



A Bayesian approach to estimate transition models for zero-inflated count data

Qusai Shakhatreh^{id}, Ayat Almomani^{id*}, Mohammed Obeidat^{id}

Department of Statistics, Yarmouk University, Irbid 21163, Jordan

Abstract

Zero-inflated count data are commonly modelled by mixing a degenerate distribution at zero with a Poisson or negative binomial component, thereby imposing a parametric form on the counts. The transition model of Berger and Tutz instead specifies the conditional probability of moving from one count to the next and treats excess zeros within a sequence of transition probabilities. We develop a Bayesian formulation of this model, derive its full conditional distributions, and address the treatment of probability above the largest represented category under finite parameterisation. The support is closed using a terminal category, and sensitivity checks show that two natural closure conventions give similar results for the data considered. We compare the Bayesian transition model with Bayesian zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) regression across three simulated mechanisms and one real dataset. Because the models' coefficients represent different functionals, comparison is restricted to prediction on the expected-count scale. In all simulated scenarios, the transition model performs at least as well as the alternatives, with pairwise differences smaller than replication variability. In the real-data application, it gives lower prediction errors on all reported criteria, including median absolute error, with the largest advantage for observations beyond the training covariate range. The results indicate that the transition model can provide competitive prediction without specifying a fixed count distribution, while its bounded support remains a limitation for counts above the largest represented category.

DOI: [10.46481/jnsps.2027.3720](https://doi.org/10.46481/jnsps.2027.3720)

Keywords: Bayesian inference, Zero-inflated models, Transition models, Count data, Simulation study

Article History:

Received: 24 July 2026

Received in revised form: 15 August 2026

Accepted for publication: 16 September 2026

Available online: 30 September 2026

© 2027 The Author(s). Published by the [Nigerian Society of Physical Sciences](#) under the terms of the [Creative Commons Attribution 4.0 International license](#). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Communicated by: B. J. Falaye

1. Introduction

Count data arise frequently in multiple scientific fields, such as medicine, industry, economics, and social sciences. The Poisson distribution is often the starting point for modeling count data [1]. It is commonly used to analyze such data under the assumption that the mean and variance of the count distribution are equal. However, real count data do not meet these conditions; instead, the data have over-dispersion problems (variance exceeding the mean) and a high proportion of zeros [2].

Zero inflation occurs when the number of observed zeros exceeds the expected zeros in the standard count models, such as the Poisson or negative binomial distributions. If these issues are not taken into account, they can lead to biased estimates and invalid inferences [3, 4]. Additionally, predictive performance may be affected because models often underestimate the probability of zero responses and overestimate low counts [5]. Failing to account for zero inflation leads to the misinterpretation of results. This usually happens in studies where the occurrence of zeros carries substantive meaning, such as healthcare, criminology, or environmental studies; see also [6].

The Poisson regression model is used instead of the linear

*Corresponding Author Tel. No.: 00-962-779-416-817.

Email address: ayat.momani@yu.edu.jo (Ayat Almomani^{id})

regression model to model count data with covariates; however, it remains misspecified in the presence of structural zeros unless the zero-generating process is explicitly modeled. Furthermore, over-dispersion can make the Poisson regression model inappropriate. The negative binomial regression model is a popular solution to account for over-dispersion [7]. The negative binomial distribution is naturally derived from a mixture of the Poisson distribution and the gamma distribution, to provide a probabilistic foundation that accommodates over-dispersion in the data [1, 8]. Bayesian estimation via Markov chain Monte Carlo (MCMC) has also been applied to other regression settings involving covariates and real-world applications [9]. Several problems have arisen in the use of negative binomial regression, as has been shown by Ridout *et al.* [10].

The negative binomial (NB) distribution has only one mode; this may lead to a problem when there are data with a large inflation of zeros, which may require a distribution with more than one mode. In addition, the NB model assumes a single data-generating process. However, the problem arises when there are two distinct data-generating mechanisms: one that determines whether the event will occur at all, and the other that determines how many times it will occur [11]. Lambert [8] introduced the zero-inflated Poisson model (ZIP) as a statistical model for count data that exhibit more zeros than the Poisson regression model can accommodate. The ZIP model combines two components: the first is the Poisson distribution, which models count data with probability $(1 - \pi)$ and mean λ , and the second is the Bernoulli distribution, which models excess zeros with probability (π) .

This model is more flexible than the Poisson and negative binomial regression models for data with excess zeros. ZIP models are relatively simple to estimate and interpret when the assumptions of the Poisson distribution are satisfied; see also [12, 13]. The ZIP model, through its mixing probability, allows variance to exceed the mean. However, it may exhibit residual over-dispersion. This limitation can bias estimates and weaken the model's predictive power, particularly in small or heterogeneous samples. To handle these limitations, Hilbe [2] proposed the zero-inflated negative binomial (ZINB) model, which is an extension of the zero-inflated Poisson (ZIP) model. The ZINB model consists of two components: the first component is the negative binomial distribution, which models the count data with probability $(1 - \pi)$ and mean λ , and the second component is the Bernoulli distribution, which models the excess zeros with probability (π) .

This model accommodates the large number of zeros and the over-dispersion resulting from these zeros. Greene [14] studied how excess zeros and sample selection can be accommodated within Poisson and negative binomial regression, and set out the zero-inflated negative binomial specification in that context. Neither the standard ZIP nor the standard ZINB formulation accounts for heterogeneity between subjects or for repeated measurements on the same subject; addressing this requires a further extension. Min and Agresti [5] developed random-effects versions of zero-inflated models for repeated measures of count data, in which subject-specific effects may enter the zero component, the count component, or both. Rid-

out *et al.* [10] considered the modeling of zero-inflation data and reviewed some of the possible models.

The zero-inflation problem is also considered in the Bayesian context; Angers and Biswas [15] presented a study on the generalized zero-inflated Poisson model within the Bayesian framework. The model comparison results showed that the model provides a better fit for intense zero data than the traditional Poisson or zero-inflated Poisson models. It also has a higher ability to handle multiple peaks in the probability function and improve the prediction of future values. Xu *et al.* [16] carried out an objective Bayesian analysis of the zero-inflated Poisson distribution, deriving non-informative priors for the zero-inflation and rate parameters and examining the frequentist properties of the resulting posterior inference. Their application to healthcare data indicated that objective priors yield reliable inference for both parameters in settings where genuine prior information is unavailable.

Garay *et al.* [17] presented a comprehensive Bayesian framework for analyzing zero-inflation count data using a negative binomial regression model, with a particular focus on developing new diagnostic tools to identify influential conditions. Applied results on real-world data have shown that this approach not only provides more accurate estimates when there is over-dispersion but also enables researchers to gain a deeper understanding of data structure.

Adesina [18] fitted Bayesian multilevel models for count data using the No-U-Turn sampler, and assessed them on simulated data exhibiting both over-dispersion and under-dispersion, reporting stable estimation across the dispersion settings considered. That study concerns multilevel Poisson-type specifications rather than zero-inflated models, but it illustrates the practical feasibility of gradient-based Markov chain Monte Carlo for count regression problems.

Taken together, this literature indicates that a Bayesian treatment is workable and informative for count models with excess zeros. Our contribution is to bring the transition model of Berger and Tutz [19], which has so far been estimated by penalized maximum likelihood, into the same framework, so that parameter uncertainty is represented directly through the posterior distribution.

What distinguishes the transition model from the alternatives considered here is what it declines to assume. The ZIP and ZINB models require the counts above zero to follow a particular parametric law, and when that law is wrong the consequences are difficult to diagnose, since a poor fit may reflect the count distribution, the zero-inflation mechanism, or both. The transition model imposes no such family. Generality of this kind is normally purchased with variance, because the sequence of transition probabilities must be estimated where a parametric model would have derived it from two or three parameters, and the question is whether the price is worth paying. To answer it, the right test is not against a competitor whose assumptions are wrong, since a flexible model should do well there in any case, but against one whose assumptions are exactly right, where a flexible model should lose. Two of our three simulation scenarios are constructed so that one competitor is exactly the mechanism that generated the data, and are in that sense designed

against the model we propose.

2. Methodology

In frequentist statistics, zero-inflation models, such as ZIP and ZINB, depend on parameter-estimation procedures, for example, maximum likelihood estimation. In contrast, the Bayesian framework combines the likelihood function with the prior distribution to derive the posterior distribution, which integrates prior knowledge and information from observed data to estimate the parameters.

2.1. Bayesian zero-inflated Poisson

Let $y_1, y_2, \dots, y_n$ be a random sample of size n with excess zeros, π represents the probability of belonging to the set of structural zeros, and λ represents the rate of occurrence of events in the Poisson process. The ZIP model is given by:

$$P(Y = y_i | \mathbf{x}_i) = \begin{cases} \pi + (1 - \pi)e^{-\lambda_i} & \text{if } y_i = 0 \\ (1 - \pi) \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} & \text{if } y_i > 0 \end{cases} \quad (1)$$

Here, π is treated as a constant probability, common to all observations, rather than being modeled as a function of covariates through a logistic regression.

