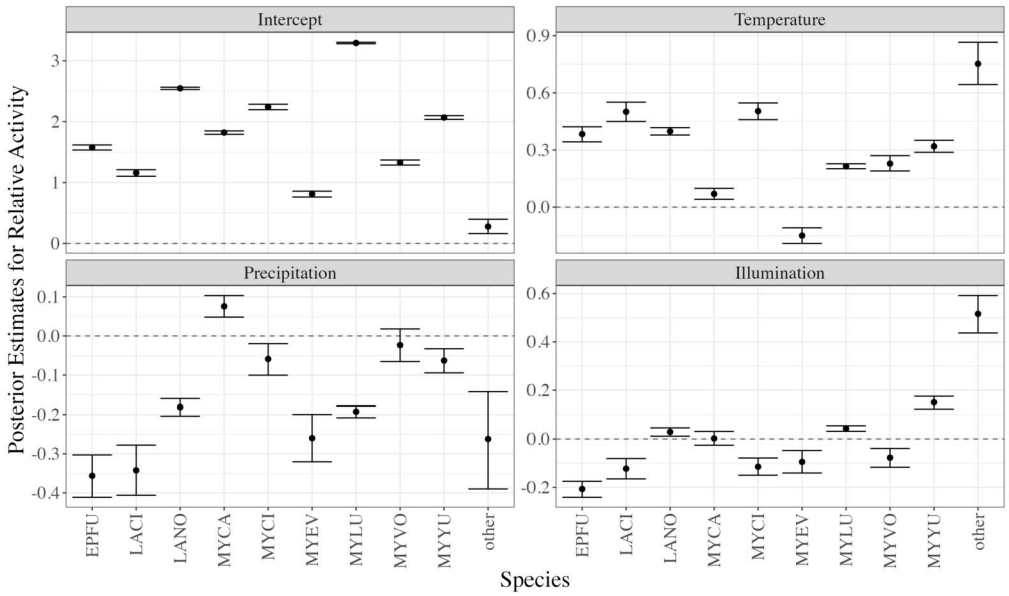

Figure 5: Bat Monitoring Application Results: Posterior estimates of regression coefficients for relative activity using the DMZIP model. Dot represents the posterior mean with error bars representing the 95% credible intervals.

One of the advantages of the proposed method is that it does not require making distributional assumptions for the latent ecological process to simultaneously model nondetection and misclassification in multispecies settings. By leveraging a combination of data augmentation techniques, the proposed method is able to scale to high-dimensional settings while additionally modeling the latent classifications and generating inference on the true relative abundances of each species at each site measurement. While missZIDM teases apart the task of modeling the true classifications and ecological process, it does not preclude embedding the proposed model into larger hierarchical frameworks aimed at simultaneously inferring the ecological process. This is similar to techniques used in integrated population modeling (Schaub and Abadi, 2011).

5.2 Application to Official Crime Data

In this analysis, we apply the proposed method to incident-level official crime data that include victim-involved offenses reported by Colorado law enforcement agencies to the National Incident-Based Reporting System (NIBRS) in 2022 (publicly available online; Bureau of Justice Statistics, 2023). Collected under the Uniform Crime Reporting Program (UCR), NIBRS data contain criminal incidents that are known to law enforcement and serve as the national standard for official crime reporting in the United States (Addington, 2019). Since the UCR inception in 1991, NIBRS data have continued to improve upon official criminal data monitoring by providing greater breadth and depth in

reporting information about criminal incidents, context about specific crime problems, and more capacity for data analysis about nation-wide crime trends.

One of the challenges of analyzing NIBRS crime data is the high occurrence of zero counts due to rarely occurring crimes (e.g., murder) (Rydberg and Carkin, 2017; Luo et al., 2022), data entry errors (Wheeler and Kovandzic, 2018), and the underreporting of criminal incidents (Skogan, 1974; Wormeli, 2018). For example, victim-involved criminal incidents including violent, domestic violence, or sexual assault offenses are particularly likely to go unreported to law enforcement (Langton et al., 2017). Another challenge of modeling NIBRS crime data is they are subject to misclassification by law enforcement agencies (Bibel, 2015). Common misclassification errors may occur when a criminal offense is coded incorrectly in comparison to qualitative incident information provided in the probable cause affidavit (Nolan et al., 2011). For example, a criminal incident may be incorrectly coded to include a simple assault offense instead of an aggravated assault when the presence of a weapon is detailed in the affidavit. Misclassification errors may also occur when the NIBRS coding category fails to include the necessary level of specificity for a certain type of offense circumstance within a criminal incident (Osborne et al., 2019; Haberman et al., 2022). For example, incidents that involve a robbery offense cannot be differentiated further by typology such as carjacking, bank robbery, or residential invasion. The amount of misclassification has been found to vary by crime type (Nolan et al., 2011) and can result in the over- and undercounting of certain crime categories. Without accurate reporting, criminal justice research using official data is vulnerable to potential biases introduced by law enforcement reporting practices that misrepresent true criminal phenomenon (Pina-Sánchez et al., 2023). Given the widespread use of NIBRS data to inform resource allocation, criminal justice policy, and the empirical understanding of crime phenomena, it is critical to account for potential classification errors to improve the accuracy of criminal incident reporting.

