

University of Groningen

## Modeling dragonfly population data with a Bayesian bivariate geometric mixed-effects model

van Oppen, Yulan B.; Milder-Mulderij, Gabi; Brochard, Christophe; Wiggers, Rink; de Vries, Saskia; Krijnen, Wim P.; Grzegorzcyk, Marco A.

*Published in:*  
Journal of Applied Statistics

*DOI:*  
[10.1080/02664763.2022.2068513](https://doi.org/10.1080/02664763.2022.2068513)

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*  
Publisher's PDF, also known as Version of record

*Publication date:*  
2023

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

van Oppen, Y. B., Milder-Mulderij, G., Brochard, C., Wiggers, R., de Vries, S., Krijnen, W. P., & Grzegorzcyk, M. A. (2023). Modeling dragonfly population data with a Bayesian bivariate geometric mixed-effects model. *Journal of Applied Statistics*, 50(10), 2171–2193. Advance online publication. <https://doi.org/10.1080/02664763.2022.2068513>

### Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

### Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

*Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.*



# Modeling dragonfly population data with a Bayesian bivariate geometric mixed-effects model

Yulan B. van Oppen, Gabi Milder-Mulderij, Christophe Brochard, Rink Wiggers, Saskia de Vries, Wim P. Krijnen & Marco A. Grzegorzcyk

To cite this article: Yulan B. van Oppen, Gabi Milder-Mulderij, Christophe Brochard, Rink Wiggers, Saskia de Vries, Wim P. Krijnen & Marco A. Grzegorzcyk (2023) Modeling dragonfly population data with a Bayesian bivariate geometric mixed-effects model, Journal of Applied Statistics, 50:10, 2171-2193, DOI: [10.1080/02664763.2022.2068513](https://doi.org/10.1080/02664763.2022.2068513)

To link to this article: <https://doi.org/10.1080/02664763.2022.2068513>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



View supplementary material [↗](#)



Published online: 06 May 2022.



Submit your article to this journal [↗](#)



Article views: 1003




View related articles [↗](#)



View Crossmark data [↗](#)

# Modeling dragonfly population data with a Bayesian bivariate geometric mixed-effects model

Yulan B. van Oppen <sup>a</sup>, Gabi Milder-Mulderij<sup>b</sup>, Christophe Brochard<sup>b</sup>, Rink Wiggers<sup>b</sup>, Saskia de Vries<sup>b</sup>, Wim P. Krijnen<sup>c</sup> and Marco A. Grzegorzczuk<sup>c</sup>

<sup>a</sup>Groningen Biomolecular Sciences and Biotechnology Institute, Groningen University, Groningen Netherlands; <sup>b</sup>Bureau Biota, Groningen, Netherlands; <sup>c</sup>Bernoulli Institute, Groningen University, Groningen, Netherlands

## ABSTRACT

We develop a generalized linear mixed model (GLMM) for bivariate count responses for statistically analyzing dragonfly population data from the Northern Netherlands. The populations of the threatened dragonfly species *Aeshna viridis* were counted in the years 2015–2018 at 17 different locations (ponds and ditches). Two different widely applied population size measures were used to quantify the population sizes, namely the number of found exoskeletons ('exuviae') and the number of spotted egg-laying females were counted. Since both measures (responses) led to many zero counts but also feature very large counts, our GLMM model builds on a zero-inflated bivariate geometric (ZIBGe) distribution, for which we show that it can be easily parameterized in terms of a correlation parameter and its two marginal medians. We model the medians with linear combinations of fixed (environmental covariates) and random (location-specific intercepts) effects. Modeling the medians yields a decreased sensitivity to overly large counts; in particular, in light of growing marginal zero inflation rates. Because of the relatively small sample size ( $n = 114$ ) we follow a Bayesian modeling approach and use Metropolis-Hastings Markov Chain Monte Carlo (MCMC) simulations for generating posterior samples.

## ARTICLE HISTORY

Received 13 October 2021  
Accepted 18 April 2022


## KEYWORDS

Bayesian modeling;  
generalized linear model (GLM); mixed effects;  
bivariate geometric distribution; count data;  
*Aeshna viridis*

## 1. Introduction

*Generalized linear mixed models (GLMMs)* are a popular statistical tool for modeling ecological data; see, e.g. [19,24,28,29]. The framework of generalized linear models (GLMs) is needed since many ecological responses are not continuous but binary, counts, or proportions [2]. Another typical characteristic is that ecological data often contain repeated observations on the same measurement units. This yields non-trivial dependencies among the observations that can be accounted for by including fixed and random (mixed) effects, leading to GLMMs. For ecological count data, two more common characteristics are the

**CONTACT** Yulan van Oppen  [y.b.van.oppen@rug.nl](mailto:y.b.van.oppen@rug.nl)

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/02664763.2022.2068513>

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

presence of unreasonably many zero counts ('zero inflation') and the presence of unreasonably large counts ('outliers'). For univariate responses, there are many well-established methods to deal with these data features. However, in some applications, the data contain bivariate (or multivariate) responses. One sub-optimal approach is then to apply a univariate GLMM to each response separately. The disadvantage is that intrinsic dependencies (correlations) between the responses are ignored. A more natural and methodologically preferred approach is to work with multivariate response distributions. Nonetheless, the conceptual problem is that many of the proposed multivariate models do not easily allow to account for zero inflation and/or large outliers and/or random effects to be integrated. Henceforth, some of the required model features might get lost when switching from a univariate to a bi- or multivariate GLMM.

In this paper, we present a new GLMM for bivariate count data. Our model assumes that the data follow a bivariate geometric distribution and we design the model such that we can easily integrate all the important features of the univariate GLMMs.

The data that motivated us to design the new model stem from a recent ecological study on *Aeshna viridis* ('green hawker') dragonfly populations in the Northern Netherlands.<sup>1</sup> For four consecutive years (2015–2018), field studies were performed at various locations (ponds and ditches). The dragonflies were counted in two distinct ways: the traditional way of counting live specimens, but also by collecting and counting the exoskeletons (exuviae) shed by the dragonflies upon transition from the larval to the adult stage. In the ecological literature, there is some disagreement on which count type yields more representative/reliable results [13]. For the data analysis, we thus designed a bivariate GLMM to model both population measures simultaneously while taking potential correlations between them into account. The available dragonfly population data set (see Section 5.1) is of relatively small size ( $n = 114$ ), includes repeated measurements (at  $m = 17$  locations), and contains many zero counts (potential zero inflation) as well as large counts (potential outliers). These data characteristics urged the need for designing a tailored model for the data analysis.

We design the model in successive steps. First, we derive the probability mass function (PMF) of a bivariate geometric (BGe) distribution. Then we show that the BGe distribution can be easily parameterized in terms of its marginal medians and a correlation parameter. Modeling the medians yields a decreased sensitivity towards large outliers. Next, we extend the BGe distribution to a zero-inflated bivariate geometric (ZIBGe) distribution, so as to be able to deal with unreasonably many zero counts. Finally, we use the ZIBGe distribution to build the likelihood of a GLMM, which we refer to as ZIBGe-GLMM. The advantage of the ZIBGe-GLMM model is that the underlying ZIBGe distribution can intrinsically cope with zero inflation and outliers, and that it is conceptually straightforward to model the marginal medians based on fixed effects (here: environmental factors) and random effects (here: location-specific intercepts). We follow a Bayesian approach and use Markov Chain Monte Carlo (MCMC) simulations for model inference. For a detailed model description, we refer to Section 2.

In the literature, many bivariate distributions for count data have been proposed, and different derivation techniques have been used to extend univariate to bivariate count distributions. The existing bivariate count distributions have different strengths and weaknesses when dealing with data features, such as zero inflation, outliers, and/or overdispersion issues. In particular, bivariate Poisson distributions have become very popular. For

example, [1] propose a bivariate Poisson distribution derived from conditional probabilities of two dependent random counts, while [9] use a multiplicative factor parameter for defining a bivariate Poisson distribution. And in [16], the so-called Sarmanov approach [25] has been used to derive a bivariate Poisson distribution. When following the Sarmanov approach, the bivariate probability mass function (PMF) is obtained by introducing a multiplicative coupling term between the univariate PMFs.

In [8], the Sarmanov approach has been used for deriving a bivariate (zero-inflated) exponentiated-exponential distribution. Although widely applicable, the Sarmanov approach comes with the technical difficulty that it can lead to negative probability masses, which can only be avoided by imposing tailored restrictions on the parameter spaces; for discussions on this see, e.g. [16]. In [6], integration was used to derive a bivariate negative binomial distribution that can additionally include zero inflation along the coordinate axes.

In this paper, we use the bivariate geometric (BGe) distribution from [18] to build a GLMM model for the dragonfly data. An important property of the BGe distribution is that it can be easily parameterized in terms of only three parameters, namely the two marginal medians and a correlation parameter (cf. Section 2). These three parameters have intuitive meanings and are thus very easy to interpret. When compared with the competing distributions for bivariate count data (see above), we see the following advantages of the BGe distribution:

- For the dragonfly data, we observe marginal modes close to zero and very large tails (cf. right panel of Figure 3). These data features are not compatible with the characteristic shape of Poisson distributions. Hence, it can be expected that the geometric distribution yields a better fit to the data. To provide empirical evidence for this claim, we compare our model with the BZIP model from [22] which employs the bivariate Poisson distribution (cf. Section 5.2.3).
- The bivariate exponentiated-exponential distribution from [8] and the bivariate negative Binomial distribution from [6] require additional overdispersion parameters (to accommodate large outliers), and both distributions cannot be parameterized such that there is an explicit correlation parameter. On the other hand, the three parameters of the bivariate geometric (BGe) distribution from [18] are easy to interpret and allow us to explicitly model the correlation between the two components.<sup>2</sup>

Our modeling approach was not only motivated by but also borrows and combines ideas from several works. For example, the idea to parameterize the BGe distribution in terms of its marginal medians has been inspired by [3], where a discretized univariate Weibull distribution has been parameterized in terms of its median, so as to make it more robust to outliers. Moreover, we follow [22] and borrow the idea from [20] when extending the bivariate geometric (BGe) distribution from [18] to a zero-inflated BGe (ZIBGe) distribution. We then use the ZIBGe distribution to build the GLMM likelihood, and like [10], we make use of random effects to account for repeated measurements on the same units.

The remainder of this paper is organized as follows. In Section 2, we derive the new ZIBGe-GLMM, and in Section 3, we provide the implementation details. The results of a small study on synthetic data, which serves as a proof of concept, are presented in

Section 4. In Section 5, we describe and statistically analyze the *A. viridis* dragonfly population data. In Section 6, we conclude with a short discussion. A formal proof, theoretical considerations, and a map of the study locations have been delegated to the Supplementary Material.

