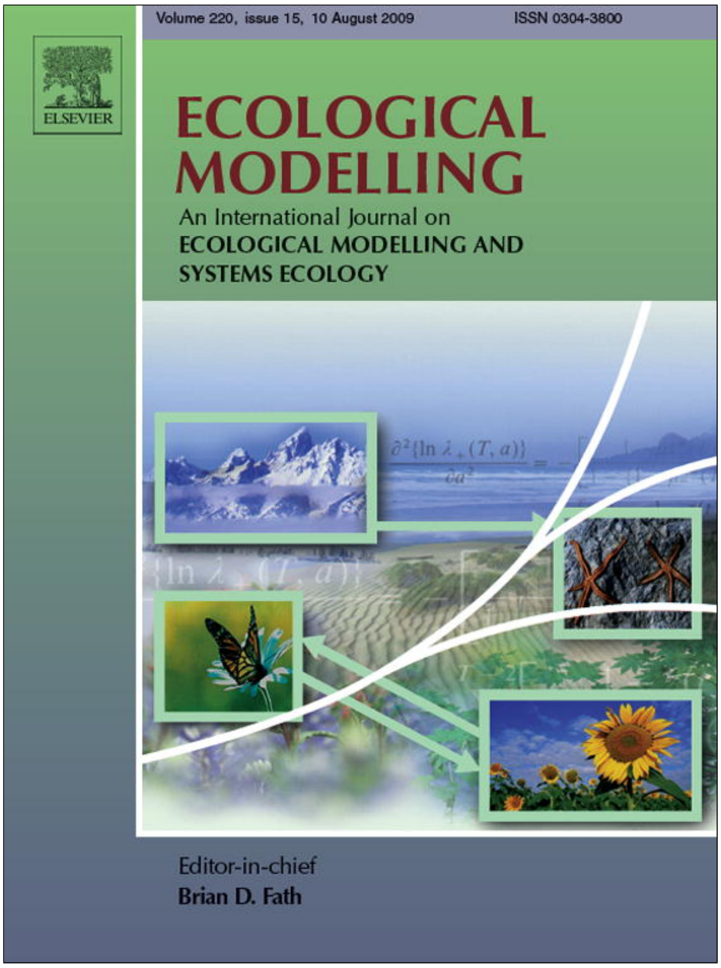
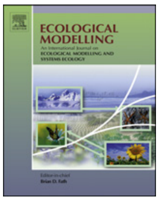

Volume 220, issue 15, 10 August 2009
ISSN 0304-3800
ELSEVIER
ECOLOGICAL MODELLING
An International Journal on
ECOLOGICAL MODELLING AND SYSTEMS ECOLOGY
Editor-in-chief
Brian D. Fath

Ecological Modelling 220 (2009) 1764–1775

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel



# Traditional occupancy–abundance models are inadequate for zero-inflated ecological count data

Gudeta Sileshi[a,*], Girma Hailu[b], Gerson I. Nyadzi[c]

[a] World Agroforestry Centre (ICRAF), SADC-ICRAF Agroforestry Programme, Chitedze Agricultural Research Station, P.O. Box 30798, Lilongwe, Malawi
[b] 2193 Espirt Dr, K4A, 4Z1, Orleans, Ontario, Canada
[c] UN Millennium Villages Project, P.O. Box 1561, Tabora, Tanzania

ARTICLE INFO

*Article history:*
Received 10 December 2007
Received in revised form 20 February 2009
Accepted 17 March 2009
Available online 21 May 2009

*Keywords:*
Model selection
Parsimony
Rare species
Repeated measures

ABSTRACT

Traditional occupancy–abundance and abundance–variance–occupancy models do not take into account zero-inflation, which occurs when sampling rare species or in correlated counts arising from repeated measures. In this paper we propose a novel approach extending occupancy–abundance relationships to zero-inflated count data. This approach involves three steps: (1) selecting distributional assumptions and parsimonious models for the count data, (2) estimating abundance, occupancy and variance parameters as functions of site- and/or time-specific covariates, and (3) modelling the occupancy–abundance relationship using the parameters estimated in step 2. Five count datasets were used for comparing standard Poisson and negative binomial distribution (NBD) occupancy–abundance models. Zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) occupancy–abundance models were introduced for the first time, and these were compared with the Poisson, NBD, He and Gaston's and Wilson and Room's abundance–variance–occupancy models. The percentage of zero counts ranged from 45 to 80% in the datasets analysed. For most of the datasets, the ZINB occupancy–abundance model performed better than the traditional Poisson, NBD and Wilson and Room's model. He and Gaston's model performed better than the ZINB in two out of the five datasets. However, the occupancy predicted by all models increased faster than the observed as density increased resulting in significant mismatch at the highest densities. Limitations of the various models are discussed, and the need for careful choice of count distributions and predictors in estimating abundance and occupancy parameter are indicated.

## 1. Introduction

A positive occupancy–abundance and abundance–variance relationship has been widely documented, both intra- and interspecifically, at a range of spatial scales for a diverse array of animal and plant species (Brown, 1984; Gaston et al., 2000, 2006; He et al., 2002; Taylor, 1961). Since the first comprehensive treatment of the occupancy–abundance relationship (Brown, 1984), it has become a general mathematical expectation that the occupancy–abundance relationship will always be positive, although occasional zero and negative correlations have been reported (Gaston et al., 2000; Wilson, 2008). This relationship has received particular attention in the context of meta-population dynamics, conservation biology, agricultural entomology and epidemiology (Anderson and May, 1985; Gaston et al., 2006; Wilson and Room, 1983).

A suite of empirical and theoretical models has been widely employed to describe the occupancy–abundance relationship in various fields (He and Gaston, 2003). However, most are special forms of the negative binomial distribution (NBD) occupancy–abundance model (He and Gaston, 2003). Recently, He and Gaston (2003) derived a general abundance–variance–occupancy model by combining the abundance–variance relationship described by Taylor's power law (TPL) (Taylor, 1961) and the NBD occupancy–abundance model. The abundance–variance–occupancy model arguably has much wider ecological significance from the perspective of pattern unification and, as such, it may help in fundamental understanding of spatial variation in abundance (He and Gaston, 2003). However, this model assumes perfect detection of species, and occupancy and abundance to be temporally and spatially invariant (He and Gaston, 2003). In their current form, all the occupancy–abundance models also do not take into account zero-inflation and its impacts on estimates of abundance, variance and occupancy from count data. Therefore, there is a need to develop more robust models that account for zero-inflation, which may arise from various sources.

A wide range of ecological count data exhibit zero-inflation (Cunningham and Lindenmayer, 2005; Gray, 2005; Martin et al., 2005; Sileshi, 2006, 2008; Warton, 2005), and such data do not readily fit standard distributions such as the NBD (Hall, 2000). Two

[*] Corresponding author. Tel.: +265 1707329.
*E-mail addresses:* sgwelde@yahoo.com (G. Sileshi), girmahailuus@yahoo.com (G. Hailu), gerson.nyadzi@undp.org (G.I. Nyadzi).