In this analysis, we demonstrate how the proposed method can be used to account for potential zero-inflation and misclassification when estimating crime incidents. We applied the proposed method to incident-level victim-involved crime data (i.e., murder, rape, sodomy, sexual assault with an object (SAO), fondling, robbery, aggravated assault, simple assault) reported by all Colorado agencies to NIBRS in 2022. These data capture 55,198 total reported incidents over $N = 216$ agencies, ranging from 318 incidences of murder to 29,747 incidents of simple assault. We observed 50% of the crime categories reported by law enforcement agencies were zero counts. Across the different categories, murder had the highest proportion of zero counts (i.e., 76%).

In the absence of validation data to inform misclassification probabilities, we evaluated the model using three different prior specifications. In the first setting (*Naive*), we assume that the probability of an at-risk observation for a given crime, η_t, is 0.50, there is a 0.05 probability of misclassification for each crime (i.e., $\psi_t = 0.05$), and each crime is equally likely to be misclassified as another crime with a small probability of a lucky guess (i.e., $\nu_{tc} = 1$ with $\nu_{tt} = 0.001$). Without knowledge regarding the amount of misclassification in the data, it is difficult to specify the concentration hyperparameters for the true relative abundances. Thus in the *Naive* setting, we take an empirical Bayes approach and set the concentration hyperparameters to the log of the average observed relative abundances across agencies scaled by 1,000.

The assumption that each crime is equally likely to be misclassified as another crime is somewhat unrealistic as some crimes are more similar than others. For example, a simple assault is more likely to be misclassified as an aggravated assault than a homicide. Therefore in a second setting (*Blocked*), we formed three groups of victim-involved crimes including murder, sex offenses (i.e., rape, sodomy, SAO, and fondling), and other violent crimes (i.e., robbery, aggravated assault, and simple assault) within which misclassification was more likely. For implementation, we assumed a block matrix for $\boldsymbol{\nu}$ with $\nu_{tc} = 1{,}000$ for the off-diagonal elements within each block and $\nu_{tc} = 1\text{e}{-5}$, otherwise. The other hyperparameters were set similar to the *Naive* model.

In a third setting, we used historical data from Nolan et al. (2011) to inform the model (*Historical*). Previously, Nolan et al. (2011) performed a validation study on 3 of the 12 largest municipal police agencies in a mostly rural southeastern state for 15 crime categories using Uniform Crime Reporting (UCR) data collected in 2002. As this information may not fully represent those found in our application data, we recommend validating the results obtained from this analysis prior to generalization. In the validation data, Nolan et al. (2011) found that 21.6% of the rape incidents, 7.3% of the other sex offenses, 5.0% of the robberies, 8.4% of the aggravated assaults, and 1.8% of the simple assaults were misclassified. We used these misclassification probabilities to specify ψ_t in our model, with 7.3% misclassification assumed for each of the sex offense incidents analyzed. Because there were no misclassifications found in the validation set for murder, we assumed a small probability of error for analysis (i.e., 0.0001). Additionally, we set the concentration hyperparameters for the true classifications to the log of the "statistically adjusted" relative abundances from Nolan et al. (2011) scaled by 1,000. We assumed a similar block matrix for $\boldsymbol{\nu}$ with $\nu_{tc} = 1{,}000$ for the off-diagonal elements within each block, $\nu_{tt} = 0.001$, and $\nu_{tc} = 1\text{e}{-8}$, otherwise. The remaining priors were specified similar to the other settings.

In each setting, the MCMC algorithm was run for 200,000 iterations, thinning to every 25th iteration and treating the first 2,000 as burn-in, leaving 2,000 iterations for inference. Estimated true classifications were obtained using the `salso` method with Binder's loss (Dahl et al., 2022). Convergence and mixing of the models was visually inspected using traceplots. See Supplementary Material for traceplots for a random subset of the parameters (Supplementary Figures S9–S12).

Table 4 reports the estimated mean incidence with 95% credible intervals for each crime using the proposed method. Under the assumptions of the *Naive* model, the misclassified incidents from the smaller crime categories gravitated towards the largest category, simple assault. Overall, we observed the estimated crime incidents for the *Blocked* model were the most similar to the observed classifications. These results were expected as the prior specification in this setting aligns closely to the observed data. The *Naive* model estimated the highest number of misclassifications (2,160), compared to 817 and 1,529 for the *Blocked* and *Historical* models, respectively (Figure 6). With the *Blocked* and *Historical* models, no misclassifications were observed outside of the blocked structure, whereas the *Naive* model estimated 134 misclassifications. In all settings, the majority of misclassifications estimated were for true simple assaults observed as aggravated assaults.