## 2. Methods

A robust Bayesian generalized linear mixed model (GLMM) for univariate count response data has been proposed in [3]. The model employs a univariate discretized Weibull distribution, which has the advantage that it can be parameterized in terms of its median. By modeling the median, the model facilitates decreased sensitivity to extreme outliers. In this paper, we adapt the idea of [3] to define a new GLMM for bivariate count responses. To this end, we replace the univariate discretized Weibull distribution from [3] by a bivariate geometric distribution, and we propose further to parameterize it in terms of its marginal medians.

### 2.1. The geometric (Ge) distribution and its ‘continuitized’ median

The *geometric distribution* describes the distribution of the number of failures when performing independent experiments with success probability  $q \in [0, 1]$  until a success is observed, symbolically we write:  $Ge(q)$ . Its probability mass function (PMF) and its cumulative distribution function (CDF) are given by

$$p(x | q) = (1 - q)^x q \quad (x \in \mathbb{N}_0) \quad \text{and} \quad F_q(x) = 1 - (1 - q)^{x+1} \quad (x \in \mathbb{N}_0),$$

where  $\mathbb{N}_0$  denotes the set of natural numbers together with 0. As every parameter  $q$  implies a unique mean,  $\mu := \frac{1}{q} - 1 \Leftrightarrow q = \frac{1}{\mu+1}$ , the geometric distribution can as well be parameterized in terms of  $\mu$ , symbolically  $Ge(\mu)$ . The median  $[M]$  of the geometric distribution is<sup>3</sup>

$$[M] = \left\lceil \frac{-1}{\log_2(1 - q)} - 1 \right\rceil = \left\lceil \frac{1}{\log_2\left(\frac{\mu+1}{\mu}\right)} - 1 \right\rceil, \tag{1}$$

where  $\lceil \cdot \rceil$  is the ceiling function. But unlike for the mean  $\mu$ , there is no one-to-one mapping between  $q$  and  $[M]$ ; different parameters  $q$  yield the same median  $[M] \in \mathbb{N}_0$ . Therefore, we introduce the concept of the ‘continuitized’ (continuous) median  $M \in \mathbb{R}_0^+$  ( $\mathbb{R}_0^+$  denotes the set of non-negative real numbers), which we define via the relationship

$$F_q(M) = \frac{1}{2} \Leftrightarrow M = \frac{1}{\log_2\left(\frac{\mu+1}{\mu}\right)} - 1.$$

Since there is a one-to-one mapping between  $q$  and the continuitized median  $M$ , we can parameterize the geometric distribution in terms of  $M$ . We have  $[M] - M < 1$ , so that  $M$  is close to the true median. In our GLMM regression framework, we therefore model  $M \in \mathbb{R}_0^+$  rather than the true median  $[M] \in \mathbb{N}_0$ .

In the same way that distributions are uniquely determined by their moment-generating functions (MGFs), distributions with sample space  $\mathcal{S} = \mathbb{N}_0$  are uniquely determined by

their *probability-generating functions (PGFs)* [12, Chapter 5]. For a random variable  $X$ , the PGF is defined as  $G(s) := \mathbb{E}\{s^X\}$  with  $|s| \leq r$ , where  $r \geq 1$  is called the radius of convergence. The PGF of the  $\text{Ge}(\mu)$  geometric distribution is given by

$$G(s) = \frac{1}{1 + \mu(1 - s)} \quad (|s| \leq 1). \tag{2}$$

**2.2. A bivariate geometric distribution (BGe)**

The PGF of a bivariate random vector  $(X, Y)$  is defined as  $G(s, t) := \mathbb{E}\{s^X t^Y\}$  with  $|s|, |t| \leq r$ . In [18], a *bivariate geometric (BGe) distribution*, denoted  $\text{BGe}(\mu, \nu, \theta)$ , has been defined by extending the PGF from (2) to the bivariate PGF

$$G(s, t) = \frac{1}{(1 + \mu(1 - s))(1 + \nu(1 - t)) - \theta\mu\nu(1 - s)(1 - t)} \quad (|s|, |t| \leq 1). \tag{3}$$

The parameters  $\mu, \nu > 0$  correspond to the marginal means and the parameter  $\theta \in [0, 1]$  affects the covariance between the two components. The marginal distributions are the  $\text{Ge}(\mu)$  and the  $\text{Ge}(\nu)$  geometric distributions since

$$\mathbb{E}\{s^X\} = G(s, 1) = \frac{1}{1 + \mu(1 - s)} \quad \text{and} \quad \mathbb{E}\{t^Y\} = G(1, t) = \frac{1}{1 + \nu(1 - t)}.$$

We have  $\text{Cov}\{X, Y\} = \frac{\partial^2 G}{\partial s \partial t}(1, 1) - \mu\nu = \theta\mu\nu$ . For large  $\mu, \nu$ , the parameter  $\theta$  resembles the correlation coefficient,  $\rho$ , between the components since

$$\rho = \frac{\text{Cov}\{X, Y\}}{\sqrt{\text{Var}X}\sqrt{\text{Var}Y}} = \frac{\theta\mu\nu}{\sqrt{\mu(\mu + 1)}\sqrt{\nu(\nu + 1)}} = \theta \sqrt{\frac{\mu\nu}{(\mu + 1)(\nu + 1)}} \xrightarrow{\mu, \nu \rightarrow \infty} \theta. \tag{4}$$

The probability mass function (PMF) of the BGe distribution is uniquely defined through the PGF in (3), but the PMF has *not* been provided in [18]. As we need the PMF for computing the likelihood of the GLMM, we derive it in Supplementary Material A.1. By taking derivatives of  $G(s, t)$ , we obtain the following result:

**Proposition 2.1:** *The PGF of the  $\text{BGe}(\mu, \nu, \theta)$  distribution provided in (3) implies the following probability mass function (PMF). For  $x, y \in \mathbb{N}_0$ ,*

$$g(x, y | \mu, \nu, \theta) = \sum_{j=0}^{\min\{x, y\}} (-1)^j \binom{x + y - j}{j, x - j, y - j} \frac{\zeta^j (\mu + \zeta)^{x-j} (\nu + \zeta)^{y-j}}{(1 + \mu + \nu + \zeta)^{x+y-j+1}}, \tag{5}$$

where  $\zeta := (1 - \theta)\mu\nu$ .

**Proof:** Since the proof is long and technical, we have delegated it to Supplementary Material A.1. ■

### 2.3. A zero-inflated bivariate geometric (ZIBGe) distribution

Data are *zero-inflated* with respect to a given distribution when they contain more zeros than the distribution can support. To be able to model zero-inflated data, one can define a zero-inflated variant of the given distribution. As [20] show, zero inflation can be included for any subspace of the support for multivariate distributions. We follow this approach to define a *Zero-Inflated Bivariate Geometric (ZIBGe)* distribution as

$$\text{ZIBGe}(\mu, \nu, \theta, \boldsymbol{\pi}) \sim \begin{cases} (0, 0) & \text{with probability } \pi_1, \\ (\text{Ge}(\mu), 0) & \text{with probability } \pi_2, \\ (0, \text{Ge}(\nu)) & \text{with probability } \pi_3, \\ \text{BGe}(\mu, \nu, \theta) & \text{with probability } 1 - \pi_1 - \pi_2 - \pi_3, \end{cases} \quad (6)$$

where  $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$  is a vector of zero inflation probabilities with  $\pi_1 + \pi_2 + \pi_3 \leq 1$ . Accounting for zero inflation is particularly important when modeling data with distributions that do not feature any dispersion parameters, like the geometric distribution. For the PMF of the  $\text{ZIBGe}(\mu, \nu, \theta, \boldsymbol{\pi})$  distribution, we obtain for  $x, y \in \mathbb{N}_0$ :

$$f(x, y \mid \mu, \nu, \theta, \boldsymbol{\pi}) = \pi_1 \mathbb{1}_{\{(0,0)\}}(x, y) + \pi_2 \mathbb{1}_{\{0\}}(y) \frac{\mu^x}{(\mu + 1)^{x+1}} + \pi_3 \mathbb{1}_{\{0\}}(x) \frac{\nu^y}{(\nu + 1)^{y+1}} + (1 - \pi_1 - \pi_2 - \pi_3)g(x, y \mid \mu, \nu, \theta), \quad (7)$$

where  $\mathbb{1}_A(\cdot)$  denotes the indicator function on a set  $A$ , and the PMF  $g(x, y \mid \mu, \nu, \theta)$  was defined in (5).

### 2.4. Reparameterization in terms of continuity marginals

For the  $\text{BGe}(\mu, \nu, \theta)$  distribution, we have a one-to-one mapping between the marginal means  $\mu, \nu$  and the respective marginal continuity marginals  $M, N$ , which we defined such that they are continuous. Henceforth, we can parameterize the  $\text{BGe}(\mu, \nu, \theta)$  distribution (and so the  $\text{ZIBGe}(\mu, \nu, \theta, \boldsymbol{\pi})$  distribution) in terms of  $M, N$ , and  $\theta$ .

A zero-inflated (univariate) geometric (ZIGe) distribution with mean  $\mu$  (without zero inflation) and zero inflation probability parameter  $\pi$  has CDF

$$F_{\mu,\pi}(x) = \pi + (1 - \pi) \left( 1 - \left( 1 - \frac{1}{\mu + 1} \right)^{x+1} \right) \quad (x \in \mathbb{N}_0)$$

and its continuity marginal  $M$  solves the equation  $F_{\mu,\pi}(M) = \frac{1}{2}$ . As  $\pi \geq \frac{1}{2}$  implies  $M = 0$ , we assume that the zero-inflation probability  $\pi < \frac{1}{2}$ . Solving  $F_{\mu,\pi}(M) = \frac{1}{2}$  for  $M$  yields

$$M = \frac{\log(2(1 - \pi))}{\log\left(\frac{\mu+1}{\mu}\right)} - 1 \Leftrightarrow \mu = ((2(1 - \pi))^{\frac{1}{M+1}} - 1)^{-1}. \quad (8)$$

From (6), it can be seen that the marginal zero inflation probabilities for the first and second component of the  $\text{ZIBGe}(\mu, \nu, \theta, \boldsymbol{\pi})$  distribution are  $\pi_1 + \pi_3$  and  $\pi_1 + \pi_2$ , respectively. Using (8), we can parameterize the distribution in terms of its marginal continuity



medians  $M, N$ , from which we can easily get the marginal means:

$$\mu = ((2(1 - \pi_1 - \pi_3))^{\frac{1}{M+1}} - 1)^{-1} \quad \text{and} \quad \nu = ((2(1 - \pi_1 - \pi_2))^{\frac{1}{N+1}} - 1)^{-1}. \quad (9)$$

This reparameterization facilitates decreased sensitivity to large outliers, see Supplementary Material A.2 for an illustrative explanation. We denote the resulting distribution by  $\text{ZIBGe}(M, N, \theta, \boldsymbol{\pi})$ , whereby we impose on the zero-inflation probability parameters  $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$  the restrictions:  $\pi_1 + \pi_2 < \frac{1}{2}$  and  $\pi_1 + \pi_3 < \frac{1}{2}$ .