0304-3800/$ – see front matter © 2009 Published by Elsevier B.V.
doi:10.1016/j.ecolmodel.2009.03.024

types of zeros are often encountered in count data: structural zeros which are inevitable, and sampling zeros which occur by chance. Structural zeros consist of a large number of true zeros which arise when presence is not tenable (Cunningham and Lindenmayer, 2005). These are caused by the real ecological effects of interest (Martin et al., 2005). For example, the study of rare organisms will often lead to the collection of data with a high frequency of zeros (Welsh et al., 1996). Within almost all communities the vast majority of species are rare. Yet such species will frequently be of ecological, conservation or management interest in part because they may be among the extinction-prone taxa in an assemblage (Cunningham and Lindenmayer, 2005). Sampling zeros are random, and arise due to sampling where conditions are potentially suitable but absence is observed. False zeros (MacKenzie et al., 2002) occur when the species under study is present at the time of sampling, but the observer does not detect it because of its cryptic or secretive nature. Therefore, for rare species with low detection probability, excess zeroes could be substantial.

In this paper we illustrate a novel method for modelling the occupancy–abundance relationship for species with patchy distributions and, therefore, zero-inflated count datasets. We also propose two new occupancy–abundance models based on the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) distributions. We then compare these models with the traditional occupancy–abundance models derived from the Poisson and negative binomial distribution (NBD), and with two abundance–variance–occupancy models derived from Taylor's power law. We illustrate the use of information criteria for model selection, and discuss the limitations of the various models.

## 2. Methods

### 2.1. The data

Five count datasets with varying levels of zero-inflation were used in this analysis. These datasets are by no means the most representative of zero-inflated counts. They were only used to illustrate the analytical methods proposed in Section 2.2. The first dataset consisted of counts of adults of the chrysomelid beetle *Mesoplatys ochroptera* Stål in western Kenya. This species was monitored in two experiments established during 1999–2000 and 2000–2001, each consisting of three agroforestry treatments consisting of pure *Sesbania sesban* (L.) Merrill, a mixture of *S. sesban* and *Tephrosia vogelii* Hook and *S. sesban* and *Crotalaria grahamiana* Whight & Arn. The sites were Dudi and Khumusalaba in Butere district, Mutumbu in Siaya district, and Lela in Kisumu district of western Kenya (Sileshi et al., 2006). In each treatment, the abundance of *M. ochroptera* was monitored on 15 randomly selected trees that were tagged using coloured plastic strings. The numbers of adults were recorded monthly from July to December in 1999 and 2000 on each tree. This constituted two years of data collected on six dates of sampling for each site and treatment. On each date, samples were taken from the same tree, and this constituted a repeated measures dataset. The total sample size was 2546 *S. sesban* trees, of which 2035 (79.9%) had zero counts of *M. ochroptera*. Years, dates, sites and treatments were used as covariates to estimate abundance parameters in Section 2.2.1.

The second dataset consisted of counts of the invasive species *Heteropsylla cubana* Craw per shoot of eight provenances of the fodder tree *Leucaena leucocephala* (Lam) de Wit in Tanzania. To assess psyllid abundance, three terminal shoots with the next three open leaves were randomly cut from three randomly selected trees per plot. Samples were bagged in polythene bags and taken to the laboratory where the shoots were examined under a dissecting microscope and the number of psyllid nymphs per shoot was recorded. Data were recorded on six sampling dates. The total sample size was 861 shoots, of which 403 (46.8%) had zero counts. Sampling dates and provenances were used as covariates to estimate abundance parameters in Section 2.2.1.

The third dataset consisted of counts of the curculionid beetle *Diaecoderus* sp. per maize plant in eastern Zambia. Beetles were counted on 10 randomly selected maize plants in 13 agroforestry treatments in February 2002 and 2003. The treatments were replicated four times and arranged in a randomized complete blocks design. The total sample size was 990 plants, of which 444 (44.9%) were zero counts. Years and treatments were used as covariates for estimation of abundance parameters in Section 2.2.1.

The fourth dataset consisted of counts of the tenebrionid *Gonocephalum simplex* (F.) in soil monoliths from agroforestry practices in eastern Zambia. The study areas, treatment, experimental design and management of the experiments have been described in detail by Sileshi and Mafongoya (2007). Sampling was conducted three times between December 2003 and July 2004. Soil samples were collected using a soil monolith (25 cm × 25 cm and 25 cm depth) placed over a randomly selected spot, and driven into the soil to ground level using a metallic mallet. Adults were hand-sorted from the soil and counts recorded per soil monolith. The total sample size was 542 monoliths, of which 414 (76.4%) had zero counts. Sites and treatments were used as covariates to estimate the abundance parameters in Section 2.2.1.

The fifth dataset consisted of counts of the leaf beetle *Ootheca bennigseni* Weis in eastern Zambia. Beetles were monitored on bean and cowpea crops in experimental fields and two nearby farmers' fields at Msekera in February 2003. Each farm was divided into homogenous (2 m × 2 m) plots and beetle counts were recorded on 15 and 30 plants of each of bean and cowpea plants per plot in farmers' field and the experimental fields, respectively. The total sample size was 420 plants, of which 240 (68.6%) had zero counts. Fields and crops were used as covariates to estimate parameters of abundance in Section 2.2.1.

### 2.2. The modelling approach

The shape and interpretation of occupancy–abundance and abundance–variance–occupancy relationships are subject to the sampling scale (He et al., 2002). In practice, these relationships are established at some sampling scale using a range of sample mean abundance ($m$), variance ($s^2$) and occupancy ($p_o$). If the sample size is sufficiently large, $m$, $s^2$ and $p_o$ are assumed to approach the true abundance ($\mu$), variance ($\sigma^2$) and occupancy ($p_p$), respectively. For clarity, sample abundance is defined as the mean density of individuals in the sampling units (habitat patches) in which a species was recorded, and the observed occupancy as the proportion of occupied patches. When the sampling scale changes, values of $m$, $s^2$ and $p_o$ will change, and this is likely to change the model that best fits the observed data. The computation of $m$, $p_o$ and $s^2$ is sometimes done without due consideration for predictors (covariates) of $\mu$, $\sigma^2$ and $p_p$. If not done according to covariates that significantly explain these parameters, they may be biased resulting in distortion of the occupancy–abundance relationship. In this paper we propose a three-step modelling approach that will account for zero-inflation and improve accuracy in parameter estimation. The steps include (1) selecting distributional assumptions and parsimonious models for the count data, (2) estimating abundance, occupancy and variance parameters as functions of site- and/or time-specific covariates, and (3) modelling the occupancy–abundance relationship using the parameters estimated in step 2.