	Murder	Rape	SAO	Sodomy	Fondling	Robbery	Aggravated Assault	Simple Assault
Observed	318	2245	620	579	2340	4215	15134	29747
Naive	260.2	2156.7	546.1	579.8	2273.8	3998.7	14195.0	31187.7
	(247, 277)	(2125, 2194)	(522, 580)	(566, 598)	(2193, 2334)	(3950, 4050)	(14119, 14269)	(31107, 31275)
Blocked	313.6	2236.1	622.6	584.6	2340.7	4268.0	15385.8	29446.6
	(257, 318)	(2230, 2242)	(617, 628)	(579, 591)	(2334, 2346)	(4249, 4288)	(15337, 15448)	(29381, 29493)
Historical	344.8	2169.7	644.9	590.0	2352.6	4290.9	14073.7	30731.5
	(336, 354)	(2160, 2178)	(635, 656)	(582, 599)	(2341, 2363)	(4271, 4312)	(13975, 14184)	(30623, 30832)

Table 4: Crime Application Results: Observed and estimated mean incidents and corresponding 95% credible intervals (below in parentheses) obtained with the proposed model using different prior formulations.

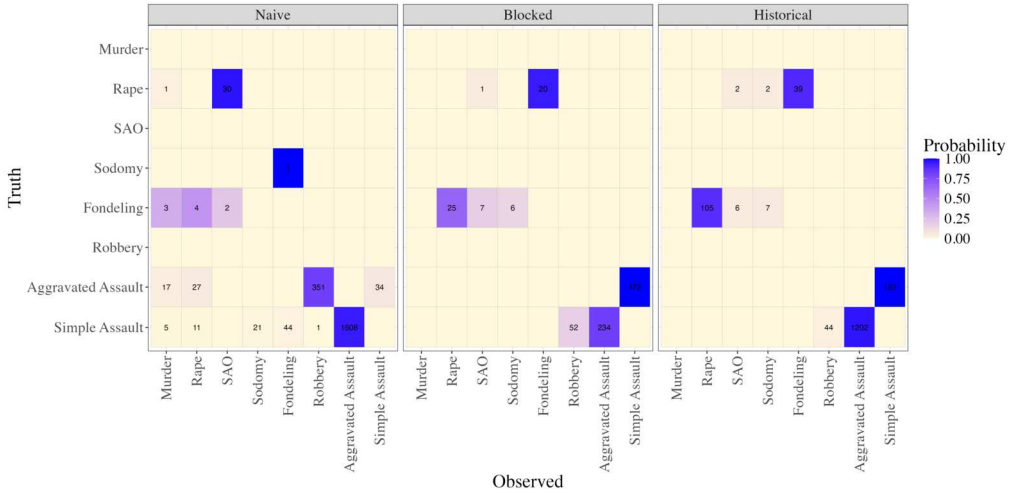

Figure 6: Crime Application Results: Probability of observed classification given the true classification estimated using the three modeling assumptions. Note that each row sums to one, and the counts represent the number of misclassified incidents.

In this analysis, we demonstrate how the proposed method can be used to account for potential biases in the estimation of criminal incidents attributable to zero-inflation in and misclassification of official crime data. While we illustrated how prior knowledge and historical data can be used to inform the model, using non-representative or inaccurate information may do more harm than good mitigating biases in official crime data. However, specifying informative priors based on historical data or expert knowledge may still be preferred over naive prior assumptions when appropriate. For example in our application, it is reasonable to justify that murders are rarely misclassified. Whereas the true crime classification probabilities obtained from the historical data collected in a mostly rural southeastern state in 2002 may not be fully representative of the crime patterns found in Colorado in 2022. In practice, we recommend collecting validation data for a subset of the data analyzed to account for spatial and temporal variation in misclassification rates.

6 Conclusions

In this work, we propose the first method for simultaneously accommodating false positives and false negatives in multinomial data. Our model can naturally incorporate existing knowledge of zero-inflation or misclassification through prior specification and/or easily accommodate validation data to inform the model. In simulation, we demonstrated that our approach obtains similar or improved estimation performance for at-risk, true classification, and misclassification probabilities compared to alternative methods that ignore one or both forms of measurement error. We further show how the proposed approach can provide more accurate estimation for at-risk and classification

probabilities than existing multivariate methods in the presence of overdispersed data because it does not require accurately specifying the latent count process. By simultaneously modeling misclassification at the individual level with potential non-detection, our approach accommodates the uncertainty of non-detection occurring if the species is present but not observed at each site visit and if each of the individuals observed for a particular species at a given site are incorrectly classified. Conceptually, the method of Wright et al. (2020) and Spiers et al. (2022) could be adjusted to accommodate potential overdispersion by replacing the Poisson distribution with a negative binomial distribution. This approach serves as a potentially viable alternative for modeling misclassification in overdispersed multivariate count data, but it would not be appropriate for multinomial or compositional data settings where the total number of counts is fixed.