In the Bayesian formulation, both π and λ are treated as unknown parameters. Prior distributions are assigned to these parameters to reflect prior uncertainty. In particular, λ_i is linked to covariates through the log-linear regression model $\log(\lambda_i) = \mathbf{x}'_i \boldsymbol{\beta}$, where $\mathbf{x}_i$ are covariates, and $\boldsymbol{\beta}$ represents the associated coefficients. Appropriate prior distributions are assigned to the regression coefficients $\boldsymbol{\beta}$. The likelihood function is given by:

$$L(\pi, \boldsymbol{\beta}, \mathbf{x}, \mathbf{y}) = \prod_{i=1}^n \left[\pi + (1 - \pi)e^{-e^{\mathbf{x}'_i \boldsymbol{\beta}}} \right]^{I(y_i=0)} \times \left[(1 - \pi) \frac{e^{(y_i) \mathbf{x}'_i \boldsymbol{\beta}} e^{-e^{\mathbf{x}'_i \boldsymbol{\beta}}}}{y_i!} \right]^{I(y_i>0)}, \quad (2)$$

where $I(\cdot)$ is the indicator function, which takes the value 1 if the specified condition is satisfied and 0 otherwise. We assume the following prior distributions: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma I)$, π is reparameterized to avoid the constrained support in the interval (0, 1). Let $\omega^* = \log(\frac{\pi}{1-\pi})$, the support of ω^* now is the whole real line. Therefore, $\omega^* \sim \mathcal{N}(\mathbf{0}, \sigma)$. Then, the corresponding value of π is obtained by applying the inverse logit transformation. $\pi = \frac{e^{\omega^*}}{1+e^{\omega^*}}$, which maps the real line back to the interval (0, 1). The prior $\mathcal{N}(\mathbf{0}, \sigma)$ is placed directly on ω^* itself, rather than obtained by transforming a prior placed on π ; consequently, no Jacobian adjustment is required. Combining the likelihood function and prior distributions, the posterior is given by:

$$P(\boldsymbol{\beta}, \omega^* | \mathbf{x}, \mathbf{y}) \propto p(\boldsymbol{\beta}, \omega^*) * \prod_{i=1}^n \left[\pi + (1 - \pi)e^{-e^{\mathbf{x}'_i \boldsymbol{\beta}}} \right]^{I(y_i=0)} \times \left[(1 - \pi) \frac{e^{(y_i) \mathbf{x}'_i \boldsymbol{\beta}} e^{-e^{\mathbf{x}'_i \boldsymbol{\beta}}}}{y_i!} \right]^{I(y_i>0)}, \quad (3)$$

where $p(\boldsymbol{\beta}, \omega^*) = \prod_{j=1}^p (p(\beta_j)) \times p(\omega^*)$ is the joint prior for $j = 1, 2, \dots, p$. To apply the MCMC algorithm, the full conditional distributions (FCD) are given as follows:

$$p(\omega^* | \mathbf{y}, \mathbf{x}, \boldsymbol{\beta}) \propto e^{-\frac{(\omega^* - \mu_0)^2}{2\sigma_0^2}} \prod_{i=1}^n \left[\frac{e^{\omega^*}}{1 + e^{\omega^*}} + \frac{1}{1 + e^{\omega^*}} e^{-e^{\mathbf{x}'_i \boldsymbol{\beta}}} \right]^{I(y_i=0)} \times \frac{1}{1 + e^{\omega^*}} [e^{(y_i) \mathbf{x}'_i \boldsymbol{\beta}} e^{-e^{\mathbf{x}'_i \boldsymbol{\beta}}}]^{I(y_i>0)}, \quad (4)$$

$$p(\beta_j | \mathbf{y}, \mathbf{x}, \boldsymbol{\beta}_{-j}, \pi) \propto e^{-\frac{(\beta_j - \mu_j)^2}{2\sigma_j^2}} \prod_{i=1}^n \left[\pi + (1 - \pi)e^{-e^{\mathbf{x}'_i \boldsymbol{\beta}}} \right]^{I(y_i=0)} \times [e^{(y_i) \mathbf{x}'_i \boldsymbol{\beta}} e^{-e^{\mathbf{x}'_i \boldsymbol{\beta}}}]^{I(y_i>0)}. \quad (5)$$

2.2. Bayesian zero-inflated negative binomial

Let $y_1, y_2, \dots, y_n$ be a random sample of size n with excess zeros, π be the probability of a structural zero, θ be the dispersion parameter, and $\lambda = \frac{\theta(1-p)}{p}$ denote the mean, where p is the probability of success. The ZINB model is given by:

$$p(Y = y_i | \mathbf{x}_i) = \begin{cases} \pi + (1 - \pi) \left(\frac{\theta}{\theta + \lambda_i} \right)^\theta & \text{if } y_i = 0 \\ (1 - \pi) \binom{y_i + \theta - 1}{y_i} \left(\frac{\theta}{\theta + \lambda_i} \right)^\theta \left(\frac{\lambda_i}{\theta + \lambda_i} \right)^{y_i} & \text{if } y_i > 0 \end{cases} \quad (6)$$

Prior distributions are assigned to the model parameters (π , θ and λ). λ_i is linked to covariates through the log-linear regression model $\log(\lambda_i) = \mathbf{x}'_i \boldsymbol{\beta}$, where $\mathbf{x}_i$ are covariates, and $\boldsymbol{\beta}$ represents the associated coefficients. As in Section 2.1, π is held constant across observations and ω^* carries the prior directly, requiring no Jacobian correction. The likelihood function is given by:

$$L(\boldsymbol{\beta}, \theta, \pi; \mathbf{x}, \mathbf{y}) = \prod_{i=1}^n \left[\pi + (1 - \pi) \left(\frac{\theta}{\theta + e^{\mathbf{x}'_i \boldsymbol{\beta}}} \right)^\theta \right]^{I(y_i=0)} \times \left[(1 - \pi) \binom{y_i + \theta - 1}{y_i} \left(\frac{\theta}{\theta + e^{\mathbf{x}'_i \boldsymbol{\beta}}} \right)^\theta \left(\frac{e^{\mathbf{x}'_i \boldsymbol{\beta}}}{\theta + e^{\mathbf{x}'_i \boldsymbol{\beta}}} \right)^{y_i} \right]^{I(y_i>0)}. \quad (7)$$

The following prior distributions are assigned to the model parameters: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma I)$, $\theta \sim \text{Inv-Gamma}(\alpha, \gamma)$, π is reparameterized. Let $\omega^* = \log(\frac{\pi}{1-\pi})$, $\omega^* \sim \mathcal{N}(\mathbf{0}, \sigma)$. Combining the likelihood function and prior distributions, the posterior is given by:

$$P(\boldsymbol{\beta}, \theta, \omega^* | \mathbf{x}, \mathbf{y}) \propto p(\boldsymbol{\beta}, \theta, \omega^*) * \prod_{i=1}^n \left[\pi + (1 - \pi) \left(\frac{\theta}{\theta + e^{\mathbf{x}'_i \boldsymbol{\beta}}} \right)^\theta \right]^{I(y_i=0)} \times \left[(1 - \pi) \binom{y_i + \theta - 1}{y_i} \left(\frac{\theta}{\theta + e^{\mathbf{x}'_i \boldsymbol{\beta}}} \right)^\theta \left(\frac{e^{\mathbf{x}'_i \boldsymbol{\beta}}}{\theta + e^{\mathbf{x}'_i \boldsymbol{\beta}}} \right)^{y_i} \right]^{I(y_i>0)}, \quad (8)$$

where $p(\boldsymbol{\beta}, \theta, \omega^*) = \prod_{j=1}^p (p(\beta_j)) \times p(\theta) \times p(\omega^*)$ is the joint prior for $j = 1, 2, \dots, p$. The full conditional distributions are given as follows:

$$p(\beta_j | \mathbf{y}, \mathbf{x}, \boldsymbol{\beta}_{-j}, \theta, \pi) \propto e^{-\frac{(\beta_j - \mu_j)^2}{2\sigma_j^2}} \prod_{i=1}^n \left[\pi + (1 - \pi) \left(\frac{\theta}{\theta + e^{\mathbf{x}'\boldsymbol{\beta}}} \right)^\theta \right]^{I(y_i=0)} \times \binom{y_i + \theta - 1}{y_i} \left[\left(\frac{\theta}{\theta + e^{\mathbf{x}'\boldsymbol{\beta}}} \right)^\theta \left(\frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{\theta + e^{\mathbf{x}'\boldsymbol{\beta}}} \right)^{y_i} \right]^{I(y_i>0)}, \quad (9)$$

$$p(\theta | \mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \pi) \propto \theta^{-\alpha-1} e^{-\frac{\gamma}{\theta}} \prod_{i=1}^n \left[\pi + (1 - \pi) \left(\frac{\theta}{\theta + e^{\mathbf{x}'\boldsymbol{\beta}}} \right)^\theta \right]^{I(y_i=0)} \times \binom{y_i + \theta - 1}{y_i} \left[\left(\frac{\theta}{\theta + e^{\mathbf{x}'\boldsymbol{\beta}}} \right)^\theta \left(\frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{\theta + e^{\mathbf{x}'\boldsymbol{\beta}}} \right)^{y_i} \right]^{I(y_i>0)}, \quad (10)$$

$$p(\omega^* | \mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \theta) \propto e^{-\frac{(\omega^* - \mu_0)^2}{2\sigma_0^2}} \prod_{i=1}^n \left[\frac{e^{\omega^*}}{1 + e^{\omega^*}} + \frac{1}{1 + e^{\omega^*}} \left(\frac{\theta}{\theta + e^{\mathbf{x}'\boldsymbol{\beta}}} \right)^\theta \right]^{I(y_i=0)} \times \frac{1}{1 + e^{\omega^*}} \binom{y_i + \theta - 1}{y_i} \left[\left(\frac{\theta}{\theta + e^{\mathbf{x}'\boldsymbol{\beta}}} \right)^\theta \left(\frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{\theta + e^{\mathbf{x}'\boldsymbol{\beta}}} \right)^{y_i} \right]^{I(y_i>0)}. \quad (11)$$