#### 2.2.1. Modelling abundance

The first two steps involved analysis of the datasets described above assuming Poisson, NBD, ZIP and ZINB, and jointly estimating the parameters of interest in occupancy–abundance models. The Poisson distribution is one of the two commonly used count distributions in ecology (Gray, 2005; Sileshi, 2006, 2008). It assumes that the observed counts have variance ($\sigma^2$) that is equal to the mean ($\mu$). However, count data often exhibit over-dispersion (i.e. $\sigma^2 > \mu$) relative to the Poisson assumption. Over-dispersion may derive from multiple sources, including unobserved heterogeneity, zeros in excess of those expected under a Poisson distribution and contagion (Gray, 2005). If the Poisson distribution is used, over-dispersion can lead to underestimation of the standard errors of estimates and confidence intervals that are too narrow (Sileshi, 2006, 2008).

The traditional alternative to the Poisson has been the NBD, which is a gamma mixture of Poisson responses (Gray, 2005; Sileshi, 2006, 2008). However, much controversy surrounds its dispersion parameter ($k$), which goes by the name aggregation parameter, dispersion parameter, shape parameter, clustering coefficient, etc. Different approaches have been employed to estimate $k$, namely, method of moments, extended quasi-likelihood, maximum likelihood, and bias-corrected maximum likelihood methods (Saha and Paul, 2005; Lloyd-Smith, 2007). If $k$ is not estimated accurately, the resulting occupancy–abundance relationship may be distorted.

The Poisson and NBD are inappropriate for data with many zeros (Hall, 2000). Therefore, ZIP and ZINB distributions have been proposed. These assume an unobserved or latent zero process that augments the zeros arising from the standard Poisson and NBD (Gray, 2005). The advantage of the ZIP and ZINB is that they predict a more realistic percentage of zeroes in count data than the Poisson and NBD (Gray, 2005; Sileshi, 2008). In the ZIP the population is considered to consist of two types of individuals. The first type gives Poisson distributed counts, which might contain zeros, while the second type always gives a zero count. The actual counts generated from the first type (i.e. Poisson) will have a probability $1-\pi$, while the probability of an individual being of the second type will be $\pi$. Therefore, $\pi$ is called the "zero-inflation probability". The probability ($P$) of observing $y = 0, 1, 2, \ldots, n$ in the ZIP distribution is (Cunningham and Lindenmayer, 2005; Martin et al., 2005):

$$P(Y_i = y_i) = \begin{cases} \pi + (1-\pi)\exp(-\mu) & \text{for} \quad y_i = 0 \\ (1-\pi)\dfrac{\exp(-\mu)\,\mu^y}{y!} & \text{for} \quad y_i = 1, 2, \ldots, n \end{cases} \tag{1}$$

with mean $E(Y_i = \mu_i)$. The parameter of interest in the ZIP occupancy–abundance model (Section 2.2.3) is $\pi$, which was estimated here as a function of covariates in each dataset. Like the ZIP, ZINB uses a mixture distribution that assigns a mass $\pi$ to the extra zeros and a mass $1-\pi$ to the NBD (Martin et al., 2005; Mwalili et al., 2008). Hence the probability of observing $y = 0, 1, 2, \ldots, n$ is

$$P(Y_i = y_i) = \begin{cases} \pi + (1-\pi)\dfrac{1}{1+\tau\mu} & \text{for} \quad y_i = 0 \\ (1-\pi)\dfrac{\Gamma\left(y + 1/\tau\right)}{y!\,\Gamma\left(1/\tau\right)}\left(\dfrac{(\tau\mu)^y}{(1+\tau\mu)^y(1+\tau\mu)^{1/\tau}}\right) & \text{for} \quad y_i = 1, 2, \ldots, n \end{cases} \tag{2}$$

where $\tau = 1/k$, and $k$ is the dispersion parameter of the NBD. ZINB approaches the ZIP and NBD as $\tau \to \infty$ and $\pi \to 0$, respectively (Mwalili et al., 2008). If both $1/\tau$ and $\pi \approx 0$ then the ZINB reduces to the Poisson distribution (Mwalili et al., 2008). The parameters of interest in the ZINB occupancy–abundance model (Section 2.2.3) are $\pi$ and $k$, which were jointly estimated as functions of the covariates in each dataset.

All data were subjected to the Poisson, NBD, ZIP and ZINB regression models. However, detailed analyses will be illustrated using the *M. ochroptera* dataset as it represents a typical example of repeated measures data with many zeros. The analysis of repeated measures data involved inclusion of fixed effects (covariates) and a single normally distributed random effect ($u$) with $\mu = 0$ and variance $\delta^2$ (Agresti et al., 2000; Hall, 2000; Tooze et al., 2002). Random effects are associated with individual experimental units (e.g. trees in the case of *M. ochroptera*) drawn at random from a population and govern the variance–covariance structure of the response variable. Parameters of the Poisson, NBD, ZIP and ZINB models were all estimated using the non-linear mixed effects model (NLMIXED) procedure of SAS (Agresti et al., 2000). This procedure maximizes the likelihood by adaptive Gaussian quadrature, which is one of the best methods that deliver exact maximum likelihood estimation (Pinheiro and Bates, 1995) of parameters such as $\pi$ and $k$. Models were fitted sequentially starting from a null (without covariates) through single-, two- and three-variable main effects models to the full model (all variables) (see Appendix Table A.1 for details).

For a given functional form of the count model, the deviance explained (%$D$) was used as a measure of goodness-of-fit. The deviance values of each model were compared with that of the null model to examine what proportion of the variation in the response was attributed to the covariates. Interpretation of results was based mainly on the 95% confidence intervals (95% CI) of parameters. If the 95% CI includes both negative and positive values, parameters were interpreted as not significantly different from zero. If the 95% CI of two or more distributions (e.g. ZIP, NBD and ZINB) overlap, the effect of distributional assumption on the parameters (e.g. $\pi$ or $k$) was interpreted as non-significant. Similarly, if the 95% CI of the null and full models overlap, the effect of covariates on parameters was interpreted as non-significant.

In the approach described here, model selection is an important step of parameter estimation because only those parameters estimated using a parsimonious model can produce accurate occupancy–abundance relationships. Here, parsimony is defined as a trade-off between bias and variance. A model with too few parameters results in high bias in parameter estimates and an under-fit model that fails to identify factors of importance. Too many parameters result in high variance, and an over-fit model with a risk of identifying spurious factors (Johnson and Omland, 2004). The traditional approach to model selection is to use a likelihood ratio test to compare nested models. This is appropriate for comparing the Poisson with the NBD or the ZIP with the ZINB as the first is nested in the second. For non-nested models we cannot use the standard likelihood ratio test. An appropriate approach in such cases is to use Akaike information criterion (AIC) (Gray, 2005; Sileshi, 2006, 2008). One could also use other criteria such as the Bayesian information criterion (BIC) to compare models. However, note that AIC and BIC arise from a more general form of information criterion (IC): $\text{IC} = -2\text{ll} + \theta c$. If $c = 0$, IC is equal to the classical likelihood-ratio statistic. If $c = 1$, IC is equal to the goodness-of-fit procedure based on plotting the deviance against degrees of freedom. If $c = 2$, IC is identical to AIC, and if $c = \log N$, IC is equal to BIC. AIC ($c = 2$) is also asymptotically equivalent to a cross-validation criterion (Sileshi, 2008 and references cited). Therefore, model selection was based on AIC computed as $\text{AIC} = -2\text{ll} + 2\theta$ from the log-likelihood (ll) and number of parameters ($\theta$) estimated.