While demonstrated in ecological and criminal justice research settings, the proposed method is applicable to other settings in which zero-inflated multinomial data with potential misclassification are collected. For example, the method could be used to model zero-inflated multivariate count data collected in human microbiome research settings that are subject to measurement error introduced at various stages of the measurement protocol (Pollock et al., 2018; Clausen and Willis, 2022). Additionally, citizen science and crowdsourcing projects often task contributors with classifying different features, such as radio technosignatures to help detect extraterrestrial life (Margot et al., 2019), types of stars based on their spectra (DeLisle and Barker, 2024), as well as images, videos, sounds, water samples, and/or sensor data for biodiversity research, Earth observation, and geography and climate change research (Schmidt et al., 2013; Pocock et al., 2014; Ficetola et al., 2016; Lahoz-Monfort et al., 2016; Willoughby et al., 2016; Fraisl et al., 2022). In the context of conservation research and wildlife monitoring studies, the proposed method could be customized to answer pressing research questions. For example, to accommodate occupancy dynamics, one could assume the probability a site is in a given occupancy state is governed by a Markov process, similar to Miller et al. (2013). In the second application study, we investigated official data of criminal incidents reported by law enforcement agencies. However, it is well known that there are discrepancies between official, victim- and self-reported crime data. In future work, our proposed method could be used to evaluate potential biases in official data reported across different sources by accommodating multiple measurements for each site, or in this case, jurisdiction, and covariate information for true relative abundances.

Acknowledgments

The opinions, findings, and conclusions expressed are those of the authors and do not necessarily reflect the views of the NSF. The authors thank Dr. Kathi Irvine for providing useful discussions about modeling bat acoustic data.

Funding

MDK gratefully acknowledges the support of NSF grant DMS-2245492. AK gratefully acknowledges the support of NSF grants DMS-2330089 and SES-2338428.

Supplementary Material

Supplementary Material (DOI: 10.1214/24-BA1477SUPPA; .pdf). The Supplementary Material contains detailed derivations of the MCMC algorithm, sensitivity analysis, and additional tables and figures.

missZIDM `R` package (DOI: 10.1214/24-BA1477SUPPB; .zip). This file contains code to generate data similar to the simulation study as well as a vignette demonstrating how to apply the method and perform inference (Koslovsky et al., 2024b).

References

Addington, L. A. (2019). NIBRS as the new normal: What fully incident-based crime data mean for researchers. *Handbook on Crime and Deviance*, pages 21–33. doi: https://doi.org/10.1007/978-3-030-20779-3_2. 18

Aitchison, J. and Ho, C. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76(4):643–653. MR1041409. doi: https://doi.org/10.1093/biomet/76.4.643. 2

Appel, G., López-Baucells, A., Ernest-Magnusson, W., and Bobrowiec, P. E. D. (2017). Aerial insectivorous bat activity in relation to moonlight intensity. *Mammalian Biology*, 85:37–46. doi: https://doi.org/10.1016/j.mambio.2016.11.005. 17

Bayley, P. B. and Peterson, J. T. (2001). An approach to estimate probability of presence and richness of fish species. *Transactions of the American Fisheries Society*, 130(4):620–633. doi: https://doi.org/10.1577/1548-8659(2001)130<0620:AATEPO>2.0.CO;2. 13

Bibel, D. (2015). Considerations and cautions regarding NIBRS data: A view from the field. *Justice Research and Policy*, 16(2):185–194. doi: https://doi.org/10.1177/1525107115623943. 19

Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., and Castells, E. (2019). What does a zero mean? Understanding false, random and structural zeros in ecology. *Methods in Ecology and Evolution*, 10(7):949–959. doi: https://doi.org/10.1111/2041-210X.13185. 13

Broms, K. M., Hooten, M. B., and Fitzpatrick, R. M. (2015). Accounting for imperfect detection in Hill numbers for biodiversity studies. *Methods in Ecology and Evolution*, 6(1):99–108. doi: https://doi.org/10.1111/2041-210X.12296. 13

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455. MR1665662. doi: https://doi.org/10.2307/1390675. 14

Bureau of Justice Statistics (2023). National Incident-Based Reporting System, 2022: Extract Files. Inter-university Consortium for Political and Social Research. doi: https://doi.org/10.3886/ICPSR38925.v1.. Accessed: 2024-06-22. 18