2.3. Transition model

Berger and Tutz [19] proposed a new semi-parametric zero-inflated model for count data, known as the transition model. This model was proposed as an alternative and flexible model to traditional models such as the Poisson, negative binomial, ZIP, and ZINB models, which depend on fixed distributional assumptions. The basic idea of the transition model is to avoid identifying a distribution for the count response. One advantage of this model is its capacity to accommodate both zero inflation and over-dispersion within a unified framework, without the need to separate the zero-generating and count-generating processes as in zero-inflated models. Let the random variable y represent the count of the event with possible values $\{y = 0, 1, 2, 3, \dots\}$, and let $\pi_r = P(Y = r)$ be the marginal probability of y . The model is then characterized by the conditional probability as follows

$$\delta_r = p(Y > r | Y \geq r) = \frac{1 - \pi_0 - \pi_1 - \pi_2 - \dots - \pi_r}{1 - \pi_0 - \pi_1 - \pi_2 - \dots - \pi_{r-1}}. \quad (12)$$

Berger and Tutz [19] assumed that the conditional probability follows a cumulative distribution function to increase flex-

ibility in modeling the data distribution, without considering a fixed distribution for the data, and is given as follows

$$p(Y > r | Y \geq r, \mathbf{x}_i) = F(\theta_r + \mathbf{x}'_i \boldsymbol{\beta}). \quad (13)$$

Here, $\mathbf{x}_i$ represents the covariate variables for the i -th observation, $\boldsymbol{\beta}$ represents the associated regression coefficients, and θ_r represents the parameter related to the distribution of the count data. $F(\cdot)$ is a known cumulative distribution function.

Let $f(r) = P(Y = r | \mathbf{x}_i)$ denote the marginal probability of observing the count r . This probability can be obtained from the transition probabilities. To observe exactly r , the process must first reach r , and then stop at r . The probability of reaching r is

$$P(Y \geq r | \mathbf{x}_i) = \prod_{s=0}^{r-1} P(Y > s | Y \geq s, \mathbf{x}_i). \quad (14)$$

After reaching r , the conditional probability of stopping at r is

$$P(Y = r | Y \geq r, \mathbf{x}_i) = 1 - P(Y > r | Y \geq r, \mathbf{x}_i). \quad (15)$$

Therefore, the conditional probability mass function is

$$f(r) = [1 - P(Y > r | Y \geq r, \mathbf{x}_i)] \prod_{s=0}^{r-1} P(Y > s | Y \geq s, \mathbf{x}_i). \quad (16)$$

By substituting Equation 13 into Equation 16, we obtain

$$f(r) = [1 - F(\theta_r + \mathbf{x}'_i \boldsymbol{\beta})] \prod_{s=0}^{r-1} F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta}). \quad (17)$$

Berger and Tutz [19] discussed maximum likelihood and penalized maximum likelihood to analyze the transition model and found that the transition model performed better than well-known count models such as Poisson, Negative Binomial, ZIP, and ZINB. Unlike those traditional models, which assume a fixed distribution, the transition model adapts to the actual data structure. The estimation of the transition model is iterative, and it requires certain conditions to have effective inference. The most important conditions are determining an appropriate maximum likelihood estimator and selecting good initial values to achieve faster convergence. Also, the results depend on the convergence of the estimation procedure and the sample size. The size of the sample significantly affects the inference; larger samples generally yield better estimates. Furthermore, maximum likelihood estimation provides point estimates, which summarize the parameter as a single value; however, uncertainty can be assessed using standard errors or confidence intervals.

To address the limitations of point estimation, a Bayesian approach can be adopted. The Bayesian approach does not treat the parameter as a fixed value but rather as a random variable with a prior distribution, thereby allowing direct representation of uncertainty through the posterior distribution. Additionally, the Bayesian approach can handle small sample sizes, providing greater flexibility. In this work, we apply the Bayesian approach to transition models to enhance their flexibility and performance.

2.4. Proposed model

In this paper, we propose a Bayesian approach for the transition model. Let $(\mathbf{x}, \mathbf{y})$ be a random sample of size n . The transition model is given by:

$$f(y_i) = \prod_{s=0}^{y_i} (1 - F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta}))^{I(s=y_i)} F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta})^{1-I(s=y_i)}, \quad (18)$$

where $F(\cdot)$ is a known cumulative distribution function, which can be the logistic distribution or the probit distribution, $\mathbf{x}'_i \boldsymbol{\beta}$ is the linear predictor in which $\mathbf{X} = (\mathbf{x}_{1i}, \mathbf{x}_{2i}, \dots, \mathbf{x}_{pi})$ is the $(n \times p)$ covariate matrix, where n is the number of observations, p the number of predictors, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$ are the associated coefficients, and the parameter θ_s represents the intercept parameter associated with the s -th transition. Let $M_{\max} = \max_i(y_i)$ denote the maximum observed count. To allow the model to account for counts larger than the maximum observed value, the transition parameters are defined up to a value M such that $M > M_{\max}$. Thus, $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \dots, \theta_M)$. Berger and Tutz [19] suggest choosing M somewhat larger than $M_{\max}$, so that the model assigns probability to counts beyond those observed. Under the terminal-category convention introduced below, however, the likelihood indexes θ only as far as $M_{\max}$, so any category above that is a parameter the data cannot inform. We therefore take $M = M_{\max} + 1$, the smallest choice that leaves the model unchanged while avoiding such parameters; the terminal category still carries $P(Y > M_{\max})$, so no probability is lost.

Because $\boldsymbol{\theta}$ is finite dimensional, the support of the model has to be closed at M . If every θ_s , $s = 0, \dots, M$, were left as a free transition parameter, the residual probability $P(Y > M | \mathbf{x}_i) = \prod_{s=0}^M F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta})$ would remain unallocated, the implied mass function would not sum to one, and the finite sum used later to compute the fitted mean would not be the expected value of a proper distribution. We therefore treat M as a *terminal* (absorbing) category and adopt the convention

$$\begin{aligned} \delta_M = P(Y > M | Y \geq M, \mathbf{x}_i) &= 0, \\ \text{equivalently } F(\theta_M + \mathbf{x}'_i \boldsymbol{\beta}) &:= 0, \end{aligned} \quad (19)$$

so that the free transition parameters are $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_{M-1})$ and θ_M is not estimated. Under (19) the mass function of the proposed model is

$$f(r) = \begin{cases} [1 - F(\theta_r + \mathbf{x}'_i \boldsymbol{\beta})] \prod_{s=0}^{r-1} F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta}) & r = 0, 1, \dots, M-1, \\ \prod_{s=0}^{M-1} F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta}), & r = M, \end{cases} \quad (20)$$

which satisfies $\sum_{r=0}^M f(r) = 1$ for every $\mathbf{x}_i$, since the terminal category absorbs all mass not allocated to $r < M$. The choice $M > M_{\max}$ guarantees that the terminal category is never observed in the sample, so (19) places no constraint on the observed data and only fixes how the unobserved upper tail is handled. All likelihood, posterior and prediction expressions below are evaluated under this convention; in particular, the likelihood

in (22) remains valid as written, because for an observation with $y_i = M$ the factor $(1 - F(\theta_M + \mathbf{x}'_i \boldsymbol{\beta}))^{I(s=y_i)}$ equals one and the remaining factors reproduce the second line of (20).

Convention (19) is not the only way to close the support, and the alternative is worth stating because the two are not equivalent in practice. One may instead retain $f(0), \dots, f(M-1)$ as in the first line of (20) and rescale them by their sum, which also yields a proper distribution but assigns no mass to the terminal category at all. The two agree whenever the residual probability

$$P(Y > M-1 | \mathbf{x}_i) = \prod_{s=0}^{M-1} F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta}), \quad (21)$$

is negligible, which is the usual case: for an observation with a moderate linear predictor the product in (21) is many orders of magnitude below one. They diverge sharply for observations whose linear predictor is large, since (21) then approaches one and the absorbing convention contributes a term $M \cdot P(Y > M-1 | \mathbf{x}_i)$ to the fitted mean. This is a real modelling decision rather than a numerical detail, and it also qualifies the sense in which the predictive mean of the transition model is bounded: it is bounded by M , which is not itself a small number. We therefore report the largest value of (21) across observations as a diagnostic. It reaches 0.004 in the zero-inflated Poisson scenario, 0.170 in the zero-inflated negative binomial scenario, 0.273 in the Poisson parameter-driven scenario and 0.039 in the real-data application.

Since these are not uniformly negligible, we do not assume the two closures agree: the absorbing convention (19) is used throughout and the rescaling alternative is reported alongside it in Table 7. As that table shows, the fitted values differ by at most 0.014 in the simulations and 0.19 on the real data, far below the between-replication variability, so the choice is immaterial for these data even where the residual mass is appreciable. It must nonetheless be specified, since without it the predictive distribution is not normalized.