Since AIC does not take into account sample size, AICc, a small sample bias adjustment of AIC was used. In order to compare different models the AICc difference ($\Delta$AICc) was calculated from AICc as in Johnson and Omland (2004). This in turn allowed calculation of the relative likelihoods of the different models. Normalizing the relative likelihood yielded the Akaike weight (AICw), which was used as the strength of evidence for each model relative to other models in the set of models considered. The AICw of any particular model varies from zero (no support) to unity (complete support) relative to the entire set of models (Johnson and Omland, 2004).

### *2.2.2. Modelling the abundance–variance relationship*

Abundance–variance–occupancy relationships are established using parameters of the variance–abundance relationship described by TPL:

$$\sigma^2 = \alpha\mu^\beta \tag{3}$$

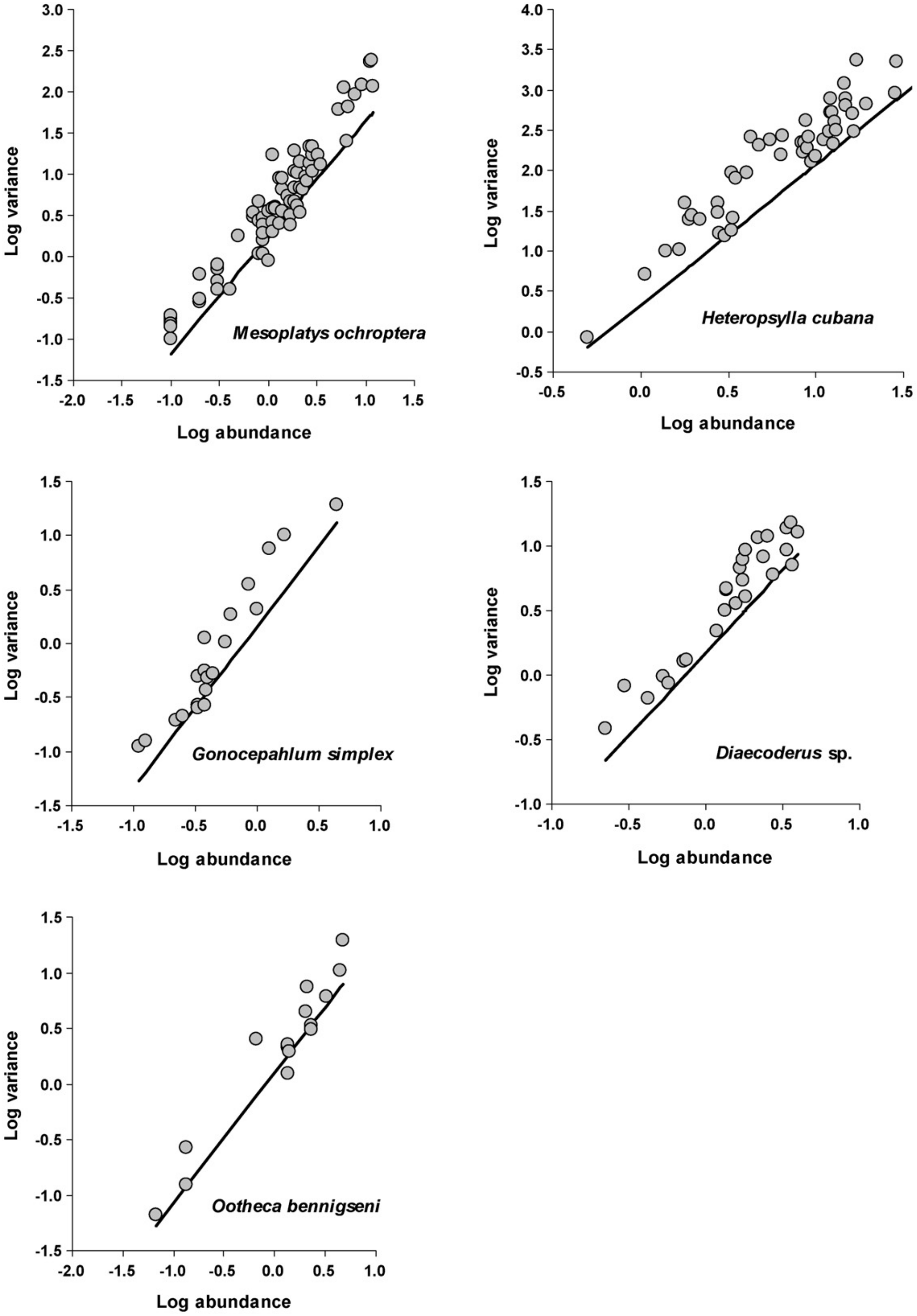


**Fig. 1.** Plots (on logarithmic scale) of the observed variances (open circles) and variances predicted using Taylor's power law (straight line) against the mean densities of various insects.

where $\alpha$ and $\beta$ are estimated by fitting a least square regression of the sample variance ($s^2$) against the sample mean density ($m$) on a logarithmic scale (Taylor, 1961). In practice, replicate count data are obtained from a variety of conditions, times, etc., the sample variance ($s^2$), mean ($m$) and occupancy ($p_o$) are computed. The assumption is that these simple values provide an estimate of the population variance ($\sigma^2$), abundance ($\mu$) and occupancy ($p$). In order to illustrate the effect of covariates on these parameters, $s^2$ and $m$ were computed according to the different covariate models in the datasets. However, only $\alpha$ and $\beta$ values derived according to the best abundance model (largest AICw) were used in generating the abundance–variance–occupancy relationships in Section 2.2.3. To that effect, the observed $s^2$, $m$ and $p_o$ values for *M. ochroptera* were computed per 15 trees, which gave a total of 91 values of $s^2$, $m$ and $p_o$ after discarding zero $s^2$ and $m$ values. For *H. cubana*, $s^2$, $m$ and $p_o$ were computed per 18 shoots (i.e. 6 dates × 3 trees per provenance), making a total of 48 values of each. In the *Diaecoderus* dataset, computation was based on 40 plants (i.e. 10 plants each from 4 replicates per treatment), and this gave a total of 26 $s^2$, $m$ and $p_o$ values after discarding zero $s^2$ and $m$ values. In the *G. simplex* dataset, computation was based on nine soil monoliths (i.e. 3 each from 3 replicates per treatment). A total of 42 $s^2$, $m$ and $p_o$ values were available after dropping those with zero values. In the *O. bennigsenni* dataset, $s^2$, $m$ and $p_o$ were computed per 15 bean or 30 cowpea plants. This gave a total of 16 $s^2$, $m$ and $p_o$ values after discarding those with zero values. Then the observed variances (=$s^2$) and the variances predicted ($\sigma^2$) using Eq. (3) were plotted against the respective $m$ value (on a log scale) to demonstrate the agreement between $s^2$ and $\sigma^2$ (Fig. 1). Since $\log(s^2) = 0$ and $\log(m) = 0$ are undefined, samples for which $m = s^2 = 0$ or $s^2 = 0$ but $m > 0$ were removed from all datasets when computing $\alpha$ and $\beta$.