### 2.5. The ZIBGe-GLMM model

Consider a set of  $n$  observations of the form  $(\mathbf{y}_i, \mathbf{x}_i, z_i)$ , where each  $\mathbf{y}_i = (y_{1,i}, y_{2,i})$  is a bivariate random vector whose elements  $y_{1,i}$  and  $y_{2,i}$  refer to two potentially correlated counts, each  $\mathbf{x}_i$  is a  $k$ -dimensional vector of covariate values that may also include a constant intercept term, and each  $z_i$  is a known and observed grouping factor that divides the  $n$  data points into  $m$  groups  $\{1, \dots, m\}$ . We assume that the distribution of the vector  $\mathbf{y}_i$  depends on the covariate values  $\mathbf{x}_i$  as well as on the group label  $z_i$ . With regard to the dragonfly data (compare Section 5.1), we assume further that the discrete covariates have only a few possible levels, while the number of groups,  $m$ , is relatively large. Therefore, to keep the number of parameters low, we use random intercepts to account for systematic offsets among the  $m$  groups.

We employ the zero inflated bivariate geometric (ZIBGe) distribution from Subsection 2.3 and its continuitized median parametrization from Section 2.4 to model the bivariate count vectors  $\mathbf{y}_1, \dots, \mathbf{y}_n$  in a GLMM, and we refer to the resulting new model as the ZIBGe-GLMM model. For the likelihood, we get

$$\mathbf{y}_i \mid (M_i, N_i, \theta, \boldsymbol{\pi}) \sim \text{ZIBGe}(M_i, N_i, \theta, \boldsymbol{\pi}) \quad (i = 1, \dots, n), \quad (10)$$

where the continuitized median parameters  $M_i, N_i \in \mathbb{R}_0^+$  depend on the covariate values  $\mathbf{x}_i$  and the group label  $z_i$  ( $i = 1, \dots, n$ ). More precisely, we assume a generalized linear relationship:

$$\begin{aligned} M_i &= \exp\{\mathbf{x}_i \cdot \boldsymbol{\beta}^{(M)} + b_{z_i}^{(M)}\}, \\ N_i &= \exp\{\mathbf{x}_i \cdot \boldsymbol{\beta}^{(N)} + b_{z_i}^{(N)}\}, \end{aligned} \quad (11)$$

where  $\boldsymbol{\beta}^{(M)}, \boldsymbol{\beta}^{(N)}$  are  $k$ -dimensional vectors of fixed effects, and the subscripts  $z_i \in \{1, \dots, m\}$  of the intercepts indicate the group to which observation  $i$  belongs. As we supposed the number of groups,  $m$ , to be large, we treat the vectors of group offsets  $\mathbf{b}_j := (b_j^{(M)}, b_j^{(N)})$  ( $j = 1, \dots, m$ ) as random intercepts, for which we assume

$$\mathbf{b}_1, \dots, \mathbf{b}_m \mid \boldsymbol{\Sigma} \stackrel{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}), \quad 0 < \boldsymbol{\Sigma} \in \mathbb{R}^{2 \times 2}. \quad (12)$$

We note that the  $2m$  random intercepts allow us to account for group-specific offsets in the two marginal medians, while keeping the effective number of parameters low; the matrix  $\boldsymbol{\Sigma}$  consists of only 3 ( $< 2m$ ) free parameters.

We model the correlation parameter  $\theta$  via the logistic function

$$\theta = \frac{1}{1 + e^{-\eta}} \iff \text{logit}(\theta) = \eta, \tag{13}$$

where  $\eta \in \mathbb{R}$  is a real-valued intercept parameter, and we employ a multinomial logistic regression approach for the zero-inflation probability parameters  $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ :

$$\pi_\ell = \frac{e^{\gamma_\ell}}{2(1 + e^{\gamma_1} + e^{\gamma_2} + e^{\gamma_3})} \quad (\ell = 1, 2, 3) \iff \text{mlogit}(2\boldsymbol{\pi}) = \boldsymbol{\gamma}, \tag{14}$$

where  $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3) \in \mathbb{R}^3$  is a vector of real-valued intercept parameters. The function  $\text{mlogit}(\cdot)$  is the generalization of the logistic link.<sup>4</sup> By multiplying its argument by 2, we ensure  $\pi_1 + \pi_2 + \pi_3 < 0.5$ , so that (9) refers to a valid parameterization.

### 2.6. Inferring the ZIBGe-GLMM model

In Section 2.5, we have kept the description of the ZIBGe-GLMM model generic, and model inference can be done by following either a frequentist or a Bayesian approach. Because of the small sample size,  $n$ , we follow the Bayesian way, whereby we have to impose prior distributions on the unknown parameters. For the fixed effect regression coefficient vectors  $\boldsymbol{\beta}^{(M)}, \boldsymbol{\beta}^{(N)}$ , the correlation parameter’s intercept  $\eta$ , and the zero inflation parameters’ intercept vector  $\boldsymbol{\gamma}$ , we use Gaussian prior distributions:

$$\boldsymbol{\beta}^{(M)} \sim N(\mathbf{0}, \sigma_M^2 \mathbf{I}_k), \quad \boldsymbol{\beta}^{(N)} \sim N(\mathbf{0}, \sigma_N^2 \mathbf{I}_k), \quad \eta \sim N(0, \sigma_\theta^2), \quad \text{and} \quad \boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_\pi^2 \mathbf{I}_3), \tag{15}$$

where  $\mathbf{I}_l$  denotes the  $l \times l$  identity matrix and  $\sigma_M^2, \sigma_N^2, \sigma_\theta^2, \sigma_\pi^2 > 0$  are fixed variance hyperparameters. In the absence of genuine prior knowledge, we will select the hyperparameters such that we get uninformative prior distributions (cf. Section 3).

For the random effect covariance matrix  $\boldsymbol{\Sigma}$  in (12), we follow [3] and impose a *Multivariate Generalized Hyperbolic t* prior distribution on its inverse:

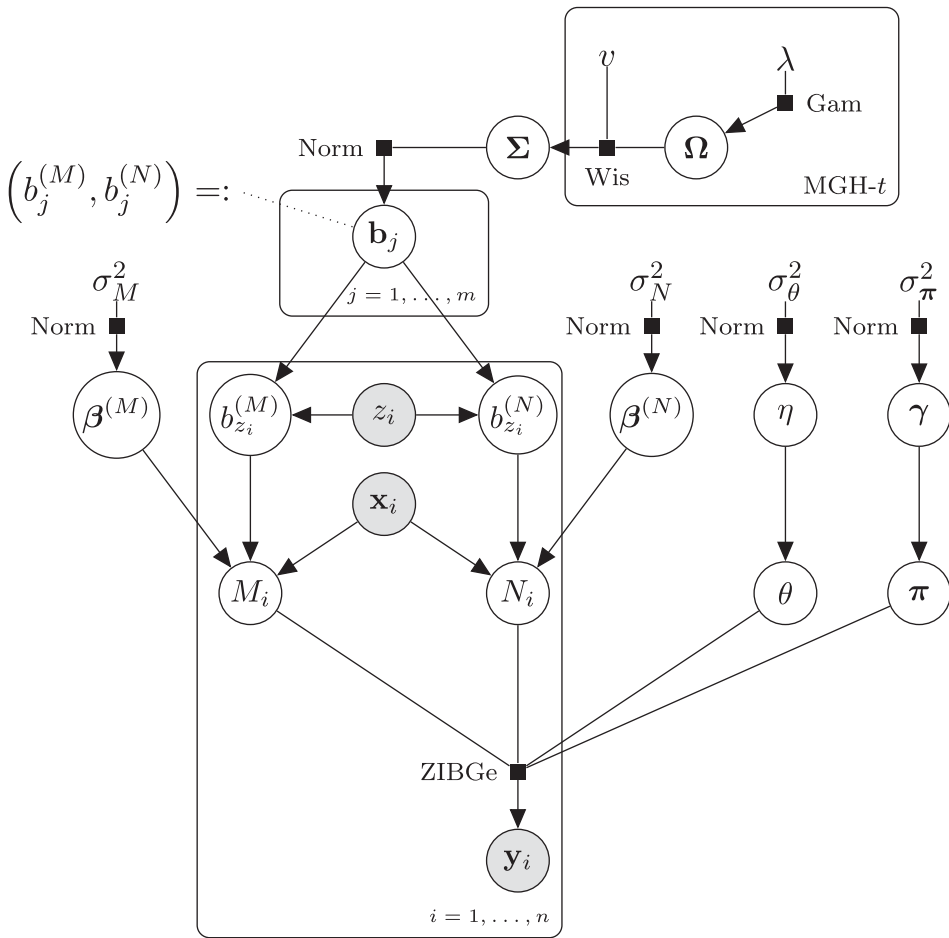
$$\boldsymbol{\Sigma}^{-1} \sim \text{MGH-t}(\lambda, \nu, d), \tag{16}$$

where  $d = 2$  is the dimension of  $\boldsymbol{\Sigma}$  and  $\lambda, \nu > 0$  are fixed hyperparameters. As argued in [17], the MGH-t distribution is less informative than an inverse-Wishart distribution in a marginal sense. The MGH-t distribution is a compound distribution defined as a Wishart distribution with scale matrix  $(2\nu\boldsymbol{\Omega})^{-1}$  and  $\nu + d - 1$  degrees of freedom, where

$$\boldsymbol{\Omega} = \text{diag}(\omega_1, \dots, \omega_d) \quad \text{with } \omega_1, \dots, \omega_d \stackrel{\text{iid}}{\sim} \Gamma(0.5, \lambda^{-2}).$$

Note that in the above,  $\lambda^{-2}$  denotes the *rate* parameter of the Gamma distribution. For the density of the joint posterior distribution of the model parameters, we obtain

$$\begin{aligned} & p(\boldsymbol{\beta}^{(M)}, \boldsymbol{\beta}^{(N)}, \eta, \boldsymbol{\gamma}, \mathbf{b}_1, \dots, \mathbf{b}_m, \boldsymbol{\Sigma}, \boldsymbol{\Omega} \mid \mathbf{y}_1, \dots, \mathbf{y}_n) \\ & \propto \left( \prod_{i=1}^n f(\mathbf{y}_i \mid M_i, N_i, \theta, \boldsymbol{\pi}) \right) p(\boldsymbol{\beta}^{(M)}) p(\boldsymbol{\beta}^{(N)}) p(\eta) p(\boldsymbol{\gamma}) \\ & \quad \times \left( \prod_{j=1}^m p(\mathbf{b}_j \mid \boldsymbol{\Sigma}) \right) p(\boldsymbol{\Sigma} \mid \boldsymbol{\Omega}) p(\boldsymbol{\Omega}), \end{aligned} \tag{17}$$