The likelihood function is given by

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{x}, \mathbf{y}) = \prod_{i=1}^n \prod_{s=0}^{y_i} (1 - F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta}))^{I(s=y_i)} F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta})^{1-I(s=y_i)}. \quad (22)$$

For Bayesian inference, a prior distribution is assigned to each model parameter; the location-scale t distribution is used in this work for $(\boldsymbol{\theta}$ and $\boldsymbol{\beta})$. For $j = 1, 2, \dots, p$, $\beta_j \sim t_\nu(\mu_{\beta_j}, \sigma_{\beta_j}^2)$ and for $k = 0, 1, 2, \dots, M-1$, $\theta_k \sim t_\kappa(\mu_{\theta_k}, \sigma_{\theta_k}^2)$ (recall that θ_M is fixed by the terminal-category convention (19) and therefore carries no prior) where ν and κ are the degrees of freedom of β_j and θ_k , respectively. Combining the likelihood function and prior distributions, the posterior is given by:

$$P(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{x}, \mathbf{y}) \propto p(\boldsymbol{\beta}, \boldsymbol{\theta}) \prod_{i=1}^n \prod_{s=0}^{y_i} (1 - F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta}))^{I(s=y_i)} F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta})^{1-I(s=y_i)}, \quad (23)$$

where $p(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{j=1}^p (p(\beta_j)) \times \prod_{k=0}^{M-1} (p(\theta_k))$. The full conditional distributions are given as follows.

$$p(\beta_j | \mathbf{y}, \mathbf{x}, \boldsymbol{\beta}_{(j)}, \boldsymbol{\theta}) \propto \left[1 + \frac{1}{\nu} \left(\frac{\beta_j - \mu_{\beta_j}}{\sigma_{\beta_j}} \right)^2 \right]^{-\frac{\nu+1}{2}} \prod_{i=1}^n \prod_{s=0}^{y_i} (1 - F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta}))^{I(s=y_i)} F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta})^{1-I(s=y_i)}, \quad (24)$$

$$p(\boldsymbol{\theta} | \mathbf{y}, \mathbf{x}, \boldsymbol{\beta}) \propto \prod_{k=0}^{M-1} \left[1 + \frac{1}{\kappa} \left(\frac{\theta_k - \mu_{\theta_k}}{\sigma_{\theta_k}} \right)^2 \right]^{-\frac{\kappa+1}{2}} \prod_{i=1}^n \prod_{s=0}^{y_i} (1 - F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta}))^{I(s=y_i)} F(\theta_s + \mathbf{x}'_i \boldsymbol{\beta})^{1-I(s=y_i)}. \quad (25)$$

It is worth noting that the intercept parameters $\theta_0, \theta_1, \dots, \theta_{M-1}$ are assigned independent priors, without imposing any structural relationship between parameters associated with adjacent count categories. As a result, categories with few observations may be weakly identified by the likelihood, leaving their posterior largely governed by the prior.

This point can be made precise, and in one region of the parameter space it is not a matter of degree. Inspection of (22) shows that θ_k enters the likelihood only through observations with $y_i \geq k$: such an observation contributes $F(\theta_k + \mathbf{x}'_i \boldsymbol{\beta})$ if $y_i > k$ and $1 - F(\theta_k + \mathbf{x}'_i \boldsymbol{\beta})$ if $y_i = k$, while observations with $y_i < k$ contribute no factor involving θ_k at all. Consequently, for any $k > \max_i y_i$ the likelihood is constant in θ_k and the posterior for that parameter is exactly its prior. Since M is chosen larger than $M_{\max}$, the transition parameters indexed between $\max_i y_i$ and M are of this kind: they are not informed by the data in any degree, yet under a prediction rule that does not close the support at M they still influence the fitted mean through the factors $F(\theta_k + \mathbf{x}'_i \boldsymbol{\beta})$. The terminal-category convention (19) limits the consequences of this, since the unallocated mass is then assigned to a single known category rather than distributed according to prior draws, but it does not remove the underlying weak identification.

A common remedy is to impose a smoothness assumption across adjacent θ_k 's (e.g., a random-walk prior), as considered in the penalized likelihood framework of Berger and Tutz [19]. Such a prior would allow the sparsely observed categories to borrow strength from their neighbours and would remove the dependence of the fitted mean on parameters the data cannot inform; incorporating this structure within the Bayesian framework is left for future work.

3. Simulation study

We compare the three models across three data-generating mechanisms, chosen so that each of the two zero-inflated models is correctly specified in one scenario and none of the three is correctly specified in the last. Each scenario is replicated $R = 100$ times at sample sizes $n = 100$ and $n = 200$.

For every replication we record the bias of the regression coefficients and the root mean squared error of the fitted counts, and we summarise the RMSE across replications by its mean and variance, the second being as informative as the first: a model whose predictions vary widely from one dataset to the next is of limited use whatever its average performance. Here $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$ and $\hat{y}_i$ is the fitted value on the expected-count scale. The mean squared error (MSE) of a regression coefficient $\hat{\beta}_j$ across the $R = 100$ replications is defined as $MSE(\hat{\beta}_j) = \frac{1}{R} \sum_{r=1}^R (\hat{\beta}_j^{(r)} - \beta_j)^2$, where β_j is the true value used to generate the data and $\hat{\beta}_j^{(r)}$ is the posterior mean estimate from replication r ; as explained in Section 3.1, this quantity is reported only for the Bayesian ZIP and Bayesian ZINB models.

Because RMSE is used as the single measure that is comparable across the three models, the fitted value must be the expected count implied by each fitted model, and not a component of it. Under a zero-inflated specification a proportion π of the observations are structural zeros, so the unconditional expected count is

$$E(Y_i | \mathbf{x}_i) = (1 - \pi)\lambda_i, \quad \lambda_i = \exp(\mathbf{x}'_i \boldsymbol{\beta}), \quad (26)$$

for the Bayesian ZIP model, and likewise $(1 - \pi)\lambda_i$ for the Bayesian ZINB model, since the negative binomial count component also has mean λ_i . The quantity λ_i on its own is the mean of the count component conditional on *not* being a structural zero, and using it as the fitted value would systematically overstate the predicted counts by a factor $1/(1 - \pi)$. Accordingly, for both zero-inflated models the fitted value is taken as the posterior mean of the expected count,

$$\hat{y}_i = \frac{1}{S} \sum_{t=1}^S (1 - \pi^{(t)}) \exp(\mathbf{x}'_i \boldsymbol{\beta}^{(t)}), \quad \pi^{(t)} = \frac{e^{\omega^{*(t)}}}{1 + e^{\omega^{*(t)}}}, \quad (27)$$

where $t = 1, \dots, S$ indexes the S retained post-burn-in draws. For the Bayesian Transition model, the fitted value is the corresponding posterior expected count under the mass function (20),

$$\hat{y}_i = \frac{1}{S} \sum_{t=1}^S \left\{ \sum_{r=0}^{M-1} r \left[(1 - F(\theta_r^{(t)} + \mathbf{x}'_i \boldsymbol{\beta}^{(t)})) \prod_{s=0}^{r-1} F(\theta_s^{(t)} + \mathbf{x}'_i \boldsymbol{\beta}^{(t)}) \right] + M \prod_{s=0}^{M-1} F(\theta_s^{(t)} + \mathbf{x}'_i \boldsymbol{\beta}^{(t)}) \right\}, \quad (28)$$

in which the final term is the contribution of the terminal category and its inclusion is what makes (28) the mean of a normalized distribution. All three models are therefore evaluated on the same expected-count scale. The first dataset is from the zero-inflated Poisson model. Assume λ such that $\log(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where the covariate variables, x_1 and x_2 , are generated using $x_1 \sim N(2, 0.2)$, $x_2 \sim N(2, 0.25)$, and the parameter values are $\beta_0 = 0, \beta_1 = 1.5, \beta_2 = -2$. The percentage of zero counts is set to $\pi = 0.2$. To generate a data set from a ZIP model, for $i = 1, \dots, n$, generate a binary indicator

$I_i \sim \text{Bernoulli}(\pi)$ and

$$y_i = \begin{cases} \text{Poisson}(\lambda_i), & I_i = 0 \\ 0, & I_i = 1. \end{cases} \quad (29)$$

For the second data set from the zero-inflated Negative Binomial model, the covariates are generated using $x_1 \sim N(0, 1)$, and $x_2 \sim N(0, 1)$, where parameter values are $\beta_0 = 0, \beta_1 = 1.2, \beta_2 = -1.2, \theta = 1.5$, and the percentage of zero count is $\pi = 0.2$. $\log(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. For $i = 1, \dots, n$, a binary indicator $I_i \sim \text{Bernoulli}(\pi)$ is generated, then

$$y_i = \begin{cases} \text{NB}(\theta, \lambda_i), & I_i = 0 \\ 0, & I_i = 1. \end{cases} \quad (30)$$

The third dataset is generated by the Poisson parameter-driven model. Let $y_1, \dots, y_n$ be a sequence of counts with $y_i | \lambda_i \sim \text{Poisson}(\lambda_i)$, let the link function be $\log(\lambda_i) = \mathbf{x}_i' \boldsymbol{\beta} + \boldsymbol{\eta}_i$ where $\boldsymbol{\beta}$ are the regression coefficients and $\boldsymbol{\eta}_i$ are unobserved latent processes, which can be modeled with an auto-regression process of order $p = 1$, denoted by $AR(1)$, that is $\eta_i = \phi \eta_{i-1} + \epsilon_i$, $i > 1$ where $\eta_0 \sim N(0, \frac{\sigma_\epsilon^2}{1-\phi^2})$ and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. To generate from this model, fix $\beta_0 = 0.1, \beta_1 = 1.5, \beta_2 = -2, \phi = 0.8, \sigma_\epsilon^2 = 0.4$. Generate x_1 from $N(-2, 0.2)$, x_2 from $N(-2, 0.25)$, ϵ_i from $N(0, \sigma_\epsilon^2)$, and η_0 from $N(0, \frac{\sigma_\epsilon^2}{1-\phi^2})$. For $i = 1, \dots, n$, let $\eta_i = \phi \eta_{i-1} + \epsilon_i$, and $\lambda_i = \exp(\mathbf{x}_i' \boldsymbol{\beta} + \boldsymbol{\eta}_i)$, then $y_i \sim \text{Poisson}(\lambda_i)$.