### 2.2.3. *Modelling occupancy–abundance and abundance–variance–occupancy relationships*

These relationships were modelled by inserting the relevant parameters estimated in Section 2.2.1, and the mean density ($m$) value into the following models. Note that $\mu$ was approximated by $m$ in all models. The predicted occupancy ($p_p$) values of the Poisson occupancy–abundance model were generated using the following relationship:

$$p_p = 1 - \exp^{-\mu} \tag{4}$$

Similarly, $p_p$ values of the NB model were generated by inserting the dispersion parameter ($k$) from the best NBD model into the following equation:

$$p_p = 1 - \left(1 + \frac{\mu}{k}\right)^{-k} \tag{5}$$

If $k$ is infinitely large, Eq. (5) results in the Poisson model (Eq. (4)).

The ZIP occupancy–abundance model was derived from equation 1 as follows:

$$p_p = (1 - \pi)\left(1 - \exp^{-\mu}\right) \tag{6}$$

Then $p_p$ values were generated by inserting $\pi$ values from the best ZIP abundance model (Section 2.2.1) into Eq. (6). As $\pi \to 0$, the occupancy predicted by Eq. (6) approaches that of the Poisson (Eq. (4)).

The ZINB occupancy–abundance model was derived from Eq. (2) as follows:

$$p_p = (1 - \pi)\left[1 - \left(1 + \frac{\mu}{k}\right)^{-k}\right] \tag{7}$$

**Table 1**
Ranking of Poisson and zero-inflated Poisson (ZIP) models of *M. ochroptera* abundance using the explained deviance (%*D*) and Akaike information criteria (ACc and AICw). Note the variation in the probability of zero-inflation ($\pi$) and their 95% confidence intervals (95% CI) with distributional assumption and covariate structure.

| | Model | $\theta$[a] | %*D* | AICc | AICw | π (95% CI) |
|---|---|---|---|---|---|---|
| Poisson | Null | 2 | 0.0 | 5289 | <0.01 | NA |
| | Site | 3 | 0.2 | 5282 | <0.01 | NA |
| | Treatment | 3 | 0.2 | 5279 | <0.01 | NA |
| | Date | 3 | 0.3 | 5277 | <0.01 | NA |
| | Date + site | 4 | 0.4 | 5273 | <0.01 | NA |
| | Site + treatment | 4 | 0.4 | 5271 | <0.01 | NA |
| | Date + treatment | 4 | 0.5 | 5266 | <0.01 | NA |
| | Date + site + treatment | 5 | 0.6 | 5261 | <0.01 | NA |
| | Year + date + site | 5 | 2.4 | 5167 | <0.01 | NA |
| | Year + date | 4 | 2.4 | 5166 | <0.01 | NA |
| | Year + site | 4 | 2.4 | 5165 | <0.01 | NA |
| | Year | 3 | 2.4 | 5164 | <0.01 | NA |
| | Year + date + site + treatment | 6 | 2.7 | 5155 | 0.10 | NA |
| | Year + date + treatment | 5 | 2.7 | 5154 | 0.17 | NA |
| | Year + site + treatment | 5 | 2.7 | 5153 | 0.28 | NA |
| | Year + treatment | **4** | **2.7** | **5152** | **0.45** | NA |
| ZIP | Null | 3 | 0.0 | 5104 | <0.01 | 0.71 (0.67–0.74) |
| | Treatment | 5 | 0.2 | 5097 | <0.01 | 0.66 (0.61–0.71) |
| | Date | 5 | 0.3 | 5093 | <0.01 | 0.73 (0.68–0.79) |
| | Date + treatment | 7 | 0.5 | 5085 | <0.01 | 0.69 (0.63–0.76) |
| | Site | 5 | 0.8 | 5067 | <0.01 | 0.73 (0.69–0.77) |
| | Site + treatment | 7 | 1.0 | 5060 | <0.01 | 0.68 (0.63–0.74) |
| | Year | 5 | 2.2 | 4999 | <0.01 | 0.62 (0.57–0.66) |
| | Year + date | 7 | 2.3 | 4997 | <0.01 | 0.59 (0.59–0.67) |
| | Year + treatment | 7 | 2.4 | 4990 | <0.01 | 0.56 (0.50–0.62) |
| | Year + date + treatment | 9 | 2.5 | 4989 | <0.01 | 0.54 (0.45–0.63) |
| | Date + site | 7 | 2.4 | 4988 | <0.01 | 0.80 (0.73–0.87) |
| | Date + site + treatment | 9 | 2.8 | 4974 | <0.01 | 0.70 (0.59–0.82) |
| | Year + site | 7 | 2.8 | 4968 | <0.01 | 0.60 (0.54–0.66) |
| | Year + site + treatment | 9 | 3.1 | 4960 | <0.01 | 0.54 (0.47–0.61) |
| | Year + date + site | 9 | 5.6 | 4833 | <0.01 | 0.63 (0.50–0.76) |
| | Year + date + site + treatment | **11** | **5.8** | **4817** | **1.00** | **0.49 (0.33–0.66)** |

The model in bold face is the best estimator of the parameters. NA = not applicable.

[a] $\theta$ is the total number of parameters estimated in each model including $\alpha_i$, $\beta_i$, and $\pi$.

Then the $p_p$ values were generated by inserting $\pi$ and $k$ values from the best ZINB abundance model into Eq. (7). Note that as $\pi \to 0$, the occupancy predicted by Eq. (7) approaches that of the NBD (Eq. (5)).

The $p_p$ values of He and Gaston's (2003) model were generated by inserting $\sigma^2$ values estimated using Eq. (3) into the following equation:

$$p_p = 1 - \left(\frac{\mu}{\sigma^2}\right)^{\mu^2/\sigma^2-\mu} \tag{8}$$

The $p_p$ values of Wilson and Room's (1983) model were generated by inserting $\alpha$ and $\beta$ values estimated using Eq (3) into the following equation:

$$p_p = 1 - \exp^{-\mu(\ln(a\mu^{b-1})(a\mu^{b-1}-1)^{-1})} \tag{9}$$

The agreement between the fitted ($p_p$) and observed ($p_o$) occupancy across the whole range of mean densities was graphically evaluated by plotting $p_p$ and $p_o$ against the respective $m$ values. The agreement between $p_o$ and $p_p$ predicted by the different models was also statistically tested using a linear regression. Here, $p_o$ values were used as the dependent and $p_p$ values as the independent ($x$) variables (Piñeiro et al., 2008). If the model under consideration predicts occupancy consistently across all mean densities, the resulting regression line will have a constant slope ($\beta = 1$). If it predicts occupancy without bias the regression line will have zero intercept ($\alpha = 0$). Deviation from this conditions was judged by examination of the 95% CI of $\beta$ and $\alpha$. Model fit was also evaluated using the $R^2$ and predicted residual sum of squares (PRESS). The smaller the PRESS value, the better is the agreement between the observed and predicted values. When several models are compared, model likelihood is more informative than model fit. Therefore, the Akaike weight (AICw) was used to evaluate the relative support for each model. For a model with a given number of predictors based on a normal regression, the log-likelihood is $-0.5n(\ln(2\theta)+(\text{RSS}/n)+1)$. Following Gagné and Dayton (2002), AIC was calculated from the residual sum of squares (RSS), the number of parameters ($\theta$) estimated and the sample size ($n$):

$$\text{AIC} = n\left(\ln\left(\frac{\text{RSS}}{n}\right)\right) + 2(\theta) \tag{10}$$

In all cases the slope and intercept were the only parameters estimated (hence $\theta = 2$), and $n$ was equal to the number of pairs of occupancy ($p_p$ and $p_o$) values. For ach model, AICw was computed as described in Section 2.2.1.