**Figure 1.** Graphical representation of the hierarchical Bayesian model specified by (10)–(17). Arrows indicated dependencies, with distributional relationships indicated by labeled squares (‘Norm’, ‘Wis’, and ‘Gam’ indicate the *normal*, *Wishart*, and *Gamma* distributions, respectively). Observed quantities are indicated in gray and hyperparameters are given as plain nodes. The top-right panel contains the hierarchical distribution that makes up the MGH– $t(\lambda, \nu, d = 2)$  prior for  $\Sigma$ .

where  $\mathbf{b}_j := (b_j^{(M)}, b_j^{(N)})$  are the random effect intercepts for group  $j$  ( $j = 1, \dots, m$ ), and  $f(y_i | M_i, N_i, \theta, \boldsymbol{\pi})$  is the PMF from (7) but parameterized through its continuitized medians from (9). The relationship between  $M_i, N_i$  and the fixed and random regression parameters were defined in (11). A graphical representation of the model is given in Figure 1. To generate parameter samples from the posterior distribution, we use a Metropolis-Hastings Markov Chain Monte Carlo (MCMC) sampling scheme. We refer to Section 3 for more technical details.

### 3. Implementation details and software availability

To generate posterior samples from the ZIBGe-GLMM model, we use the Metropolis-Hastings Markov Chain Monte Carlo (MCMC) algorithm, as implemented in the JAGS

(*Just Another Gibbs Sampler*) software [23]. JAGS is written in C++ and uses the BUGS language for model definitions [21]. JAGS has a graphical user interface with the R software environment so that it can be easily called from R. We invoke JAGS in R using the package `runjags` [5]. Our R/JAGS code is available from our GitHub repository<sup>5</sup> [27].

We employ the following hyperparameter setting (cf. Section 2.6). To achieve very uninformative fixed effect priors, we choose  $\sigma_M^2 = \sigma_N^2 = 100$ . To ensure that the priors of  $\theta$  and  $\pi$  resemble uniform distributions on  $[0, 1]$  and on the simplex on which  $\pi$  is supported, respectively, we set  $\sigma_\theta^2 = \sigma_\pi^2 = 4$ . To obtain a uniform distribution for the correlation between the components of each  $\mathbf{b}_j$ , we follow [17] and set  $\nu = 2$  in (12). Finally, setting  $\lambda = 10$  ensures moderately uninformative standard deviations.<sup>6</sup>

For generating posterior samples for the dragonfly data (cf. Section 5), we proceeded as follows: After a short adaption phase of 1 k (1000) iterations, in which JAGS automatically optimized the proposal moves, we ran eight independent Markov chains in parallel to be able to monitor and assess convergence. Each chain was run for 260 k iterations, and by setting the burn-in phase to 10 k and the thinning factor to 2500, we obtained 100 posterior samples per chain (i.e. 800 posterior samples in total). To assess convergence, we relied on trace plot diagnostics and potential scale reduction factors (PSRFs); see [11] for details. The PSRFs provided in Table 5 are below the widely applied threshold  $\psi = 1.05$ , and hence suggest sufficient convergence.

For the small simulated data sets in Section 4, 100 iterations of adaptation were satisfactory for JAGS to optimize the proposal moves. We found that 550 iterations with a burn-in phase length of 50 and thinning factor 5 already leads to sufficient convergence.

The MCMC simulations for the ZIBGe-GLMM model were run on a PC with an Intel® Core™ i7-7700HQ processor (eight 2.80 GHz cores, running one chain per core) with 16 GB of RAM running R version 4.1.0 and JAGS 4.3.0 on Ubuntu 20.04.2 LTS. Generating the simulation study results (cf. Section 4) took around 40 min. Generating the dragonfly results (cf. Section 5) took around 36 h.

## 4. Simulation study

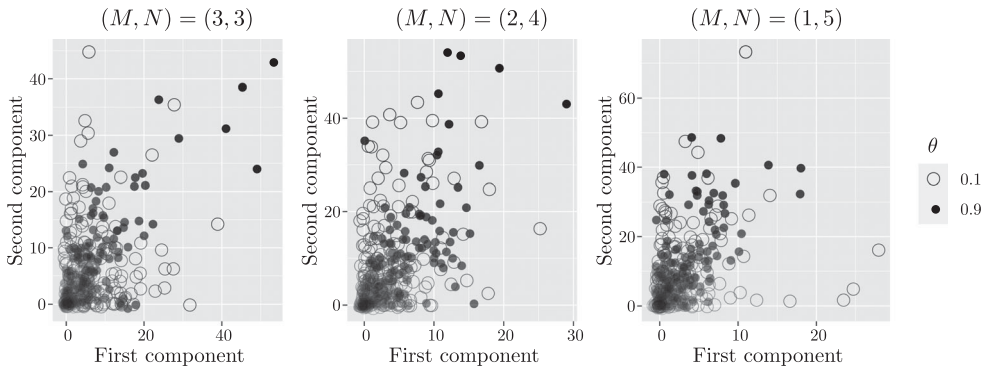
To familiarize with the proposed ZIBGe-GLMM, we performed various studies on simulated data. Here, we report the results of a small simulation study to show that the ZIBGe distribution can be inferred from data.

In (10), we use constant (i.e. without covariates or a grouping factor) continuity medians  $M_i = M$  and  $N_i = M$  across all  $n$  observations, and we use three median settings crossed with five correlation parameters as distribution parameters:

$$(M, N) \in \{(3, 3), (4, 2), (1, 5)\}, \quad \theta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}.$$

We set the three zero-inflation probability parameters to

$$\pi = (0.1, 0.05, 0.025).$$



**Figure 2.** Scatter plots of simulated data (cf. Section 4). For three  $(M, N)$  combinations, we overlaid the scatter plots for  $\theta = 0.1$  (circles) and  $\theta = 0.9$  (dots). It can be seen that the dots ( $\theta = 0.9$ ) are more strongly correlated than the circles ( $\theta = 0.1$ ). To visualize overlaid points as clusters, we added jitter and reduced the opacity of points closer to  $(0, 0)$ . The theoretical and the empirical correlations are provided in Table 1.

**Table 1.** ZIBGe parameter combinations  $(M, N, \theta)$  (in bold) and the resulting correlation coefficient  $\rho$ ; cf. (4) and (8).

$(M, N)$	<b>(3, 3)</b>		<b>(2, 4)</b>		<b>(1, 5)</b>	
	<b>0.10</b>	<b>0.90</b>	<b>0.10</b>	<b>0.90</b>	<b>0.10</b>	<b>0.90</b>
$\rho$	0.09	0.79	0.09	0.78	0.08	0.75
$\hat{\rho}$	0.13	0.76	0.31	0.70	0.23	0.66

Note: The bottom row lists the sample correlation coefficients  $\hat{\rho}$  for the data shown in Figure 2.

This translates (rounded to 2 decimal places) into the regression parameters

$$\begin{aligned}
 (\beta^{(M)}, \beta^{(N)}) &=: (\beta^{(M)}, \beta^{(N)}) \in \{(1.10, 1.10), (0.69, 1.39), (0, 1.61)\} \\
 \eta &\in \{-2.20, -0.85, 0, 0.85, 2.20\} \\
 \gamma &= (\gamma_1, \gamma_2, \gamma_3) = (-1.18, -1.87, -2.56).
 \end{aligned}$$

For each of the 15 model configurations, we generate<sup>7</sup> a data set with  $n = 200$  data points. Scatter plots of the sampled bivariate count data for the lowest ( $\theta = 0.1$ ) and highest ( $\theta = 0.9$ ) correlation parameter are shown in Figure 2. The corresponding theoretical and empirical correlation coefficients are provided in Table 1.

For each of the 15 data sets, we then generated a posterior sample (cf. Section 3). Each model has 6 parameters  $(\beta^{(M)}, \beta^{(N)}, \eta, \gamma_1, \gamma_2, \gamma_3)$  and Table 2 provides their inferred posterior medians along with 95% confidence intervals. Most of the posterior medians are close to the true parameter values and 87 out of 90 confidence intervals cover the true parameter, suggesting that the model can be inferred from data.<sup>8</sup>

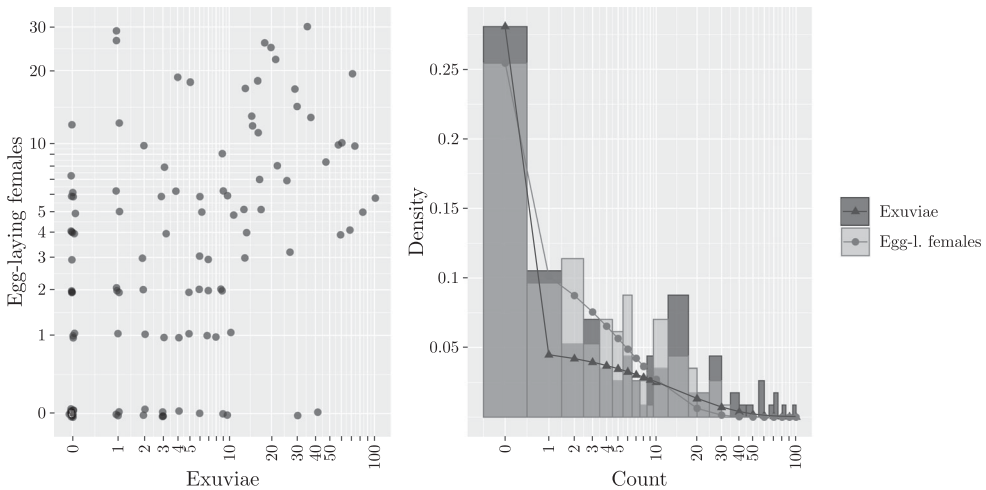
### 5. Statistical analysis of dragonfly population data

This section treats the statistical analysis of dragonfly population data from the Northern Netherlands, for which we designed the ZIBGe-GLMM model (cf. Section 2). The data are