It should be noted that the three fitted models (Bayesian ZIP, Bayesian ZINB, and Bayesian Transition) assume conditional independence between observations and do not explicitly include the $AR(1)$ structure used in data generation. Therefore, this scenario aims to evaluate the robustness of the models under model misspecification, where the models do not account for the temporal dependencies present in the data generation process.

The three models are compared under the three scenarios described above. None of the full conditional distributions derived in Section 2 is available in closed form, so each is sampled by Metropolis–Hastings within a Gibbs scheme. Every model is run for $N_{\text{iter}} = 100,000$ iterations with the first 20,000 discarded as burn-in and the remainder thinned by ten. We use the same settings for all three models so that the comparison is not confounded by differences in the sampling effort each receives; the value was chosen from preliminary runs, in which shorter chains gave effective sample sizes below the conventional threshold for some replications of the Transition model.

The prior specification is the same in every scenario and in the real-data application. For the two zero-inflated models we take $\beta_j \sim N(0, 10)$ for each coefficient and $\omega^* \sim N(0, 10)$ on the logit of the zero-inflation probability, and for the negative binomial dispersion $\theta \sim \text{Inverse-Gamma}(4, 13)$. For the Transition model F is the standard normal distribution function and the coefficients and transition parameters are given independent $t_3(0, 3.5)$ priors. These are weakly informative: they restrain the sampler in regions the data do not reach without materially influencing the posterior where the data are informative.

Two features of the implementation are worth setting out, because both affect how well the samplers mix and neither is incidental to the results.

The covariates are centred and scaled before sampling and the draws are transformed back to the original scale afterwards. This is a reparameterisation and leaves the posterior unchanged, but it matters a good deal in practice. In the first scenario the covariates are drawn as $x_{1i} \sim N(2, 0.2)$ and $x_{2i} \sim N(2, 0.25)$, which are nearly constant, so the design matrix $[\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2]$ has a condition number close to 1700. The posterior for the regression coefficients is then a long, narrow ridge along which a componentwise random walk moves very slowly. Centring and scaling reduces the condition number to about 1.1.

The regression coefficients are also updated as a block, using a multivariate proposal whose covariance is learned during burn-in and scaled towards a target acceptance rate, rather than one coefficient at a time. A componentwise random walk cannot traverse a correlated posterior efficiently however carefully each individual scale is tuned. The transition parameters θ_k are updated componentwise, since each enters the likelihood through a distinct subset of the observations. Taken together these two choices raised the minimum effective sample size in preliminary runs by roughly two orders of magnitude.

Convergence is assessed by the split Gelman–Rubin statistic ($\hat{R}$), obtained by halving each chain, and by the bulk effective sample size, reported as the minimum across the parameters of each fitted model. For the Bayesian Transition model both are computed over the transition parameters the likelihood identifies, that is those with $k \leq \max_i y_i$; as shown in Section 2, components with a higher index have posterior equal to prior, so a diagnostic computed on them would describe the prior rather than the sampler. Table 1 reports these together with the acceptance rates; the minimum effective sample size there is taken over all parameters and all replications, and so describes the worst case rather than the typical one. The regression coefficients of the Transition model are the slowest-mixing quantities in the study, which is why its minimum is the smallest of the three, but all values exceed the conventional threshold of 400. All computations were carried out in Python using numpy and scipy.

3.1. Which quantities are comparable across the three models

Before presenting the results we make explicit which of the reported measures admit a cross-model comparison, because the three models are not parameterised on a common scale.

In the Bayesian ZIP and Bayesian ZINB models the coefficient vector enters the log mean of the count component, $\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$, and the values β_1 and β_2 fixed in each data-generating mechanism are defined on precisely that scale. In the Transition model the same symbol $\boldsymbol{\beta}$ plays a different role: it enters the conditional transition probabilities $P(Y > r \mid Y \geq r, \mathbf{x}_i) = F(\theta_r + \mathbf{x}_i' \boldsymbol{\beta})$ through a probit link, so β_j measures the change in a sequence of conditional continuation probabilities rather than a multiplicative change in a mean. The two coefficient vectors are not two estimators of one estimand; they are parameters of two different functionals of the data-generating distribution, and no exact transformation maps one onto the other.

Consequently, the difference between a Transition-model posterior mean and a generating ZIP or ZINB coefficient is

Table 1: MCMC convergence diagnostics across the simulated datasets and the real-data application: number of chains, Gelman–Rubin statistic $\hat{R}$, minimum bulk effective sample size (ESS) over the parameters of each fitted model, and Metropolis–Hastings acceptance rates. Values are summarised over replications.

Model	Chains	Max $\hat{R}$	Min ESS	Median ESS	Median acceptance
Bayesian ZIP	1	1.000	1511	8000	0.338
Bayesian ZINB	1	1.000	1041	8000	0.435
Bayesian Transition	1	1.002	427	1014	0.440

The minimum ESS is taken over all parameters and all replications, and is therefore a worst case. Sampling settings and the treatment of the unidentified transition parameters are given in Section 3.

not a bias, and its square is not a mean squared error, because the quantity being estimated is undefined. In Tables 2–4 the coefficient-level bias and MSE are therefore reported only for the Bayesian ZIP and Bayesian ZINB models, whose coefficients share the estimand used to generate the data; the corresponding Transition-model cells are marked “n.a.” and are not to be read as poor or missing estimates.

A related but distinct point concerns the fixed term. The Transition model has no single intercept: the parameters $\theta_0, \theta_1, \dots, \theta_{M-1}$ jointly describe the baseline of the count distribution and collectively play the role that β_0 plays in the two zero-inflated models, but they differ from it in both dimension and interpretation. The dash (–) reported for β_0 under the Transition model in Tables 2–4 records this structural absence.

What can be compared across all three models is any functional of the fitted predictive distribution, since such a functional is invariant to how the model is parameterised. We use the RMSE computed on the expected-count scale, with $\hat{y}_i$ defined by (27) for the zero-inflated models and by (28) for the Transition model. Within each model, the coefficient bias and MSE that are reported still convey how stably that model recovers its own parameters as the sample size grows; they simply do not license a ranking of the three models against one another.

3.2. Results under zero-inflated Poisson data

Table 2 summarises the first scenario. At $n = 100$ the correctly specified Bayesian ZIP model recovers β_1 and β_2 with small bias and small MSE, and the Bayesian ZINB model, which contains the ZIP as a limiting case, behaves comparably though with slightly larger values throughout. At $n = 200$ both models improve, with bias and MSE close to zero for both coefficients, which is the expected pattern under correct specification.

On the expected-count scale the three models are indistinguishable. At $n = 100$ the mean RMSE is 0.6006 for the Bayesian ZIP model, 0.6009 for the Bayesian ZINB model and 0.5978 for the Bayesian Transition model, a spread of 0.003 against a between-replication variance of roughly 0.008; at $n = 200$ the three values are 0.6408, 0.6415 and 0.6419, with identical variances of 0.0025. No ordering among them is supported at either sample size.

This deserves more emphasis than a statement of equivalence usually would, because the scenario is not symmetric between the models. The data were generated from a zero-inflated Poisson mechanism, so the Bayesian ZIP model is cor-

rectly specified and serves as a benchmark: no estimator can do systematically better than the model that generated the data. The Transition model assumes no distributional family for the counts at all, and performs at least as well as that benchmark at both sample sizes, with the lower mean RMSE of the two at $n = 100$. Making no assumption about the distribution of the counts therefore costs nothing here, and it costs nothing in the hardest case: the competing model’s assumption is exactly right, so if anything were going to expose the price of the weaker assumption, this scenario would.

3.3. Results under zero-inflated negative binomial data

Table 3 reports the second scenario, in which the data are generated from a zero-inflated negative binomial mechanism and the Bayesian ZINB model is correctly specified. At $n = 100$ both the Bayesian ZIP and the Bayesian ZINB models estimate β_1 and β_2 with small bias and small MSE, with the correctly specified ZINB model showing the expected advantage in the intercept once the over-dispersion is accounted for. On the expected-count scale the mean RMSE at $n = 100$ is 5.3718 for the Bayesian ZIP model, 5.9015 for the Bayesian ZINB model and 5.5154 for the Bayesian Transition model. The differences are small relative to the between-replication variance, which lies between 4.7 and 5.8 and implies a standard error of roughly 0.23, so they do not support an ordering. Two things are nonetheless worth recording.