Burles, D., Brigham, R., Ring, R., and Reimchen, T. (2009). Influence of weather on two insectivorous bats in a temperate Pacific Northwest rainforest. *Canadian Journal of Zoology*, 87(2):132–138. doi: https://doi.org/10.1139/Z08-146. 17

Chambert, T., Grant, E. H. C., Miller, D. A., Nichols, J. D., Mulder, K. P., and Brand, A. B. (2018a). Two-species occupancy modelling accounting for species misidentification and non-detection. *Methods in Ecology and Evolution*, 9(6):1468–1477. doi: https://doi.org/10.1111/2041-210X.12985. 13

Chambert, T., Miller, D. A., and Nichols, J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology*, 96(2):332–339. doi: https://doi.org/10.1890/14-1507.1. 3, 13

Chambert, T., Waddle, J. H., Miller, D. A., Walls, S. C., and Nichols, J. D. (2018b). A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing. *Methods in Ecology and Evolution*, 9(3):560–570. doi: https://doi.org/10.1111/2041-210X.12910. 13

Chiquet, J., Mariadassou, M., and Robin, S. (2021). The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9:188. 2

Clausen, D. S. and Willis, A. D. (2022). Evaluating replicability in microbiome data. *Biostatistics*, 23(4):1099–1114. MR4496370. doi: https://doi.org/10.1093/biostatistics/kxab048. 23

Copas, J. and Hilton, F. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 153(3):287–312. doi: https://doi.org/10.2307/2982975. 2

Dahl, D. B., Johnson, D. J., and Müller, P. (2022). Search algorithms and loss functions for Bayesian clustering. *Journal of Computational and Graphical Statistics*, 31(4):1189–1201. MR4513380. doi: https://doi.org/10.1080/10618600.2022.2069779. 20

Dai, Z., Wong, S. H., Yu, J., and Wei, Y. (2019). Batch effects correction for microbiome data with Dirichlet-multinomial regression. *Bioinformatics*, 35(5):807–814. doi: https://doi.org/10.1093/bioinformatics/bty729. 6

Datta, A., Fiksel, J., Amouzou, A., and Zeger, S. L. (2021). Regularized Bayesian transfer learning for population-level etiological distributions. *Biostatistics*, 22(4):836–857. MR4325730. doi: https://doi.org/10.1093/biostatistics/kxaa001. 1

DeLisle, T. and Barker, T. (2024). SCOPE stellar classification online public exploration. http://scope.pari.edu/. Accessed: 2024-03-24. 23

Devarajan, K., Morelli, T. L., and Tenan, S. (2020). Multi-species occupancy models: Review, roadmap, and recommendations. *Ecography*, 43(11):1612–1624. doi: https://doi.org/10.1111/ecog.04957. 13

Dorazio, R. M., Gotelli, N. J., and Ellison, A. M. (2011). Modern methods of estimating

biodiversity from presence-absence surveys. *Biodiversity Loss in a Changing Planet*, pages 277–302. 13

Dorazio, R. M., Royle, J. A., Söderström, B., and Glimskär, A. (2006). Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, 87(4):842–854. doi: https://doi.org/10.1890/0012-9658(2006)87[842:ESRAAB]2.0.CO;2. 13

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40:1–18. doi: https://doi.org/10.18637/jss.v040.i08. 8

Ficetola, G., Taberlet, P., and Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? *Molecular Ecology Resources*, 16(3):604–607. doi: https://doi.org/10.1111/1755-0998.12508. 23

Fraisl, D., Hager, G., Bedessem, B., Gold, M., Hsing, P.-Y., Danielsen, F., Hitchcock, C. B., Hulbert, J. M., Piera, J., Spiers, H., et al. (2022). Citizen science in environmental and ecological sciences. *Nature Reviews Methods Primers*, 2(1):64. 23

Frénay, B. and Verleysen, M. (2013). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869. MR4378310. doi: https://doi.org/10.1109/tnnls.2021.3070843. 2

Griffin, D. R. (1971). The importance of atmospheric attenuation for the echolocation of bats (chiroptera). *Animal Behaviour*, 19(1):55–61. doi: https://doi.org/10.1016/S0003-3472(71)80134-3. 17

Guillera-Arroita, G., Lahoz-Monfort, J. J., van Rooyen, A. R., Weeks, A. R., and Tingley, R. (2017). Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods in Ecology and Evolution*, 8(9):1081–1091. doi: https://doi.org/10.1111/2041-210X.12743. 3

Haberman, C. P., Clutter, J. E., and Lee, H. (2022). A robbery is a robbery is a robbery? Exploring crime specificity in official police incident data. *Police Practice and Research*, 23(4):429–443. doi: https://doi.org/10.1080/15614263.2021.2009345. 19

Hoeting, J. A., Leecaster, M., and Bowden, D. (2000). An improved model for spatially correlated binary responses. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 102–114. MR1817027. doi: https://doi.org/10.2307/1400634. 13