**Table 2.** Simulation study posterior results (cf. Section 4).

Parameter	$\{\beta^{(M)}, \beta^{(N)}\}$	$\{\eta\}$				
		$\{-2.20\}$	$\{-0.85\}$	$\{0.00\}$	$\{0.85\}$	$\{2.20\}$
$\beta^{(M)}$	$\{\mathbf{1.10}, 1.10\}$	1.20 (0.95, 1.42)	0.78 (0.47, 1.08)*	0.94 (0.70, 1.18)	0.89 (0.65, 1.13)	1.04 (0.78, 1.29)
	$\{\mathbf{0.69}, 1.39\}$	0.21 (-0.26, 0.67)*	0.60 (0.30, 0.89)	0.50 (0.19, 0.84)	0.58 (0.17, 0.83)	0.59 (0.35, 0.86)
	$\{\mathbf{0.00}, 1.61\}$	-0.37 (-1.29, 0.32)	-0.17 (-0.93, 0.34)	-0.19 (-0.69, 0.28)	-0.04 (-0.44, 0.34)	-0.14 (-0.48, 0.24)
$\beta^{(N)}$	$\{1.10, \mathbf{1.10}\}$	1.07 (0.79, 1.32)	1.07 (0.75, 1.33)	0.97 (0.70, 1.27)	0.97 (0.71, 1.22)	1.10 (0.82, 1.35)
	$\{0.69, \mathbf{1.39}\}$	1.46 (1.21, 1.73)	1.33 (1.07, 1.58)	1.32 (1.08, 1.54)	1.28 (1.04, 1.51)	1.29 (1.00, 1.52)
	$\{0.00, \mathbf{1.61}\}$	1.49 (1.19, 1.75)	1.68 (1.42, 1.95)	1.59 (1.38, 1.82)	1.65 (1.44, 1.88)	1.65 (1.48, 1.85)
$\eta$ {see top row}	$\{1.10, 1.10\}$	-2.36 (-4.30, -0.83)	-0.71 (-1.74, 0.12)	-0.30 (-1.18, 0.46)	1.07 (0.38, 1.71)	1.73 (1.18, 2.34)
	$\{0.69, 1.39\}$	-0.95 (-2.16, -0.03)*	-1.22 (-2.83, -0.29)	-0.57 (-1.41, 0.24)	0.95 (0.43, 1.58)	2.13 (1.44, 2.86)
$\gamma_1$ {-1.18}	$\{0.00, 1.61\}$	-1.37 (-3.09, -0.29)	-0.81 (-2.15, 0.25)	-0.42 (-1.33, 0.38)	0.60 (0.08, 1.25)	1.51 (0.69, 2.32)
	$\{1.10, 1.10\}$	-1.24 (-1.97, -0.60)	-0.89 (-1.54, -0.08)	-1.06 (-1.61, -0.41)	-1.35 (-2.17, -0.66)	-1.17 (-1.83, -0.49)
	$\{0.69, 1.39\}$	-0.87 (-1.54, -0.14)	-1.03 (-1.64, -0.34)	-1.22 (-1.88, -0.54)	-0.95 (-1.57, -0.24)	-1.00 (-1.75, -0.34)
$\gamma_2$ {-1.87}	$\{0.00, 1.61\}$	-0.63 (-1.37, 0.26)	-0.80 (-1.67, -0.06)	-1.62 (-2.82, -0.68)	-0.92 (-1.48, -0.31)	-1.95 (-3.32, -1.09)
	$\{1.10, 1.10\}$	-1.47 (-2.69, -0.47)	-1.05 (-2.21, -0.00)	-1.78 (-3.64, -0.63)	-1.66 (-2.97, -0.60)	-1.60 (-2.48, -0.80)
	$\{0.69, 1.39\}$	-1.76 (-3.52, -0.65)	-1.30 (-2.53, -0.36)	-2.08 (-3.97, -1.04)	-2.89 (-4.93, -1.43)	-1.38 (-2.28, -0.65)
$\gamma_3$ {-2.56}	$\{0.00, 1.61\}$	-1.87 (-3.93, -0.61)	-1.41 (-2.64, -0.21)	-1.46 (-2.60, -0.52)	-3.12 (-5.04, -1.38)	-2.49 (-4.97, -1.43)
	$\{1.10, 1.10\}$	-3.34 (-5.37, -1.52)	-1.25 (-2.58, -0.04)	-3.02 (-5.46, -1.43)	-2.35 (-4.12, -0.93)	-2.38 (-4.21, -1.17)
	$\{0.69, 1.39\}$	-1.70 (-4.38, -0.32)	-2.49 (-4.99, -0.76)	-2.21 (-4.26, -0.64)	-2.41 (-4.13, -1.02)	-3.19 (-4.72, -1.70)
	$\{0.00, 1.61\}$	-1.41 (-3.49, 0.47)	-1.57 (-3.70, 0.48)	-1.82 (-4.61, -0.14)	-2.87 (-5.18, -1.35)	-2.56 (-4.61, -1.13)

Notes: For 15 models with 6 parameters each, the table provides posterior medians along with 95% posterior confidence intervals (CIs). True parameter values are in curly braces and the ones relevant for the row are in bold. Three entries, where the CI did not cover the true parameter, have been marked with asterisks. The true parameters  $\{\beta^{(M)}, \beta^{(N)}, \eta, \gamma_1, \gamma_2, \gamma_3\}$  refer to:  $(M, N)$  varying in between (3, 3), (2, 4), and (1, 5),  $\theta \in \{0.1, 0.3, \dots, 0.9\}$ , and  $\pi = (0.1, 0.05, 0.025)$ ; cf. (11), (13), and (14) for the mathematical relationships. Independent replicates of the experiment led to comparable results.



**Figure 3.** *Aeshna viridis* count distribution. Left: Scatter plot of the dragonfly count data in log-log scale; the points have been slightly jittered to reveal clusters of observations. The sample Pearson correlation coefficient is  $\hat{\rho} = 0.299$ . Right: Overlaid histograms showing the marginal count distributions along with fitted univariate zero-inflated geometric distributions in semi-log scale. The fitted zero-inflation parameter is  $\pi = 0.233$  ( $\pi = 0.137$ ) and the fitted continuitized median is  $M = 5.68$  ( $M = 2.74$ ) for the exuviae (egg-laying female) counts.

described in Section 5.1, and in Section 5.2 we report the posterior results. The dragonfly population data can be downloaded from our GitHub repository<sup>9</sup> [27].

### 5.1. Ecological background and data

*Aeshna viridis* (‘green hawker’) is a rare and threatened dragonfly species. To conserve and protect the species, in 2001, the Dutch Ministry for Agriculture, Nature, and Fisheries published a national protection plan for *A. viridis* [4]. Unlike other dragonfly species, *A. viridis* only lays its eggs into the host plant *Stratiotes aloides* (‘water soldier’), so that its presence is essential for *A. viridis*. In the Netherlands, the water soldier is commonly found in ditches that separate patches of agricultural land. But increased agricultural activity has sped up the growth of the water soldier, and its increased proliferation causes a thickened layer of sludge (formed by the decaying plants) at the bottom of the ditches. This sludge layer leads to a deteriorated water quality, damaging the ecosystem, and a decreased water depth. The decreased water depth reduces the natural frost protection of the water soldier.

To prevent the destruction and to conserve the *A. viridis* species, ecological managers have to intervene. Periodically the amount of water soldier plants has to be reduced by cleaning a maximum of 50% of the water surface.

The data contain counted *A. viridis* population sizes from 5 ecological managers covering  $m = 17$  locations (areas, ditches, or ponds) across the provinces Groningen, Friesland, and Drenthe of the Northern Netherlands. The exact locations are marked on a map in Figure B.1 of Supplementary Material B. The data stem from field studies that were performed by Bureau Biota (Groningen, NL) and financially supported by the provinces and ecological managers. To cross-compare the effects of two water-surface cleaning strategies,

**Table 3.** *Aeshna viridis* data overview: We have  $n = 114$  observations from 5 ecological managers covering  $m = 17$  locations (i.e. ditches).

Per ecological manager		Per location	Per year	
Groninger Landschap:	24	8 + 8 + 8	2015:	32
Gemeente Veendam:	22	6 + 2 + 6 + 8	2016:	28
Staatsbosbeheer Twijzel:	24	8 + 8 + 8	2017:	28
Staatsbosbeheer Groningen:	2	2	2018:	26
Waterschap Hunze en Aa's:	42	8 + 8 + 2 + 8 + 8 + 8		

Notes: The sums in the 2nd column indicate the numbers of observations per location. We have 8 observations (one per treatment per year) only if the location participated each year and all field studies took place. For example, in Groninger Landschap, there are three locations and for each, we have all 8 observations (24 in total). Gemeente (municipality) Veendam covers 4 locations, but we have only 2 (6) observations from the 2nd (3rd) location. The last column lists the number of observations per year.

which we refer to as ‘treatments’, each location was divided into two equally spaced parts. At each location, the treatments were randomly assigned to the two parts and differed in the way how 50% of the water surfaces were cleaned during the study period: (T1) ‘Clean one large rectangle-shaped area’. vs. (T2) ‘Clean small rectangle-shaped areas that are arranged in the form of a chessboard’.<sup>10</sup> The dragonflies were counted in 4 consecutive years (2015 to 2018), and in each year for each location, multiple counting sessions were scheduled.

The numbers of dragonflies were quantified in two different ways: 1. The exoskeletons (‘exuviae’) shed during metamorphosis (from larvae to adult) were collected and counted ( $y_1$ ). 2. Flying dragonflies were spotted and the egg-laying females were identified and counted ( $y_2$ ).<sup>11</sup> During each session, biologists from Bureau Biota (Groningen, NL) searched in both location parts for 45 minutes and recorded the numbers of skeletons ( $y_1$ ) and the number of egg-laying females ( $y_2$ ), yielding for each location part a bivariate count response vector  $\mathbf{y} = (y_1, y_2)$  per session. In each session, the percentage of covered water surface (*host plant emersion*) was recorded as well. From the recorded data, we computed the yearly total sums of dragonfly counts and the yearly average host plant emersion.

In addition to the dragonfly population counts, the binary treatment, and the *host plant emersion*, once per year, seven more water quality factors were measured: the *pH value*, the *redox potential*, the *oxygen concentration*, the *electrical conductivity (EC)*, the *water temperature*, the *water depth*, and the *sludge layer thickness*.

Table 3 lists the numbers of observations per ecological manager, per location, and per year. For most locations, we have 8 measurements, one for each year-treatment combination, but some have less, indicating that locations did not participate for the whole study period or that scheduled field studies could not take place.

Summary statistics on the 10 numeric variables (2 responses and 8 covariates) can be found in Table 4. The measurement units of the 8 covariates are conventional and produce covariate value ranges in comparable orders of magnitude. The right-skewed empirical distributions of the two dragonfly count measures are shown in Figure 3.