The first is that the correctly specified Bayesian ZINB model does not attain the lowest RMSE in its own scenario: the additional dispersion parameter is estimated at a cost in variance that, at these sample sizes, exceeds the reduction in bias it delivers. The second is that the Transition model attains a lower mean RMSE than that correctly specified model at both sample sizes, 5.5154 against 5.9015 at $n = 100$ and 7.6604 against 7.7748 at $n = 200$. We do not claim an advantage from these gaps, since both lie inside the replication variability. What they establish is that a model making no assumption whatever about the distribution of the counts performs at least as well as the model that generated them, even under substantial over-dispersion.

As set out in Section 3.1, no coefficient-level comparison against the Transition model is made, since the generating values are defined on the log-mean scale alone.

At the larger sample size $n = 200$ the ZIP and ZINB coefficient estimates become nearly unbiased, with bias values

Table 2: Simulation results for the three models under zero-inflated Poisson data, $R = 100$ replications.

	Metric		ZIP	ZINB	Transition
$n = 100$	β_0	Bias	-0.5165	-0.5532	–
		MSE	9.0321	9.1065	–
	β_1	Bias	0.1775	0.1952	n.a.
		MSE	1.2836	1.2847	n.a.
	β_2	Bias	-0.0082	-0.0320	n.a.
		MSE	0.5071	0.5066	n.a.
	$RMS E$	Mean	0.6006	0.6009	0.5978
		Var	0.0092	0.0082	0.0067
	β_0	Bias	-0.1150	-0.1683	–
		MSE	2.5253	2.6128	–
$n = 200$	β_1	Bias	0.0172	0.0305	n.a.
		MSE	0.3792	0.3894	n.a.
	β_2	Bias	-0.0383	-0.0560	n.a.
		MSE	0.2577	0.2542	n.a.
	$RMS E$	Mean	0.6408	0.6415	0.6419
		Var	0.0025	0.0025	0.0025

Key: “–” marks the absence of a single intercept in the Transition model; “n.a.” marks a quantity that is not defined for it. See Section 3.1.

Table 3: Simulation results for the three models under zero-inflated negative binomial data, $R = 100$ replications.

	Metric		ZIP	ZINB	Transition
$n = 100$	β_0	Bias	0.1931	-0.0074	–
		MSE	0.1224	0.0475	–
	β_1	Bias	-0.1046	-0.0035	n.a.
		MSE	0.0924	0.0415	n.a.
	β_2	Bias	0.1285	0.0237	n.a.
		MSE	0.0736	0.0348	n.a.
	$RMS E$	Mean	5.3718	5.9015	5.5154
		Var	4.6930	5.7819	5.1594
	β_0	Bias	0.1555	-0.0153	–
		MSE	0.0732	0.0254	–
$n = 200$	β_1	Bias	-0.0454	0.0288	n.a.
		MSE	0.0327	0.0142	n.a.
	β_2	Bias	0.0518	-0.0139	n.a.
		MSE	0.0264	0.0122	n.a.
	$RMS E$	Mean	7.5534	7.7748	7.6604
		Var	3.1934	3.4183	3.5192

Key: “–” marks the absence of a single intercept in the Transition model; “n.a.” marks a quantity that is not defined for it. See Section 3.1.

close to zero for both models, so accuracy and stability improve with sample size in the way one would expect for models that share the generating parameterisation. The same pattern holds at $n = 200$, where the mean RMSE is 7.5534, 7.7748 and

7.6604 respectively, again well within the replication variability. Any difference between the three models in this scenario is therefore below the resolution of the study, and we make no claim about how well the Transition model recovers individual

coefficients, a question that is not well posed for it given the generating mechanism.

3.4. Results under the Poisson parameter-driven data

The third scenario differs from the other two in that no fitted model is correctly specified: the data carry an AR(1) latent process that none of the three represents. It is therefore the setting in which the transition model's freedom from a distributional assumption about the count component might be expected to show to advantage, and the one that bears most directly on whether that freedom is worth its cost.

Table 4 reports the third scenario, in which the data are generated by a Poisson process with an AR(1) latent component that none of the three fitted models represents. At $n = 100$ all three models are misspecified, and this shows in the coefficient results for the two models where such results are defined: both ZIP and ZINB exhibit substantially larger bias for β_0 and β_1 than in the earlier scenarios. On the expected-count scale the mean RMSE at $n = 100$ is 8.6473, 8.7976 and 8.7589 for the ZIP, ZINB and Transition models, a spread of 0.15 against a between-replication variance near 27, corresponding to a standard error of roughly 0.52.

At $n = 200$ the bias of the ZIP and ZINB coefficients moves closer to zero, but does not disappear, which is consistent with the omitted temporal dependence acting as a persistent source of misspecification rather than a small-sample effect. At $n = 200$ the values are 9.4739, 9.5091 and 9.5043, again indistinguishable. Overall, in Tables 2–4 the RMSE of every model decreases as the sample size increases. This scenario is the only one in which no fitted model is correctly specified, so it was included to test whether the transition mechanism's freedom from a distributional assumption confers an advantage when the parametric alternatives are wrong. It does not: the three models again predict the expected count equally well, the transition model neither gaining nor losing when all three specifications are wrong in the same way. The mechanism's value, on this evidence, does not lie in tolerating misspecification of a kind that all three models share.

Taken across the three scenarios, however, the pattern of equivalence is itself informative, and it is worth being explicit about why. In the first two scenarios one of the competing models is the mechanism that generated the data, and is therefore the best predictor available to any method; in the third, none is correct but all three are wrong in the same way. The Bayesian Transition model performs at least as well as the competing models throughout, including the correctly specified one in each of the first two scenarios, while assuming nothing about the distribution of the counts beyond the sequence of transition probabilities it estimates. The usual price of such generality is variance: the model estimates M transition parameters where its competitors estimate three or four, many of them informed by only a handful of observations. That price is evidently small enough here not to show against a correctly specified parametric alternative, which is the most demanding comparison available and the one a flexible model would be expected to lose.

None of this is evidence about parameter estimation, and the design of the study does not permit such a comparison.

4. Real data application

Fish data are used in this section to verify the performance of the proposed model in realistic problems. These data were obtained from the UCLA Office of Academic Resources and Consulting (OARC) [20], and the dataset represents the counts of visitor groups in a state park. For each group, the number of fish caught was recorded, along with the number of children in the group, the total number of persons, and whether the group brought a camper. The data set includes approximately 250 observations and is characterized by many zero counts. The reason for the emergence of zero inflation is that some visitors do not fish, while others who do catch no fish. Consequently, the dataset contains excess zeros, and there is no information on whether a person fished. This feature makes it particularly suitable for applying count models for zero-inflation. The focus in this work is to use the count data with covariate variables xb and zg to apply our model and compare its performance with other competition models: Bayesian ZIP and Bayesian ZINB.

Priors, proposal scheme and sampling settings are exactly as in the simulation study of Section 3, so that the two analyses are directly comparable.

The observations are divided into a training set and a test set before any model is fitted. The test set consists of 20% of the observations, drawn by simple random sampling without replacement using the `sample()` function in R with the random seed 20260101 + r for replication r , giving 200 training and 50 test observations. The same split is used for all three models, so the comparison is not confounded by differences in the data each model sees.

Predictive accuracy is evaluated on the expected-count scale defined in Section 3, using exactly the same fitted values as in the simulation study: (27) for the Bayesian ZIP and Bayesian ZINB models, so that $\hat{y}_i$ estimates $(1 - \pi)\lambda_i$ rather than λ_i , and (28) for the Bayesian Transition model, evaluated under the terminal-category convention (19) with $M = M_{\max} + 1$ computed from the training data. The largest residual probability (21) across the observations is 0.039, which is not negligible, so we report both closures of Section 2 rather than assuming they agree. Table 7 collects the comparison for the simulations and the real data together: although the residual mass reaches 0.27 in one scenario, the two closures differ by at most 0.014 in the simulations and 0.19 here, in both cases far below the variability between replications.

The choice is therefore immaterial for these data, though it must still be made for the predictive distribution to be normalized at all. The results of the analysis of the real data using the three models are presented in Table 5, Table 6 and Figure 1.

Predictive performance is assessed over repeated splits rather than a single one. On a test set of this size a single badly predicted observation can dominate a squared-error criterion, so a comparison based on one partition measures which model that partition happened to favour as much as it measures predictive ability. Two pieces of evidence are therefore reported.

First, we repeat the train–test procedure over $R = 100$ independent random splits of the same 80/20 proportion and summarise the distribution of the resulting test errors. Sec-

Table 4: Simulation results for the three models under Poisson parameter-driven data with an AR(1) latent process, $R = 100$ replications.

Metric			ZIP	ZINB	Transition
$n = 100$	β_0	Bias	0.5922	0.3595	—
		MSE	4.9410	3.5783	—
	β_1	Bias	-0.1757	0.0443	n.a.
		MSE	0.6314	0.3988	n.a.
	β_2	Bias	0.1644	-0.1082	n.a.
		MSE	0.5127	0.3315	n.a.
	$RMSE$	Mean	8.6473	8.7976	8.7589
		Var	26.7295	28.0576	28.0256
	β_0	Bias	0.8062	0.5786	—
		MSE	2.6011	2.3328	—
$n = 200$	β_1	Bias	-0.1425	0.0558	n.a.
		MSE	0.3540	0.2682	n.a.
	β_2	Bias	0.2246	-0.0195	n.a.
		MSE	0.2746	0.1489	n.a.
	$RMSE$	Mean	9.4739	9.5091	9.5043
		Var	22.1850	22.6520	22.7129

Key: “—” marks the absence of a single intercept in the Transition model; “n.a.” marks a quantity that is not defined for it. See Section 3.1.