## 3. Results

### 3.1. *Effect of distributional assumptions and model choice on parameter estimates*

The count distribution model used and covariates in that model had significant effects on the explained deviance. According to the AIC, the ZIP and ZINB abundance models were superior to the standard Poisson and NBD models (Tables 1 and 2). However, the covariates included in each model explained only a small portion of the explained devaince (%$D$ < 10) (Tables 1 and 2). Among the range of *M. ochroptera* abundance models considered under the Poisson, NBD, ZIP and ZINB distributions, those ranked first accounted for only 2.7, 2.0, 5.8 and 6.6% of the explained deviance, respectively (Tables 1 and 2). Models of *Diaecoderus* abundance accounted for 3–12.3% of the explained deviance. In the case of *H. cubana* and *G.*

**Table 2**
Ranking of negative binomial distribution (NBD) and zero-inflated negative binomial (ZINB) models of *M. ochroptera* abundance using the explained deviance (%$D$) and Akaike information criteria (ACc and AICw). Note the variation in the probability of zero-inflation ($\pi$) and dispersion parameter ($k$) estimates and their 95% CI with distributional assumption and covariate structure.

| | Nested model | $\theta$[a] | %$D$ | AICc | AICw | $\pi$ (95% CI) | $k$ (95% CI) |
|---|---|---|---|---|---|---|---|
| NBD | Null | 3 | 0.0 | 5138 | <0.01 | NA | 10.8 (9.3–12.3) |
| | Treatment | 4 | 0.0 | 5138 | <0.01 | NA | 10.6 (9.1–12.1) |
| | Date | 4 | 0.5 | 5116 | <0.01 | NA | 10.9 (9.4–12.3) |
| | Date + treatment | 5 | 0.6 | 5113 | <0.01 | NA | 10.7 (9.3–12.2) |
| | Site | 4 | 0.6 | 5109 | <0.01 | NA | 10.9 (9.4–12.3) |
| | Site + treatment | 5 | 0.7 | 5108 | <0.01 | NA | 10.7 (9.3–12.1) |
| | Date + site | 5 | 0.7 | 5105 | <0.01 | NA | 10.8 (9.4–12.2) |
| | Date + site + treatment | 6 | 0.8 | 5102 | <0.01 | NA | 10.7 (9.2–12.1) |
| | Year + date | 5 | 1.9 | 5046 | 0.01 | NA | 9.3 (7.9–10.6) |
| | Year | 4 | 1.8 | 5046 | 0.01 | NA | 9.2 (7.9–10.6) |
| | Year + date + site | 6 | 2.0 | 5045 | 0.02 | NA | 9.3 (7.9–10.6) |
| | Year + site | 5 | 1.7 | 5044 | 0.04 | NA | 9.3 (7.9–10.6) |
| | Year + date + site + treatment | 7 | 2.0 | 5043 | 0.06 | NA | 9.9 (8.8–11.1) |
| | Year + treatment | 5 | 1.9 | 5042 | 0.10 | NA | 9.1 (7.7–10.5) |
| | Year + date + treatment | 6 | 2.0 | 5040 | 0.28 | NA | 9.1 (7.8–10.5) |
| | **Year** + site + **treatment** | **6** | **2.0** | **5039** | **0.46** | **NA** | **9.1 (7.7–10.5)** |
| ZINB | Null | 4 | 0.0 | 5107 | <0.01 | 0.73 (0.70–0.76) | 0.3 (−0.3–0.8) |
| | Treatment | 6 | 0.2 | 5100 | <0.01 | 0.69 (0.64–0.73) | 0.3 (−0.3–0.8) |
| | Date | 6 | 0.4 | 5093 | <0.01 | 0.74 (0.68–0.79) | 0.3 (−0.3–0.9) |
| | Date + treatment | 8 | 0.6 | 5086 | <0.01 | 0.70 (0.63–0.76) | 0.3 (−0.3–0.8) |
| | Site | 6 | 0.8 | 5068 | <0.01 | 0.73 (0.69–0.77) | 0.2 (−0.2–0.6) |
| | Site + treatment | 8 | 1.1 | 5061 | <0.01 | 0.68 (0.63–0.74) | 0.2 (−0.2–0.6) |
| | Year | 6 | 2.2 | 5000 | <0.01 | 0.65 (0.61–0.69) | 0.3 (−0.2–0.8) |
| | Year + date | 8 | 2.3 | 4998 | <0.01 | 0.60 (0.51–0.69) | 0.3 (−0.4–0.9) |
| | Year + treatment | 8 | 2.4 | 4992 | <0.01 | 0.60 (0.54–0.65) | 0.3 (−0.2–0.8) |
| | Year + date + treatment | 10 | 2.6 | 4989 | <0.01 | 0.55 (0.45–0.64) | 0.3 (−0.4–0.9) |
| | Year + site | 8 | 2.9 | 4969 | <0.01 | 0.60 (0.54–0.66) | 0.2 (−0.3–0.7) |
| | Year + site + treatment | 10 | 3.1 | 4961 | <0.01 | 0.54 (0.47–0.61) | 0.2 (−0.3–0.9) |
| | Date + site | 8 | 3.8 | 4919 | <0.01 | 0.70 (0.57–0.82) | 6.3 (3.3–5.8) |
| | Date + site + treatment | 10 | 4.1 | 4909 | <0.01 | 0.52 (0.33–0.71) | 6.1 (4.8–7.3) |
| | Year + date + site | 10 | 6.3 | 4797 | <0.01 | 0.50 (0.33–0.68) | 4.6 (3.3–5.8) |
| | **Year + Date + site + treatment** | **12** | **6.6** | **4786** | **1.00** | **0.33 (0.12–0.53)** | **4.3 (3.1–5.6)** |

The model in bold face is the best estimator of the parameters. NA = not applicable.

[a] $\theta$ is the number of parameters estimated in each model including $\alpha_i$, $\beta_i$, $\pi$ and $k$.