## 5.2. Posterior results for *A. viridis* dragonfly data

The data contain  $n = 114$  observations of the form  $(\mathbf{y}_i, \mathbf{x}_i, z_i)$ , each consisting of a bivariate count response vector  $\mathbf{y}_i = (y_{1,i}, y_{2,i})$ ,  $k = 13$  covariate values in  $\mathbf{x}_i$ , and a grouping factor  $z_i$  that indicates at which location  $j \in \{1, \dots, 17\}$  the observation  $i$  was made. The  $k = 13$  covariate values cover an initial ‘1’ for the intercept, the binary treatment variable,



**Table 4.** Summary statistics for the two responses ( $y_1$  and  $y_2$ ) and the 8 numeric covariates ( $x_1, \dots, x_8$ ).

		Mean	SD	Min	Median	Max
$y_1$	Exuviae	11.61	19.33	0	3	99
$y_2$	Egg-laying females	5.50	6.84	0	3	31
$x_1$	Emersion (fraction)	0.77	0.25	0.00	0.84	1.00
$x_2$	pH	7.01	0.55	4.90	6.96	8.10
$x_3$	Redox (V)	0.10	0.07	-0.09	0.11	0.28
$x_4$	Oxygen (fraction)	0.56	0.29	0.00	0.56	1.44
$x_5$	EC (mS/cm)	0.52	0.24	0.13	0.49	1.08
$x_6$	Temperature ( $^{\circ}$ C)	13.95	4.20	5.50	15.20	21.00
$x_7$	Water depth (m)	0.62	0.18	0.23	0.65	1.10
$x_8$	Sludge thickness (m)	0.31	0.15	0.06	0.28	0.75

Notes: The table provides the mean, standard deviation (SD), minimum, median, and the maximum. In addition, our model includes the binary surface-cleaning treatment variable ( $57 \times T1, 57 \times T2$ ), the categorical variable *year* with the distribution shown in the last column of Table 3, and random intercepts for the  $m = 17$  locations.

8 numeric variables, and the categorical variable *year* with four levels {2015, . . . , 2018}, which we encode via 3 dummy variables with 2015 being the reference year. The data thus fit into the framework of the ZIBGe-GLMM model from Section 2.5. For both responses ( $y_1$ : exuviae and  $y_2$ : egg-laying females), the covariates are included as fixed effects and the  $m = 17$  locations are included as random intercepts.

We generate posterior samples (cf. Section 3), and the inference results are summarized in Table 5. The table lists posterior medians along with 95% confidence intervals, potential scale reduction factors (PSRF), and fractions of positive samples ( $\mathcal{F}_+$ ).

### 5.2.1. Covariate effects

To assess how the covariates affect the two responses, we focus our attention on the response-specific fractions of positive parameters,  $\mathcal{F}_+$ . A high (low) value of  $\mathcal{F}_+$  means that the majority of sampled parameters is positive (negative), and hence suggests a consistently positive (negative) effect. Figure 4 shows a scatter plot and a grouped bar chart of the response-specific fractions  $\mathcal{F}_+$ . Its most important implications are:

- The surface cleaning *treatment* (T1 vs. T2) parameters have no systematic signs, and thus the treatment seems not to affect the dragonfly population sizes. This is consistent among both responses.
- The results show that there is a *yearly trend*. And the *year* effects seem to be rather consistent among the two responses. In particular, almost all the posterior sampled parameters for 2018 were negative, indicating that the dragonfly population sizes in 2018 were significantly smaller than in 2015 (reference year).
- We also observe consistent effects for the *oxygen* level. The results suggest for both response measures that the dragonfly population size decreases with the *oxygen* level.
- Moreover, both response measures agree in suggesting that the two covariates *redox* and *water depth* have no effect on the dragonfly population sizes.
- However, for two covariates, we observe different effects on the two count measures. This applies to the *electrical conductivity* (EC) and the *temperature*. For these two covariates, the response-specific parameter medians have different signs (cf. Table 5), and the fractions of positive parameters are relatively high for the one and relatively low for the other response. This difference suggests that the opposite direction of the effects is rather systematic.

**Table 5.** Results for dragonfly population data (cf. Section 5.1).

	Component	Lower95	Median	Upper95	PSRF	$\mathcal{F}_+$	
$(y_1)$ Exuviae:	(Intercept)	-4.11	1.20	6.38	1.017	0.672	
	Treatment	-0.86	-0.13	0.58	1.001	0.384	
	2016	-1.79	0.57	2.75	1.008	0.716	
	2017	-2.64	-1.39	-0.32	1.004	0.009	
	2018	-3.52	-1.95	-0.66	1.007	0.004	
	Emersion (fraction)	-1.09	0.85	3.11	0.999	0.795	
	pH	-0.68	0.11	0.88	1.013	0.595	
	Redox (V)	-4.31	-0.12	4.20	1.013	0.474	
	Oxygen (fraction)	-4.47	-2.66	-0.68	1.005	0.005	
	EC (mS/cm)	-3.01	-0.71	1.59	0.999	0.264	
	Temperature (°C)	-0.12	0.10	0.31	1.014	0.812	
	Water depth (m)	-3.13	0.01	2.68	1.005	0.504	
	Sludge thickness (m)	-4.92	-1.27	1.59	1.003	0.199	
	$(y_2)$ Egg-laying females:	(Intercept)	-6.53	-1.43	3.31	1.029	0.281
		Treatment	-0.77	-0.05	0.78	1.004	0.463
		2016	-3.02	-0.74	1.90	1.015	0.266
2017		-1.29	-0.22	0.76	0.998	0.355	
2018		-4.66	-2.65	-0.97	1.014	0.000	
Emersion (fraction)		-2.97	-0.68	1.99	1.020	0.319	
pH		-0.49	0.62	1.51	1.037	0.892	
Redox (V)		-3.96	0.80	4.93	1.002	0.620	
Oxygen (fraction)		-3.04	-1.18	0.81	1.015	0.101	
EC (mS/cm)		0.30	2.46	4.83	1.012	0.973	
Temperature (°C)		-0.41	-0.13	0.14	1.013	0.159	
Water depth (m)		-2.00	0.26	2.27	0.999	0.593	
Sludge thickness (m)		-2.94	-0.47	2.27	0.999	0.352	
$(\theta)$ Correlation:			0.22	0.53	0.80	1.010	
$(\pi)$ Zero inflation:		$\pi_1: (0, 0)$	0.00	0.07	0.14	1.012	
		$\pi_2: (\text{Exuviae}, 0)$	0.00	0.04	0.13	1.002	
	$\pi_3: (0, \text{Egg.females})$	0.00	0.05	0.14	1.005		
$(\Sigma)$ Random effects:	$\Sigma_{11}$ : Variance (Ex.)	0.23	1.94	5.24	0.999		
	$\Sigma_{12}$ : Covariance	-0.53	0.11	1.75	1.012		
	$\Sigma_{22}$ : Variance (Egg-l. f.)	0.00	0.09	1.76	1.012		

Notes: The 1st and 3rd columns give the bounds of 95% posterior confidence intervals, computed as the 0.025 and the 0.975 empirical quantiles of the posterior samples; the 2nd column lists the posterior median. The 4th column lists the Gelman-Rubin potential scale reduction factors (PSRFs). The last column gives the fraction of posterior samples that were positive. High (low) fractions indicate that the majority of posterior samples were positive (negative), indicating a systematic effect. Covariates for which at least 80% of the sampled parameters had the same sign have been put in bold. The relative covariate effects on the marginal medians are listed in Table 6.

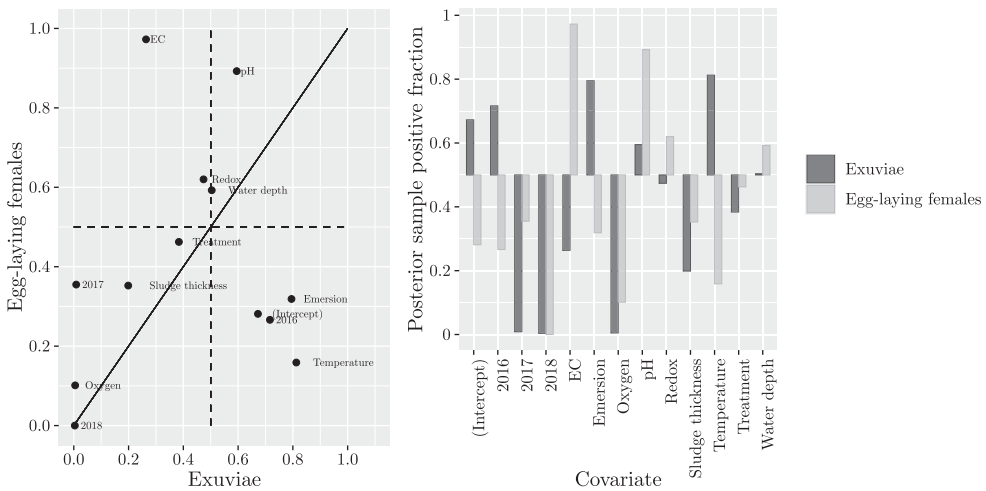
- For two covariates, we have identical signs of the parameter medians (cf. Table 5), but observe only for one of the two responses a systematic trend. The *sludge thickness* seems to have a negative effect on exuviae ( $y_1$ ) but a less pronounced negative effect on the egg-laying females ( $y_2$ ), while increasing *pH values* seem to have a positive effect on egg-laying females ( $y_2$ ) but no noteworthy effect on exuviae ( $y_1$ ).

For completeness, we provide the estimated relative covariate effects on the population size medians in Table 6. From the percentages, it can be seen that the predominantly positive ( $\mathcal{F}_+ > 0.8$ ) and negative ( $\mathcal{F}_+ < 0.2$ ) effects refer to potentially relevant changes in the dragonfly populations sizes. Most notably, for both count types, the relative difference in the population medians between 2015 and 2018 was around 90%, indicating a strong yearly trend.

**Table 6.** Results for dragonfly population data continued.

	Covariate unit (u)	Relative effects on medians	
		Exuviae	Egg-laying females
Treatment (T2)	–	–11.9%	–4.6%
Year (2016)	–	76.5%	–52.5%
Year (2017)	–	<b>–75.2%</b>	–19.6%
Year (2018)	–	<b>–85.8%</b>	<b>–92.9%</b>
Emersion	10%	8.8%	–6.6%
pH	1	12.1%	<b>86.0%</b>
Redox	100 mV	–1.2%	8.3%
Oxygen	10%	<b>–23.4%</b>	<b>–11.1%</b>
EC	100 μS/cm	–6.8%	<b>27.9%</b>
Temperature	1°C	<b>10.4%</b>	<b>–12.2%</b>
Water depth	10 cm	0.1%	2.6%
Sludge thickness	10 cm	<b>–12.0%</b>	–4.6%

Estimated relative effects on the marginal medians per covariate unit increase. Notes: For each covariate with median posterior parameter  $\beta$  (cf. Table 5), an increase of  $u$  covariate units yields the relative change  $(e^{u\beta} - 1) \cdot 100\%$ . Percentages in bold refer to covariate effects that were consistently positive ( $\mathcal{F}_+ > 0.8$ ) or negative ( $\mathcal{F}_+ < 0.2$ ); cf. Table 5.