Jasra, A., Holmes, C., and Stephens, D. (2005). Markov chain Monte Carlo Methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67. MR2182987. doi: https://doi.org/10.1214/088342305000000016. 3

Jiang, S., Xiao, G., Koh, A. Y., Kim, J., Li, Q., and Zhan, X. (2021). A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics*, 22(3):522–540. MR4287166. doi: https://doi.org/10.1093/biostatistics/kxz050. 2

Kellner, K. F. and Swihart, R. K. (2014). Accounting for imperfect detection in ecology: A quantitative review. *PloS One*, 9(10):e111436. doi: https://doi.org/10.1371/journal.pone.0111436. 12

Klüg-Baerwald, B. J., Gower, L. E., Lausen, C., and Brigham, R. (2016). Environmental correlates and energetics of winter flight by bats in southern Alberta, Canada. *Canadian Journal of Zoology*, 94(12):829–836. doi: https://doi.org/10.1139/cjz-2016-0055. 17

Koslovsky, M. D. (2023). A Bayesian zero-inflated Dirichlet-multinomial regression model for multivariate compositional count data. *Biometrics*. MR4680718. doi: https://doi.org/10.1111/biom.13853. 2, 5, 6, 8

Koslovsky, M. D., Hoffman, K. L., Daniel, C. R., and Vannucci, M. (2020). A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. *The Annals of Applied Statistics*, 14(3):1471–1492. MR4152142. doi: https://doi.org/10.1214/20-AOAS1354. 6

Koslovsky, M. D., Kaplan, A., Terranova, V. A. and Hooten, M. B. (2024a). Supplementary Material for "A unified Bayesian framework for modeling measurement error in multinomial data". *Bayesian Analysis*. doi: https://doi.org/10.1214/24-BA1477SUPPA. 6

Koslovsky, M. D., Kaplan, A., Terranova, V. A. and Hooten, M. B. (2024b). Supplementary Material for "A unified Bayesian framework for modeling measurement error in multinomial data". *Bayesian Analysis*. doi: https://doi.org/10.1214/24-BA1477SUPPB. 24

Lahoz-Monfort, J. J., Guillera-Arroita, G., and Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. *Molecular Ecology Resources*, 16(3):673–685. doi: https://doi.org/10.1111/1755-0998.12486. 23

Langton, L., Planty, M., and Lynch, J. P. (2017). Second major redesign of the National Crime Victimization Survey (NCVS). *Criminology & Public Policy*, 16:1049. doi: https://doi.org/10.1111/1745-9133.12335. 19

Lele, S. R., Moreno, M., and Bayne, E. (2012). Dealing with detection error in site occupancy surveys: What can we do with a single survey? *Journal of Plant Ecology*, 5(1):22–31. doi: https://doi.org/10.1093/jpe/rtr042. 4

Loeb, S., Rodhouse, T., Ellison, L., Lausen, C., Reichard, J., Irvine, K., Ingersoll, T., Coleman, J., Thogmartin, W., Sauer, J., et al. (2015). A plan for the North American Bat Monitoring Program (NABat). *General Technical Report-Southern Research Station, USDA Forest Service*. 13

Luo, L., Deng, M., Shi, Y., Gao, S., and Liu, B. (2022). Associating street crime incidences with geographical environment in space using a zero-inflated negative binomial regression model. *Cities*, 129:103834. 19

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., and Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species

is detected imperfectly. *Ecology*, 84(8):2200–2207. doi: https://doi.org/10.1890/02-3090. 13

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., and Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255. doi: https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2. 4, 13

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L., and Hines, J. E. (2017). *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Elsevier. 13

Margot, J.-L., Croft, S., Lazio, J., Tarter, J., and Korpela, E. (2019). The radio search for technosignatures in the decade 2020–2030. *Bulletin of the American Astronomical Society*, 51(3):298. 23

McClintock, B. T., Bailey, L. L., Pollock, K. H., and Simons, T. R. (2010). Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology*, 91(8):2446–2454. doi: https://doi.org/10.1890/09-1287.1. 13

Miller, D. A., Nichols, J. D., Gude, J. A., Rich, L. N., Podruzny, K. M., Hines, J. E., and Mitchell, M. S. (2013). Determining occurrence dynamics when false positives occur: Estimating the range dynamics of wolves from public survey data. *PLoS One*, 8(6):e65808. doi: https://doi.org/10.1371/journal.pone.0065808. 23

Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., and Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, 92(7):1422–1428. doi: https://doi.org/10.1890/10-1396.1. 13

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, 144(1):81–117. MR2439923. doi: https://doi.org/10.1016/j.jeconom.2007.12.003. 1