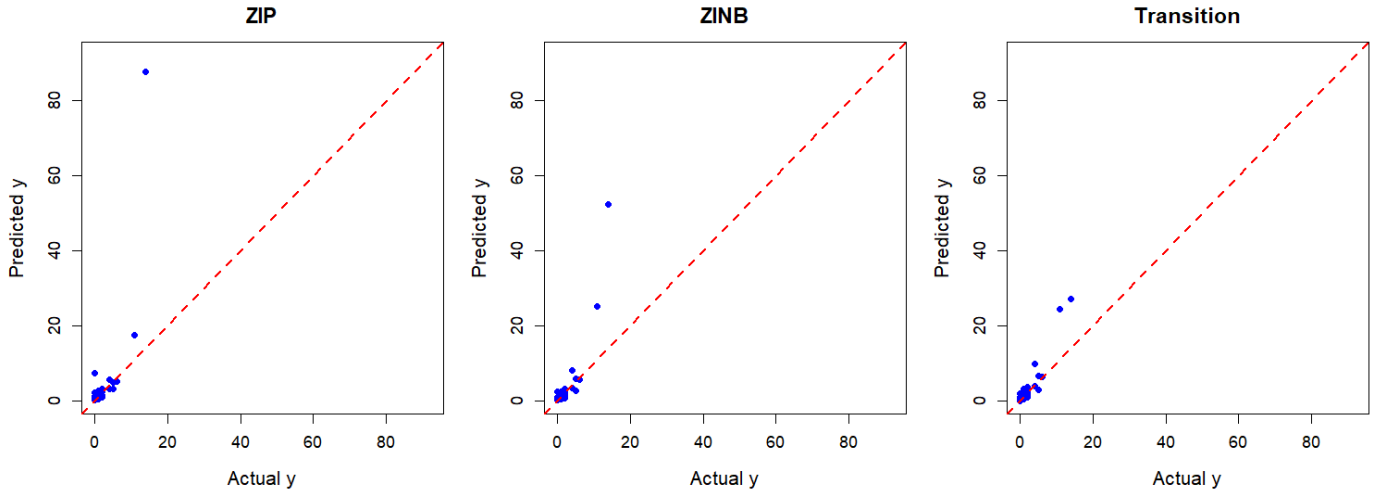


Figure 1: Prediction for real data (held-out test set) by the three Bayesian models. All three panels share a common axis range. In each panel, the dashed diagonal line represents the line of perfect prediction ($\hat{y}_i = y_i$); points closer to this line indicate more accurate predictions.

ond, alongside RMSE we report the mean absolute error, $MAE = n_{\text{test}}^{-1} \sum_i |\hat{y}_i - y_i|$, and the median absolute error, $MdAE = \text{median}_i |\hat{y}_i - y_i|$. The median absolute error is a robust criterion with a breakdown point of 50%, so a conclusion that survives under $MdAE$ cannot be produced by one or a few extreme observations. Table 5 reports these results.

The mechanism behind the single-split result is nevertheless worth stating, since it is structural rather than acciden-

tal. The ZIP and ZINB models link the mean to the covariates through an exponential function, so a test observation whose covariates lie outside the range well represented in the training data can produce a very large predicted count. The Transition model instead averages over the bounded set of count categories $0, 1, \dots, M$ defined in Section 2, so its predicted mean cannot exceed M regardless of the covariate values. This bounded support is a genuine property of the model and explains why its

Table 5: Out-of-sample predictive performance over $R = 100$ repeated random 80/20 splits of the real data: mean (standard deviation) across splits, and the percentage of splits in which each model attains the smallest test error.

Criterion	Summary	Bayesian ZIP	Bayesian ZINB	Bayesian Transition
RMSE (test)	Mean (SD)	11.74 (16.57)	18.30 (30.94)	5.01 (4.00)
	% splits best	19%	29%	52%
MAE (test)	Mean (SD)	2.68 (2.51)	3.56 (4.58)	1.60 (0.78)
MdAE (test)	Mean (SD)	0.43 (0.09)	0.34 (0.07)	0.27 (0.09)

Table 6: Mean out-of-sample RMSE over the $R = 100$ splits, separated according to whether the observation with count 149 falls in the test set.

Splits	Bayesian ZIP	Bayesian ZINB	Bayesian Transition
Without the extreme observation (80 splits)	3.64	3.08	3.10
With the extreme observation (20 splits)	44.13	79.16	12.66

Table 7: Sensitivity of the Bayesian Transition model to the treatment of the upper tail. Only this model is affected: the two zero-inflated models do not involve a terminal category and their entries are identical under both conventions.

Analysis	Absorbing, (19)	Rescaled	Max. residual mass
Simulation, ZIP data, $n = 100$	0.5978	0.5979	0.004
Simulation, ZIP data, $n = 200$	0.6419	0.6419	0.004
Simulation, ZINB data, $n = 100$	5.5154	5.5292	0.170
Simulation, ZINB data, $n = 200$	7.6604	7.6616	0.061
Simulation, AR(1) data, $n = 100$	8.7589	8.7564	0.273
Simulation, AR(1) data, $n = 200$	9.5043	9.5036	0.021
Real data, test RMSE	5.01	5.20	0.039

Entries are mean RMSE. The residual mass $P(Y > M - 1 | \mathbf{x}_i)$ is given as its maximum over observations and replications.

predictions degrade gracefully on unusual observations. It is also a limitation in the opposite direction: if the true count for a test observation exceeded M , the Transition model could not predict it.

We therefore state the conclusion of this section in terms of what the repeated splits support. Over the 100 splits the Bayesian Transition model attains the lowest mean test RMSE (5.01 against 11.74 and 18.30), the lowest mean absolute error, the lowest median absolute error, and the smallest dispersion across splits; it is also the best model in 52 of the 100 splits individually. The advantage is therefore not an artefact of averaging, as a single-split comparison could not have established.

Its source, however, is specific. Table 6 shows that on the eighty splits in which the largest observed count falls in the training set, the three models predict almost identically, with mean RMSE of 3.64, 3.08 and 3.10 – the Bayesian ZINB model marginally ahead. On the twenty splits in which that observation falls in the test set, the mean RMSE rises to 44.13, 79.16 and 12.66. The stratification therefore locates the advantage rather than explaining it away: the margin widens sharply with the extremity of the observation being predicted, from a small difference on squared error once the largest count is removed to a factor of three or six when it is present.

It does not, however, confine the advantage to the tail. The

median absolute error, which is the median of the errors within a split and so reflects the typical observation rather than the extremes, is 0.27 for the Transition model against 0.43 and 0.34 for the ZIP and ZINB models. With one hundred splits the standard error of that difference is about 0.011, so the gap is on the order of six standard errors and cannot reasonably be attributed to a minority of large errors. The Transition model predicts the bulk of the observations more accurately as well, and its robustness on extreme cases is not purchased at the cost of accuracy on ordinary ones.

One further feature deserves comment. The Bayesian ZINB model fails more severely than the Bayesian ZIP model on the extreme splits, by roughly a factor of two, even though it is the more flexible of the two specifications; the additional dispersion parameter widens the fitted predictive distribution and so amplifies rather than moderates the extrapolation.

It is worth asking why the real data separate the models when none of the three simulation scenarios does. The most visible difference is the range of the counts: the fish data run from zero to 149, whereas the largest counts in the three simulated mechanisms are of order three, sixty and one hundred and ten, with the covariates in the first scenario very tightly distributed. A model that represents the count distribution through a sequence of transitions has more to represent, and correspond-

ingly more to gain, when the counts are widely spread. We do not test this explanation here, and offer it as the reading of the evidence we find most plausible rather than as a result.

5. Conclusion

This study proposed a Bayesian formulation of the transition model for count data and compared it with the Bayesian ZIP and Bayesian ZINB models across three simulated data-generating scenarios and one real dataset, all characterised by zero inflation and, in two of the four cases, over-dispersion. The finding is that on the simulated data the transition model predicts the expected count at least as well as both alternatives, including in the two scenarios where one of them generated the data, and that on the real data it predicts better on every criterion examined, by a margin that widens with the extremity of the observation. This concerns prediction on the expected-count scale only, and is not a claim of superior parameter estimation.

That restriction is not a matter of caution but of what the comparison can support. The coefficients of the Transition model enter conditional transition probabilities through a probit link, whereas the ZIP and ZINB coefficients enter the log mean of a count component, and the values fixed in each data-generating mechanism are defined on the latter scale only. The two coefficient vectors are parameters of different functionals of the data-generating distribution and do not share an estimand, so differences between a Transition-model posterior mean and a generating ZIP or ZINB coefficient are not biases. Coefficient-level bias and MSE are therefore reported in Tables 2–4 only for the two zero-inflated models, and no ranking of the three models at the coefficient level is offered.

Comparison across all three is carried out through RMSE on the expected-count scale, which is a functional of the fitted predictive distribution and hence invariant to parameterisation. In computing it, the fitted value for the zero-inflated models is the posterior mean of $(1 - \pi)\lambda_i$, the unconditional expected count, rather than of λ_i , which is the mean of the count component conditional on the observation not being a structural zero.