**Table 3**
The effect of distributional assumptions and covariates on the zero-inflation probability ($\pi$) and the dispersion parameter ($k$) in counts of *Diaecoderus* sp., *Heteropsylla cubana*, *Gonocephalum simplex* and *Ootheca bennigseni*. The null model has no covariates (intercept only), while the best model is the one with the highest likelihood according to the AICw.

| Species | Distribution | Model | $\pi$ (95% CI) | $k$ (95% CI) |
|---|---|---|---|---|
| *Diaecoderus* sp. | NBD | Null | NA | 1.7 (1.5–2.0) |
| | | Best | NA | 1.4 (1.2–1.6) |
| | ZIP | Null | 0.42 (0.39–0.45) | NA |
| | | Best | 0.34 (0.27–0.41) | NA |
| | ZINB | Null | 0.05 (−0.16–0.26) | 1.6(0.8–2.3) |
| | | Best | 0.00001 (−0.001–0.001) | 1.0 (0.8–1.2) |
| *H. cubana* | NBD | Null | NA | 5.1 (4.6–5.7) |
| | | Best | NA | 4.9 (4.4–5.5) |
| | ZIP | Null | 0.47 (0.44–0.50) | NA |
| | | Best | 0.45 (0.37–0.53) | NA |
| | ZINB | Null | 0.07 (−0.17–0.31) | 4.5 (2.5–6.6) |
| | | Best | 0.004 (−0.01–0.02) | 4.0 (3.3–4.7) |
| *G. simplex* | NBD | Null | NA | 1.7 (1.5–2.0) |
| | | Best | NA | 1.4 (1.2–1.6) |
| | ZIP | Null | 0.71 (0.66–0.75) | NA |
| | | Best | 0.50 (0.32–0.68) | NA |
| | ZINB | Null | 0.00004 (−0.001–0.001) | 3.0 (0.7–5.2) |
| | | Best | 0.05 (−0.17–0.26) | 4.3 (3.0–5.5) |
| *O. bennigseni* | NBD | Null | NA | 4.2 (3.1–5.3) |
| | | Best | NA | 1.7 (1.2–2.2) |
| | ZIP | Null | 0.67 (0.62–0.72) | NA |
| | | Best | 0.002 (−0.0001–0.006) | NA |
| | ZINB | Null | 0.56 (0.45–0.68) | 0.8 (0.2–1.5) |
| | | Best | 0.0002 (−0.0001–0.0001) | 0.7 (0.3–1.0) |

NA = parameter not applicable.

*simplex* less than 6% of the deviance was explained, while 16–28% was explained in the case of *O. bennigseni*.

The zero-inflation probability ($\pi$) in the data significantly varied with the distributional assumption and covariates included in each model (Tables 1–3). The 95% CI showed differences in estimates of $\pi$ among various models of *M. ochroptera* within ZIP (Table 1) and ZINB distributions (Table 2). Zero inflation was significantly lower in the best ZIP model compared with the null model (Table 1). Similarly, the zero-inflation probability of was significantly reduced (by 54.8%) in the best ZINB model compared with the null model. Most of the single- and two-variable models of *M. ochroptera* abundance had significantly higher zero-inflation probability ($\pi > 0.60$) compared to the full ZINB model ($\pi < 0.54$). In the *Diaecoderus* dataset, zero-inflation probability dropped from 0.42 in the null ZIP model to 0.34 in the best ZIP model. In the ZINB model, the reduction was even more dramatic (Table 3). Similar trends were observed in the *H. cubana*, *G. simplex* and *O. bennigseni* datasets, where the null model had significantly higher zero-inflation probability than the best model. In all cases, the zero probability dropped from over 0.45 in the ZIP models to less than 0.10 in the ZINB models with the same covariates (Table 3).

The dispersion parameter ($k$) of the NBD varied with the count distribution used and covariate in the model (Tables 1–3). The 95% CI showed that $k$ values were significantly larger in the standard NBD models than in the ZINB models (Table 2). Confidence intervals of $k$ were narrower in the best model compared with the null model of all the datasets (Tables 2 and 3).

The regression parameters of TPL and their 95% CI significantly changed with the manner in which variances and means were calculated (Table 4). For example, $\alpha$ values estimated using $s^2$ and $m$

**Table 4**
Effect of model selection (covariate structure) on Taylor's power law parameters of counts of *Mesoplatys ochroptera*, *Diaecoderus* sp., *Heteropsylla cubana*, *Gonocephalum simplex* and *Ootheca bennigseni*.

| Species | Model | $\alpha$ (95% CI) | $\beta$ (95% CI) | $R^2$ |
|---|---|---|---|---|
| *M. ochroptera* | Best | 0.53 (0.48–0.58) | 1.46 (1.39–1.54) | 0.950 |
| | Date + site + treatment | 0.59 (0.54–0.64) | 1.49 (1.42–1.57) | 0.961 |
| | Year + date + site | 0.70 (0.61–0.79) | 1.43 (1.31–1.55) | 0.949 |
| | Date + site | 0.76 (0.66–0.86) | 1.48 (1.35–1.62) | 0.955 |
| | Year + site + treatment | 0.84 (0.72–0.95) | 1.54 (1.32–1.76) | 0.922 |
| | Year + date + treatment | 0.85 (0.76–0.94) | 1.52 (1.39–1.66) | 0.944 |
| | Year + date | 0.99 (0.82–1.17) | 1.43 (1.22–1.64) | 0.957 |
| *Diaecoderus* sp. | Best | 0.40 (0.33–0.46) | 1.29 (1.12–1.46) | 0.904 |
| | Treatment | 0.43 (0.36–0.49) | 1.29 (1.16–1.62) | 0.942 |
| *H. cubana* | Best (all covariates) | 0.75 (0.59–0.92) | 1.75 (1.57–1.93) | 0.892 |
| | Date | 1.15 (0.01–2.28) | 1.55 (0.34–2.76) | 0.760 |
| | Provenance | 0.01 (−0.98–0.96) | 2.73 (1.73–2.74) | 0.881 |
| *G. simplex* | Best | 0.36 (0.26–0.46) | 1.49 (1.34–1.64) | 0.909 |
| | Month, treatment | 0.40 (0.29–0.51) | 1.52 (1.35–1.69) | 0.927 |
| | Site, treatment | 0.42 (0.28–0.57) | 1.54 (1.34–1.75) | 0.902 |
| *O. bennigseni* | Best | 0.25 (0.15–0.35) | 1.18 (1.01–1.34) | 0.943 |
| | Farm | 0.25 (0.15–0.35) | 1.18 (1.01–1.34) | 0.943 |
| | Treatment | 0.40 (0.27–0.52) | 1.39 (1.09–1.69) | 0.894 |

calculated according to the best *M. ochroptera* abundance model were significantly smaller than those based on year and date of sampling (Table 4). Although not significantly different, $\alpha$ values based on the best model were generally small than those based on single-variable models of the other insects (Table 4). Although the point estimates of the slope ($\beta$) did not vary much, their 95% CI were narrower when means and variances were calculated according to the best model than single variable models (Table 4).