Mulick, A. R., Oza, S., Prieto-Merino, D., Villavicencio, F., Cousens, S., and Perin, J. (2022). A Bayesian hierarchical model with integrated covariate selection and misclassification matrices to estimate neonatal and child causes of death. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(4):2097–2120. MR4537808. doi: https://doi.org/10.1111/rssa.12853. 1

Neelon, B. (2019). Bayesian zero-inflated negative binomial regression based on Pólya-Gamma mixtures. *Bayesian Analysis*, 14(3):829. MR3960773. doi: https://doi.org/10.1214/18-BA1132. 2

Nolan, J. J., Haas, S. M., and Napier, J. S. (2011). Estimating the impact of classification error on the "statistical accuracy"' of uniform crime reports. *Journal of Quantitative Criminology*, 27:497–519. doi: https://doi.org/10.1007/s10940-011-9135-9. 19, 20

Osborne, D. L., Swartz, K., and Stover, A. (2019). Utilizing the national incident-based reporting system to further our understanding of agricultural theft. *International*

Journal of Rural Criminology, 4(2):240–257. doi: https://doi.org/10.18061/1811/87908. 19

Parker Jr, K. A., Li, H., and Kalcounis-Rueppell, M. C. (2020). Species-specific environmental conditions for winter bat acoustic activity in North Carolina, United States. *Journal of Mammalogy*, 101(6):1502–1512. doi: https://doi.org/10.1093/jmammal/gyaa133. 16

Pérez, C. J., Girón, F. J., Martín, J., Ruiz, M., and Rojano, C. (2007). Misclassified multinomial data: A Bayesian approach. *RACSAM*, 101(1):71–80. MR2324581. 1, 2

Pina-Sánchez, J., Buil-Gil, D., Brunton-Smith, I., and Cernat, A. (2023). The impact of measurement error in regression models using police recorded crime rates. *Journal of Quantitative Criminology*, 39(4):975–1002. doi: https://doi.org/10.1007/s10940-022-09557-6. 19

Pocock, M. J., Chapman, D. S., Sheppard, L. J., and Roy, H. E. (2014). *Choosing and Using Citizen Science: a guide to when and how to use citizen science to monitor biodiversity and the environment*. NERC/Centre for Ecology & Hydrology. 23

Pollock, J., Glendinning, L., Wisedchanwet, T., and Watson, M. (2018). The madness of microbiome: Attempting to find consensus "best practice" for 16s microbiome studies. *Applied and Environmental Microbiology*, 84(7):e02627–17. doi: https://doi.org/10.1128/AEM.02627-17. 23

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 6

Rodríguez-San Pedro, A., Allendes, J. L., Bruna, T., and Grez, A. A. (2024). Species-specific responses of insectivorous bats to weather conditions in central Chile. *Animals*, 14(6):860. 17

Royle, J. A. and Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87(4):835–841. doi: https://doi.org/10.1890/0012-9658(2006)87[835:GSOMAF]2.0.CO;2. 13

Royle, J. A. and Nichols, J. D. (2003). Estimating abundance from repeated presence–absence data or point counts. *Ecology*, 84(3):777–790. doi: https://doi.org/10.1890/0012-9658(2003)084[0777:EAFRPA]2.0.CO;2. 13

Ruiz-Gutierrez, V., Hooten, M. B., and Campbell Grant, E. H. (2016). Uncertainty in biological monitoring: A framework for data collection and analysis to account for multiple sources of sampling bias. *Methods in Ecology and Evolution*, 7(8):900–909. doi: https://doi.org/10.1111/2041-210X.12542. 13

Rydberg, J. and Carkin, D. M. (2017). Utilizing alternate models for analyzing count outcomes. *Crime & Delinquency*, 63(1):61–76. doi: https://doi.org/10.1177/0011128716678848. 19

Saldaña-Vázquez, R. A. and Munguía-Rosas, M. A. (2013). Lunar phobia in bats

and its ecological correlates: a meta-analysis. *Mammalian Biology*, 78(3):216–219. doi: https://doi.org/10.1016/j.mambio.2012.08.004. 17

Scharf, H. R., Lu, X., Williams, P. J., and Hooten, M. B. (2022). Constructing flexible, identifiable and interpretable statistical models for binary data. *International Statistical Review*, 90(2):328–345. MR4481438. doi: https://doi.org/10.1111/insr.12485. 13

Schaub, M. and Abadi, F. (2011). Integrated population models: a novel analysis framework for deeper insights into population dynamics. *Journal of Ornithology*, 152:227–237. doi: https://doi.org/10.1007/s10336-010-0632-7. 18

Schmidt, B. R., Kéry, M., Ursenbacher, S., Hyman, O. J., and Collins, J. P. (2013). Site occupancy models in the analysis of environmental DNA presence/absence surveys: A case study of an emerging amphibian pathogen. *Methods in Ecology and Evolution*, 4(7):646–653. doi: https://doi.org/10.1111/2041-210X.12052. 23