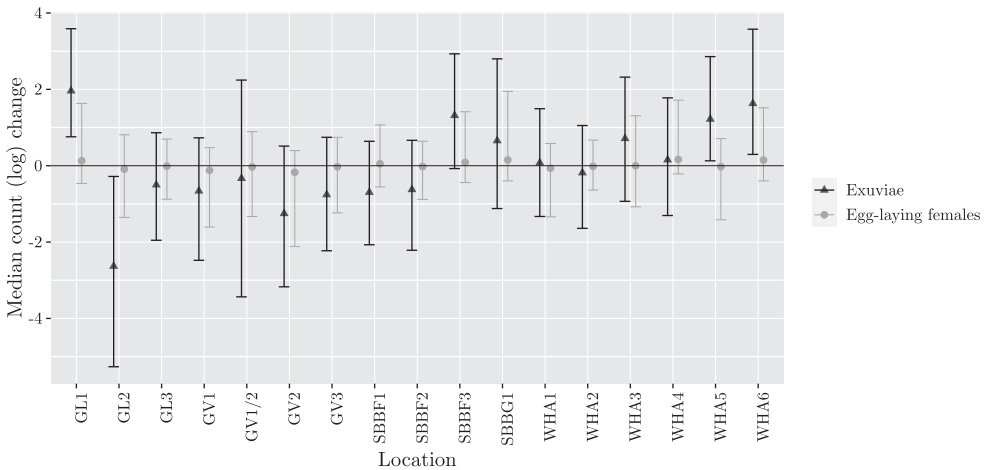


**Figure 4.** Graphical comparison of the covariate effects on the two population measures. Left: Scatter plot of the fractions of positive posterior samples for all covariates (egg-laying females vs. exuviae). A covariate effect is consistent across both population measures if the point is close to the diagonal. The dashed lines separate the positive and negative effects. Right: A grouped bar chart of the positive fractions. Values close to 1 (0) indicate significant positive (negative) effects. Covariate effects are consistent when the two bars point in the same direction and have approximately the same height.

**5.2.2. Other model parameters**

Table 5 also shows the posterior results for the correlation parameter  $\theta$ , the zero inflation probabilities  $\pi$ , and the random effect covariance matrix  $\Sigma$ .

- For the correlation parameter  $\theta$ , we have the posterior median 0.53 with confidence interval [0.22, 0.80]. This shows that the covariate effects do not fully explain the



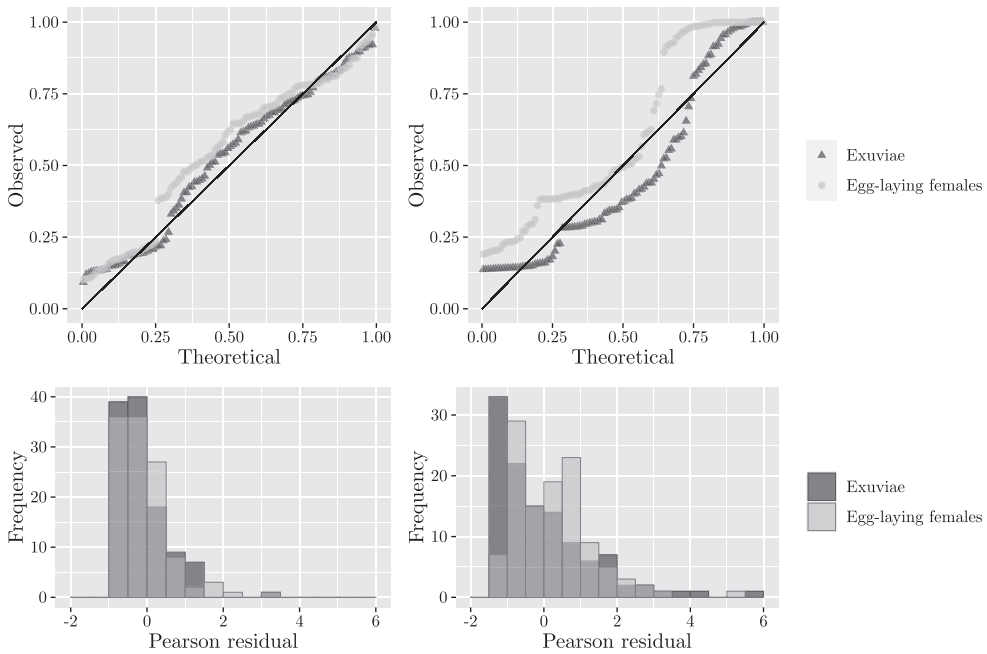
**Figure 5.** Location effects on the dragonfly population sizes. For each of the  $m = 17$  locations ( $x$ -axis), the figure shows 95% confidence intervals for the random intercept parameters for exuviae (in dark gray) and egg-laying females (in light gray). The triangles and dots mark the posterior medians.

correlation but leave a significant amount of (unexplained) correlation between the two components. This suggests that the two count types are likely to be subject to additional unobserved factors or trends. We note that the correction factor from (4) has to be taken into account when ‘translating’ the correlation parameter  $\theta$  into an actual correlation. Using the median covariate parameters from the second column of Table 5, we get that  $\theta = 0.53$  refers to the correlation  $\rho = 0.44$ , which is even higher than the sample correlation of the raw data ( $\hat{\rho} = 0.30$ ); cf. caption of Figure 3.

- The posterior zero inflation probabilities are moderate but significant. We observe the median probability  $\pi_1 = 0.07$  for both counts to be zero-inflated, and the marginal zero-inflation probability medians  $\pi_1 + \pi_2 = 0.11$  and  $\pi_1 + \pi_3 = 0.12$  for the exuviae and the egg-laying females, respectively. This shows that the data are indeed zero-inflated w.r.t. the ZIBGe-GLMM model.
- The random location intercepts for exuviae vary with a posterior median variance of 1.94, while the posterior median variance of egg-laying females is rather small (0.09). The posterior median variances (1.94 and 0.09) and covariance (0.11), yield a correlation coefficient of 0.26. Figure 5 shows posterior confidence parameters for all location intercept parameters. It can be seen that the exuviae populations are relatively large in 3–4 locations (GL1, WHA5, and WHA6 and maybe SBBF3) and relatively low in only one location (GL2). This shows that the exuviae-based population measure needs adjustment for the locations. In agreement with the small posterior variance, all location intercepts for the egg-laying females are close to 0, indicating that this population measure is not subject to location-specific effects.

**5.2.3. Model diagnostics and comparison with bivariate Poisson distribution**

To assess the fit of the ZIBGe-GLMM model to the dragonfly data, we employ the concept of randomized quantile residuals (RQRs) from [7]. For Bayesian models, the RQRs can be



**Figure 6.** Diagnostic plots. Top: Q-Q plots of response-specific randomized quantile residual (RQR) quantiles w.r.t.  $U([0, 1])$  quantiles. Bottom: Histograms of response-specific Pearson residuals, where the observations’ expectations and variances have been approximated using the respective model’s posterior sample. Left: Model diagnostics for the proposed ZIBGe-GLMM model. Right: Model diagnostics for the model from [22], which is very akin to our ZIBGe-GLMM model except that it uses a bivariate Poisson distribution to build the GLMM likelihood.

computed along the lines of [14]. Since the ZIBGe-GLMM model is a bivariate response model, we assess the fit for both response components by computing separate RQRs for the two marginal distributions.

For both response components, we have  $n = 114$  observations, and we use the posterior sampled parameters to Monte Carlo approximate the marginal predictive cumulative distribution function (CDF) at the  $n = 114$  observed response values. If there is no mismatch between model and data, the RQR quantiles are samples from a uniform distribution on the interval  $[0, 1]$ , symbolically  $U([0, 1])$ ; see [7] for details. The bottom-left panel of Figure 6 shows the two quantile-quantile (Q-Q) plots of the marginal RQR quantiles against the theoretical quantiles of the  $U([0, 1])$  distribution. Since it is hard to assess whether the two curves are ‘satisfactorily straight’, we cross-compare with the Q-Q plots of a related model. As a competitor, we use the bivariate zero-inflated Poisson (BZIP) model from [22], which is akin to our ZIBGe-GLMM model except that it uses a bivariate Poisson distribution rather than the proposed bivariate geometric (BGe) distribution from (7) to build the likelihood. Like for the ZIBGe-GLMM model, we include the environmental covariates as fixed effects and random intercepts to account for location-specific effects. Because of the resulting similarity between the two models, we refer to this variant of the BZIP model as the BZIP-GLMM model.

We generate a posterior sample for the BZIP-GLMM model (posterior results not shown) and use the sample to approximate the marginal RQR quantiles. The two resulting Q-Q plots are shown in the bottom-right panel of Figure 6. A cross-comparison reveals that the proposed ZIBGe-GLMM model (left) yields ‘more straight curves’ and so a better fit to the theoretical quantiles than the BZIP-GLMM model (right panel).

The bottom panels of Figure 6 show response-specific Pearson residual<sup>12</sup> histograms for the ZIBGe-GLMM model (left) and the BZIP-GLMM model (right). The residuals are clearly smaller for the ZIBGe-GLMM model, so this alternative diagnostic is in line with the better model fit suggested by the Q-Q plots. In the context of mixed effects, however, the Pearson residuals’ distribution can be considerably disrupted by random offsets [15], possibly leading to less reliable conclusions pertaining to model fit. To confirm the better model fit with a quantitative measure, we also use the posterior samples to compute the model-specific deviance information criteria (DIC); see [26] for details. The DIC of the proposed ZIBGe-GLMM model (DIC=1492.6) is much lower than the DIC of the BZIP-GLMM model (DIC=1925.2) so that the ZIBGe-GLMM model is also preferred in terms of the widely applied DIC criterion.