Under zero-inflated Poisson data the mean RMSE of the three models differed by 0.003 at $n = 100$ and 0.001 at $n = 200$. Under zero-inflated negative binomial data the differences were likewise small relative to the replication variance, and the correctly specified Bayesian ZINB model did not attain the lowest value in its own scenario. Under the Poisson parameter-driven mechanism, in which an AR(1) latent process is present in the data and represented by none of the fitted models, the three again predicted equivalently; this is the scenario in which the transition model's freedom from a distributional assumption might have been expected to tell, and it did not.

Across all six scenario-by-sample-size combinations the Bayesian ZIP model, the simplest of the three, attained the lowest mean RMSE in five, including both in which it was misspecified, which we read as the variance cost of estimating M transition parameters offsetting the bias reduction that weaker assumptions permit. Within the two zero-inflated models, coefficient bias and MSE behaved as expected: the correctly specified model dominated in the ZINB scenario, the two were indis-

tinguishable under ZIP data, and under the AR(1) mechanism the bias of β_0 grew with sample size while its mean squared error fell, the signature of a misspecified estimator converging to the wrong value.

The real-data application is where the models part company. Over one hundred random splits the Transition model attained the lowest mean test RMSE, 5.01 against 11.74 and 18.30, the lowest mean and median absolute error, much the smallest dispersion across splits, and the best performance in fifty-two of the hundred splits taken individually. The advantage is not confined to the tail: the median absolute error, which is insensitive to a minority of large errors, is 0.27 against 0.43 and 0.34, a difference of roughly six standard errors, so the transition model predicts the typical observation more accurately as well. What the stratification shows is that the margin widens sharply with the extremity of the observation.

On the eighty splits in which the largest observed count falls in the training set, the mean RMSE is 3.64, 3.08 and 3.10, so on squared error the three models are close once the most extreme case is removed; on the twenty splits in which it falls in the test set the values rise to 44.13, 79.16 and 12.66. The reason is structural: the exponential mean function of the ZIP and ZINB models permits unbounded extrapolation, whereas the Transition model averages over the bounded set of categories $0, \dots, M$ and cannot. That bound is also a limitation, since counts above M cannot be predicted at all. We note in passing that the more flexible of the two zero-inflated models fails the more severely, the additional dispersion parameter widening the predictive distribution and amplifying rather than moderating the extrapolation.

Several limitations remain. The transition parameters $\theta_0, \dots, \theta_{M-1}$ are assigned independent priors, so categories with few observations are weakly identified and their posteriors are governed largely by the prior; a smoothness structure across adjacent categories, such as a random-walk prior in the spirit of the penalized likelihood approach of Berger and Tutz [19], is a natural next step. The terminal-category convention that closes the support at M is a modelling choice; Table 7 shows that it does not matter for these data, but we have not established that it is generally innocuous, and the residual mass reached 0.27 in one scenario. Finally, the zero-inflation probability π was held constant across observations in both zero-inflated models rather than linked to covariates, and relaxing this would give the competitors additional flexibility.

What the study supports is narrower than we expected when we began it, though not as narrow as the simulations alone would suggest. A Bayesian treatment of the transition model is straightforward to construct and to sample from, and it makes no distributional assumption about the count component. On data whose counts occupy a short range, as in all three simulation scenarios, that freedom is free: the transition model performs at least as well as both competitors everywhere, which means it equals a correctly specified parametric model in each of the first two scenarios without sharing its knowledge of the generating family. The extra transition parameters it must estimate, many of them informed by few observations, evidently cost little enough not to show even against the best predictor

available.

On the fish data, whose counts span two orders of magnitude and whose covariates push the fitted mean well outside the range seen in training, it buys a good deal, and it does so across the distribution rather than only in the tail. The natural reading is that the transition mechanism earns its parameters when the counts are widely spread, and that its bounded support, which is a liability when a count above M is the correct answer, is an asset when the alternative is unbounded extrapolation. A study designed to vary the count range systematically would test that reading directly, and we regard it as the most useful next step.

Data availability

The real dataset analysed is publicly available through the UCLA Statistical Consulting Group (OARC) at <https://stats.oarc.ucla.edu/r/dae/zip/> [20]. The simulation code used to generate the simulated datasets and apply the Bayesian MCMC algorithms described in this study is available from the corresponding author upon reasonable request.

Declaration of competing interest

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

Funding

The authors received no specific funding from any public, commercial, or not-for-profit organisation for this research.

References

- [1] A. C. Cameron & P. K. Trivedi, *Regression analysis of count data*, 2nd ed., Cambridge University Press, Cambridge, UK, 2013. <https://doi.org/10.1017/CBO9781139013567>.
- [2] J. M. Hilbe, *Negative binomial regression*, 2nd ed., Cambridge University Press, Cambridge, UK, 2011. <https://doi.org/10.1017/CBO9780511973420>.
- [3] C. X. Feng, "A comparison of zero-inflated and hurdle models for modeling zero-inflated count data", *Journal of Statistical Distributions and Applications* **8** (2021) 8. <https://doi.org/10.1186/s40488-021-00121-4>.
- [4] J. M. Hilbe, "The statistical analysis of count data/El análisis estadístico de los datos de recuento", *Culture and Education* **29** (2017) 409. <https://doi.org/10.1080/11356405.2017.1368162>.
- [5] Y. Min & A. Agresti, "Random effect models for repeated measures of zero-inflated count data", *Statistical Modelling* **5** (2005) 1. <https://doi.org/10.1191/1471082X05st084oa>.
- [6] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev & G. M. Smith, "Zero-truncated and zero-inflated models for count data", in *Mixed effects models and extensions in ecology with R*, Springer, New York, USA, 2009, pp. 261–293. https://doi.org/10.1007/978-0-387-87458-6_11.
- [7] S. Yang & G. Berdine, "The negative binomial regression", *The South-west Respiratory and Critical Care Chronicles* **3** (2015) 50. <https://doi.org/10.12746/swrccc2015.0310.135>.
- [8] D. Lambert, "Zero-inflated Poisson regression, with an application to defects in manufacturing", *Technometrics* **34** (1992) 1. <https://doi.org/10.2307/1269547>.
- [9] Y. Aliyu & U. Usman, "On bivariate Nadarajah–Haghighi distribution derived from Farlie–Gumbel–Morgenstern copula in the presence of covariates", *Journal of the Nigerian Society of Physical Sciences* **5** (2023) 871. <https://doi.org/10.46481/jnsps.2023.871>.
- [10] M. Ridout, C. G. B. Demétrio & J. Hinde, *Models for count data with many zeros*, Proceedings of the XIXth International Biometric Conference, Cape Town, South Africa, 1998, pp. 179–192. https://www.kent.ac.uk/IMS/personal/msr/webfiles/zip/ibc_fin.pdf.
- [11] J. Mullahy, "Specification and testing of some modified count data models", *Journal of Econometrics* **33** (1986) 341. [https://doi.org/10.1016/0304-4076\(86\)90002-3](https://doi.org/10.1016/0304-4076(86)90002-3).
- [12] M. Xie, B. He & T. N. Goh, "Zero-inflated Poisson model in statistical process control", *Computational Statistics & Data Analysis* **38** (2001) 191. [https://doi.org/10.1016/S0167-9473\(01\)00033-0](https://doi.org/10.1016/S0167-9473(01)00033-0).
- [13] T. Loeys, B. Moerkerke, O. De Smet & A. Buysse, "The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression", *British Journal of Mathematical and Statistical Psychology* **65** (2012) 163. <https://doi.org/10.1111/j.2044-8317.2011.02031.x>.
- [14] W. H. Greene, "Accounting for excess zeros and sample selection in Poisson and negative binomial regression models", NYU Working Paper No. EC-94-10, Department of Economics, Stern School of Business, New York University, New York, USA, 1994. <https://ssrn.com/abstract=1293115>.
- [15] J.-F. Angers & A. Biswas, "A Bayesian analysis of zero-inflated generalized Poisson model", *Computational Statistics & Data Analysis* **42** (2003) 37. [https://doi.org/10.1016/S0167-9473\(02\)00154-8](https://doi.org/10.1016/S0167-9473(02)00154-8).
- [16] H.-Y. Xu, M. Xie & T. N. Goh, "Objective Bayes analysis of zero-inflated Poisson distribution with application to healthcare data", *IIE Transactions* **46** (2014) 843. <https://doi.org/10.1080/0740817X.2013.770190>.
- [17] A. M. Garay, V. H. Lachos & H. Bolfarine, "Bayesian estimation and case influence diagnostics for the zero-inflated negative binomial regression model", *Journal of Applied Statistics* **42** (2015) 1148. <https://doi.org/10.1080/02664763.2014.995610>.
- [18] O. S. Adesina, "Bayesian multilevel models for count data", *Journal of the Nigerian Society of Physical Sciences* **3** (2021) 224. <https://doi.org/10.46481/jnsps.2021.168>.
- [19] M. Berger & G. Tutz, "Transition models for count data: a flexible alternative to fixed distribution models", *Statistical Methods & Applications* **30** (2021) 1259. <https://doi.org/10.1007/s10260-021-00558-6>.
- [20] UCLA Statistical Consulting Group, "Zero-inflated Poisson regression", R Data Analysis Examples, University of California, Los Angeles, n.d. Available online: <https://stats.oarc.ucla.edu/r/dae/zip/>.