Shuler, K., Verbanic, S., Chen, I. A., and Lee, J. (2021). A Bayesian nonparametric analysis for zero-inflated multivariate count data with application to microbiome study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(4):961–979. MR4318016. doi: https://doi.org/10.1111/rssc.12493. 2

Skogan, W. G. (1974). The validity of official crime statistics: An empirical investigation. *Social Science Quarterly*, pages 25–38. 19

Spiers, A. I., Royle, J. A., Torrens, C. L., and Joseph, M. B. (2022). Estimating species misclassification with occupancy dynamics and encounter rates: A semi-supervised, individual-level approach. *Methods in Ecology and Evolution*, 13(7):1528–1539. doi: https://doi.org/10.1111/2041-210X.13858. 4, 5, 13, 23

Steorts, R. C., Hall, R., and Fienberg, S. E. (2016). A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516):1660–1672. MR3601725. doi: https://doi.org/10.1080/01621459.2015.1105807. 2

Stratton, C. (2022). Strattonch/CoupledUncoupled: Coupling validation effort manuscript release (v1.0.0). Zenodo. doi: https://doi.org/10.5281/zenodo.6040068.. 8, 13

Stratton, C., Irvine, K. M., Banner, K. M., Wright, W. J., Lausen, C., and Rae, J. (2022). Coupling validation effort with in situ bioacoustic data improves estimating relative activity and occupancy for multiple species with cross-species misclassifications. *Methods in Ecology and Evolution*, 13(6):1288–1303. doi: https://doi.org/10.1111/2041-210X.13831. 3, 9, 13, 14

Swartz, T. B., Haitovsky, Y., Vexler, A., and Yang, T. Y. (2004). Bayesian identifiability and misclassification in multinomial data. *Canadian Journal of Statistics*, 32(3):285–302. MR2101757. doi: https://doi.org/10.2307/3315930. 1, 2, 3, 8

Tancredi, A. and Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5(2B):1553–1585. MR2849786. doi: https://doi.org/10.1214/10-AOAS447. 2

Thies, W., Kalko, E. K., and Schnitzler, H.-U. (2006). Influence of environment and resource availability on activity patterns of *Carollia castanea* (Phyllostomidae) in Panama. *Journal of Mammalogy*, 87(2):331–338. 17

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., and Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. *Ecological Applications*, 13(6):1790–1801. doi: https://doi.org/10.1890/02-5078. 13

Vásquez, D. A., Grez, A. A., and Rodríguez-San Pedro, A. (2020). Species-specific effects of moonlight on insectivorous bat activity in central Chile. *Journal of Mammalogy*, 101(5):1356–1363. doi: https://doi.org/10.1093/jmammal/gyaa095. 17

Voigt, C. C., Schneeberger, K., Voigt-Heucke, S. L., and Lewanzik, D. (2011). Rain increases the energy cost of bat flight. *Biology Letters*, 7(5):793–795. doi: https://doi.org/10.1098/rsbl.2011.0313. 17

Wadsworth, W. D., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, S. A., and Vannucci, M. (2017). An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics*, 18(1):94. doi: https://doi.org/10.1186/s12859-017-1516-0. 4

Wang, S., Wang, L., and Swartz, T. B. (2020). Inference for misclassified multinomial data with covariates. *Canadian Journal of Statistics*, 48(4):655–669. MR4187763. doi: https://doi.org/10.1002/cjs.11556. 2

Wheeler, A. P. and Kovandzic, T. V. (2018). Monitoring volatile homicide trends across US cities. *Homicide Studies*, 22(2):119–144. doi: https://doi.org/10.1177/1088767917740171. 19

Willoughby, J. R., Wijayawardena, B. K., Sundaram, M., Swihart, R. K., and DeWoody, J. A. (2016). The importance of including imperfect detection models in eDNA experimental design. *Molecular Ecology Resources*, 4(16):837–844. doi: https://doi.org/10.1111/1755-0998.12531. 23

Wormeli, P. (2018). Criminal justice statistics — An evolution. *Criminology & Public Policy*, 17(2):483–496. doi: https://doi.org/10.1111/1745-9133.12369. 19

Wright, W. J., Irvine, K. M., Almberg, E. S., and Litt, A. R. (2020). Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. *Methods in Ecology and Evolution*, 11(1):71–81. doi: https://doi.org/10.1111/2041-210X.13315. 3, 4, 8, 13, 23

Xu, L., Paterson, A. D., Turpin, W., and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PloS One*, 10(7):e0129606. doi: https://doi.org/10.1371/journal.pone.0129606. 2

Zhang, X. and Yi, N. (2020). NBZIMM: Negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. *BMC Bioinformatics*, 21(1):1–19. doi: https://doi.org/10.1186/s12859-020-03922-7. 2