## 6. Conclusions and discussion

For the statistical analysis of an *Aeshna viridis* dragonfly population data set from the Northern Netherlands, we have proposed a new generalized linear mixed-effects (GLMM) model for bivariate count data. The proposed model uses a zero-inflated bivariate geometric (ZIBGe) distribution for building the likelihood. We have shown that the bivariate geometric (BGe) distribution can be parameterized easily in terms of three parameters, namely the two marginal medians and a correlation parameter. The advantage of this parameterization is that the three parameters have intuitive meanings and are thus very easy to interpret. Moreover, the possibility to model the medians (rather than the means) makes the model less sensitive to large counts (potential outliers). For modeling the medians, we have included environmental factors as fixed effects and we have made use of random effect intercepts to account for repeated measurements at the same locations. Given the relatively small sample size of the dragonfly data set, we have selected a Bayesian approach with Markov Chain Monte Carlo (MCMC) simulations for model inference. After a small simulation study on synthetic data, whose main purpose was to demonstrate that the model parameters can be inferred from data, we have analyzed the dragonfly data set, where the two responses refer to the number of collected exoskeletons (exuviae) and the number of spotted living specimens (egg-laying females). Both measures have been used to quantify dragonfly population sizes [13], but in the literature, we could not find any ecological study in which both quantification measures were used simultaneously and cross-compared.

The results of our statistical analysis show that the two response types can lead to different conclusions. That is, although we found the two responses (response medians) to be slightly correlated, the two responses seem to be subject to different covariate effects. Only two covariates (*yearly trend*, *oxygen*) were found to affect the two responses in the same way, while some covariates (*temperature* and *electrical conductivity*) even had opposite effects on them. The most important content-wise finding is that the surface cleaning treatment appears to have no noteworthy effect on the dragonfly population sizes. This

result is consistent between both responses. A proper ecological interpretation is beyond the scope of our work, but our results suggest that the two quantification measures differ in nature and are thus not exchangeable. Relying just on one single measure might not show the full picture and lead to biased results and erroneous conclusions. Therefore, we advise to use both measures – or at least to carefully choose the measure to quantify dragonfly population sizes – in future ecological field studies.

When studying the related literature, we found that hardly any software for modeling bivariate count data is available. To fill this gap, we decided to make our R/JAGS software available on our GitHub repository [27]. A computational bottleneck of our algorithm is the expensive computation of the bivariate geometric (BGe) likelihood. Our future work might aim to reduce the computational inference costs. One potential idea is to switch to the marginal likelihood and to approximate it by a Laplace approximation. Another interesting route of research might be to try to extend the bivariate geometric distribution to multivariate geometric distributions. In principle, the probability generating function (PGF) from [18] can easily be extended to define multivariate geometric distributions, but the mathematical challenge would be to derive the corresponding probability mass function (PMF).

## Notes

1. We provide the dragonfly population data set in our GitHub repository: <https://github.com/yulanvanoppen/ZIBGe-GLMM/tree/main/data>.
2. We note that, unlike the univariate case, the bivariate geometric (BGe) distribution from [18] is not a special case of the bivariate negative binomial (BNB) distribution from [6]. This is because the different methods to extend univariate to bivariate distributions also yield different bivariate distribution variants. Hence, the BNB distribution from [6] has a bivariate geometric distribution as special case, but this distribution differs from the BGe distribution of [18]. Only the BGe distribution from [18] can be easily parameterized such that there are two marginal median parameters and an explicit correlation parameter (cf. Section 2).
3. When  $1/\log_2(1 - q)$  is integer, the median is not unique. In this case, both  $[M]$  and  $[M] + 1$  are considered medians.
4. Specifically, the `mlogit()` link maps the simplex  $\{\mathbf{x} \in \mathbb{R}_{>0}^d \mid \|\mathbf{x}\|_1 < 1\}$  to  $\mathbb{R}^d$  through the mapping  $\mathbf{x} \mapsto \log(\mathbf{x}/(1 - \|\mathbf{x}\|_1))$ , where  $\|\cdot\|_1$  is the absolute-value norm and the logarithm is applied component-wise.
5. GitHub repository: <https://github.com/yulanvanoppen/ZIBGe-GLMM> Instructions and technical details are provided in the repository's README file.
6. For slightly varied hyperparameters, we observed very similar posterior results.
7. The script `/generate.R` in our GitHub repository [27] generates data for  $(M, N) = (3, 3)$  and  $\theta = 0.5$ .
8. We repeated the study. For newly generated data sets, we obtained similar (consistent) posterior results.
9. See <https://github.com/yulanvanoppen/ZIBGe-GLMM/tree/main/data>
10. While (T1) is easier to achieve/cheaper, (T2) is supposed to create a more natural irregular water surface.
11. Only egg-laying females tend to stay in one district/location, so their numbers can be assumed to be proportional to the true population sizes.
12. As Pearson residuals of the observed counts  $y_i$ , we use  $r_i := (y_i - \tilde{\mu}_i)/\tilde{\sigma}_i$ , where  $\tilde{\mu}_i$  and  $\tilde{\sigma}_i^2$  are Monte Carlo approximations of the expectation and the variance of  $y_i$ , which are computed from the posterior samples.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## ORCID

Yulan B. van Oppen  <http://orcid.org/0000-0001-8214-2907>

## References

- [1] P. Berkhout and E. Plug, *A bivariate Poisson count data model using conditional probabilities*, Stat. Neerl. 58 (2004), pp. 349–364.
- [2] B.M. Bolker, M.E. Brooks, C.J. Clark, S.W. Geange, J.R. Poulsen, M.H.H. Stevens, and J-S.S. White, *Generalized linear mixed models: A practical guide for ecology and evolution*, Trends Ecol. Evol. 24 (2009), pp. 127–135.
- [3] D.A. Burger, R. Schall, J.T. Ferreira, and D-G. Chen, *A robust Bayesian mixed effects approach for zero inflated and highly skewed longitudinal count data emanating from the zero inflated discrete Weibull distribution*, Stat. Med. 39 (2020), pp. 1275–1291.
- [4] T. de Jong, P.J.M. Verbeek, A.J.P. Smolders, and R.H.W. Griffioen, *Beschermingsplan groene glazenmaker 2002–2006*, Landbouw, Natuurbeheer en Visserij, 2001. Available at <https://assets.vlinderstichting.nl/docs/8d5380e8-1dd1-45a6-b5ab-c02eadcb1f77.pdf>.
- [5] M.J. Denwood, *runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS*, J. Stat. Softw. 71 (2016), pp. 1–25.
- [6] C. Dong, S.S. Nambisan, S.H. Richards, and Z. Ma, *Assessment of the effects of highway geometric design features on the frequency of truck involved crashes using bivariate regression*, Transportation Research Part A: Policy and Practice 75 (2015), pp. 30–41.
- [7] P.K. Dunn and G.K. Smyth, *Randomized quantile residuals*, J. Comput. Graph. Stat. 5 (1996), pp. 236–244.
- [8] F. Famoye, *Bivariate exponentiated-exponential geometric regression model*, Stat. Neerl. 73 (2019), pp. 434–450.
- [9] P. Faroughi and N. Ismail, *Bivariate zero-inflated generalized Poisson regression model with flexible covariance*, Commun. Stat.-Theor. Meth. 46 (2017), pp. 7769–7785.
- [10] A. Ganesalingam, A.B. Smith, C.P. Beeck, W.A. Cowling, R. Thompson, and B.R. Cullis, *A bivariate mixed model approach for the analysis of plant survival data*, Euphytica 190 (2013), pp. 371–383.
- [11] A. Gelman and D.B. Rubin, *Inference from iterative simulation using multiple sequences*, Stat. Sci. 7, Oxford University Press, Oxford, England, (1992), pp. 457–472.
- [12] G. Grimmett and D. Stirzaker, *Probability and random processes*, 2001.
- [13] S. Hardersen, S. Corezzola, G. Gheza, 2A. Dell’Otto, and G. La Porta, *Sampling and comparing odonate assemblages by means of exuviae: Statistical and methodological aspects*, J. Insect Conserv. 21 (2017), pp. 207–218.
- [14] F. Hartig, *DHARMA for Bayesians*, 2021. Available at <https://cran.r-project.org/web/packages/DHARMA/vignettes/DHARMAForBayesians.html>.
- [15] F. Hartig, *DHARMA: Residual diagnostics for hierarchical (multi-level/mixed) regression models*, 2021. Available at <https://cran.r-project.org/web/packages/DHARMA/vignettes/DHARMA.html>.
- [16] V. Hofer and J. Leitner, *A bivariate Sarmanov regression model for count data with generalised Poisson marginals*, J. Appl. Stat. 39 (2012), pp. 2599–2617.
- [17] A. Huang and M.P. Wand, *Simple marginally noninformative prior distributions for covariance matrices*, Bayesian Anal. 8 (2013), pp. 439–452.
- [18] K. Jayakumar and D.A. Mundassery, *On bivariate geometric distribution*, Statistica 67 (2007), pp. 389–404.



- [19] A. Kruger and P.J. Morin, *Predators induce morphological changes in tadpoles of *Hyla andersonii**, *Copeia* 108 (2020), pp. 316–325.
- [20] C-S. Li, J.C. Lu, J. Park, K. Kim, P.A. Brinkley, and J.P. Peterson, *Multivariate zero-inflated Poisson models and their applications*, *Technometrics* 41 (1999), pp. 29–38.
- [21] D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter, *The BUGS Book: A Practical Introduction to Bayesian Analysis*, CRC Press, Boca Raton, Florida, United States, 2012.
- [22] A. Majumdar and C. Gries, *Bivariate zero-inflated regression for count data: A Bayesian approach with application to plant counts*, *Int. J. Biostat.* 6 (2010), Article 27.
- [23] M. Plummer, *JAGS Version 4.3.0 user manual*, 2017. Available at <https://sourceforge.net/projects/mcmc-jags/files/Manuals/4x>.
- [24] T.E. Reimchen and C.A. Bergstrom, *The ecology of asymmetry in stickleback defense structures*, *Evol. Int. J. Org. Evol.* 63 (2009), pp. 115–126.
- [25] O.V. Sarmanov, *Generalized normal correlation and two-dimensional Frechet classes*, in *Doklady Akademii Nauk*, Vol. 168, Russian Academy of Sciences, St. Petersburg, Russia, 1966, pp. 32–35.
- [26] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. Van Der Linde, *Bayesian measures of model complexity and fit*, *J. R. Stat. Soc. B (Stat. Methodol.)* 64 (2002), pp. 583–639.
- [27] Y. van Oppen, *ZIBG-GLMM software package*, 2021. Available at <https://github.com/yulanvanoppen/ZIBGe-GLMM>.
- [28] X. Zhang, H. Mallick, Z. Tang, L. Zhang, X. Cui, A.K. Benson, and N. Yi, *Negative binomial mixed models for analyzing microbiome count data*, *BMC Bioinform.* 18 (2017), Article number 4.
- [29] A. Zuur, E.N. Ieno, N. Walker, A.A. Saveliev, and G.M. Smith, *Mixed Effects Models and Extensions in Ecology with R*, Springer Science & Business Media, Berlin, Berlin, Germany, 2009.