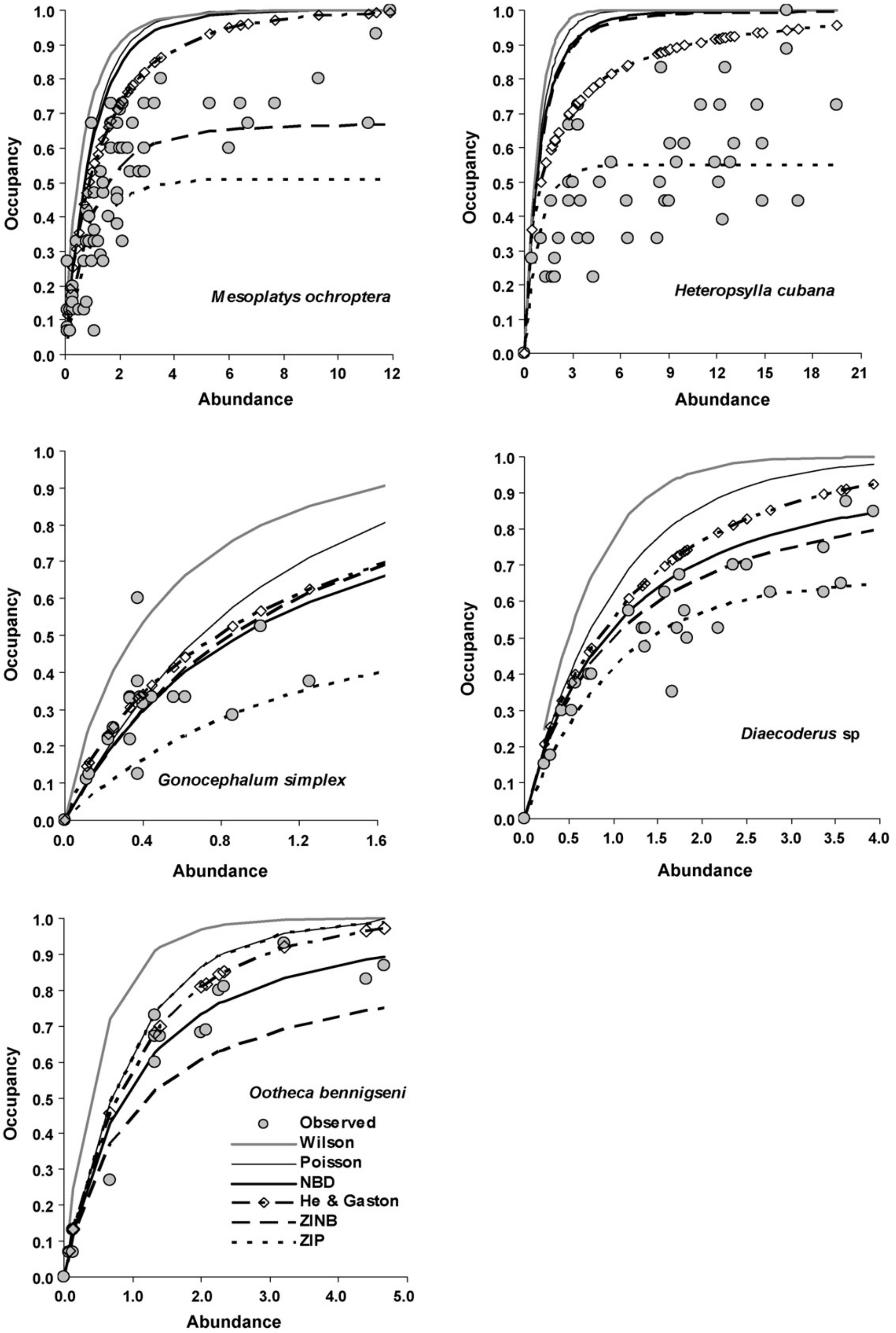


**Fig. 2.** Plots of the observed occupancy (open circles) and occupancy predicted by the standard Poisson, ZIP, NBD, ZINB, Wilson and Room, and He and Gaston's models against mean densities of various insects.

Fig. 1 shows the agreement between the observed variances and those predicted from TPL across a range of mean densities on a logarithmic scale. The predicted variances were consistently smaller than the observed across the range of densities. In all cases, regression of the observed variance against the predicted resulted in a line with a slope ($\beta$) close to 1 indicating that TPL predicted the variance consistently across all mean densities. However, the 95% CI of $\alpha$ (0.21–0.30 for *M. ochroptera*, 0.23–0.62 for *H. cubana*, 0.11–0.29 for *G. simplex*, 0.15–0.30 for *Diaecodersus* sp., and 0.04–0.24 for *O. bennigseni*) were significantly greater than 0 indicating bias in the prediction of the variance by TPL.

### 3.2. Relative performance of occupancy–abundance models

The occupancy predicted by all models closely agreed with the observed occupancy at low densities, while the agreement generally decreased as density increased (Fig. 2). The agreement was also poorer in species whose mean densities varied widely such as *M. ochroptera* and *H. cubana* compared with those that had narrower density ranges (Fig. 2). Comparisons of the various models with respect to model fit and likelihood are presented in Table 5. The intercepts of the regression of observed vs predicted occupancy did not significantly differ from zero (i.e. 95% CI included zero), indicating lack of bias in most models of *M. ochroptera* (Table 5). The only exception was Wilson and Rooms' model that was significantly biased in predicting occupancy. The regression slopes were significantly smaller than unity indicating overestimation of *M. ochroptera* occupancy by the Poisson, Wilson and Room's, He and Gaston's and the NBD occupancy–abundance models. On other hand, the slope of the ZIP model was significantly greater than unity ($\beta > 1.0$) indicating significant underestimation of the predicted occupancy of *M. ochroptera* relative to the observed (Table 5). The Poisson and Wilson and Room's models overestimated occupancy, and provided the worst predictions for all datasets (Fig. 2; Table 5).

According to the $R^2$ and PRESS, ZINB was the first ranked occupancy–abundance model of *M. ochroptera* and *Diaecoderus* sp., where as He and Gaston's model was the first ranked for *H. cubana* and *G. simplex*. However, among the candidate models examined, these best-fitting models were not overwhelmingly supported by the data. For example, given the *M. ochroptera* data, the model ranked first had only 51% (AICw = 0.511) chance of being the best among the set of candidate models. Similarly, for *Diaecoderus* sp., *H. cubana* and *G. simplex*, the model ranked first had 32, 66 and 55% chance of being the best, respectively. For *O. bennigseni* and *G. simplex* where $\pi \approx 0$ (Table 3) the Poisson model was as good as the ZIP model (Fig. 2, Table 5). The first, second and third ranked models had only 29, 25 and 22% chance of being the best of *O. bennigseni* occupancy–abundance models (Table 5).

## 4. Discussion

The percentage of zero counts ranged from 45 to 80% in the datasets analysed. This is typical of many ecological counts (Martin et al., 2005; Sileshi, 2006, 2008; Warton, 2005). In addition, spatially and temporally correlated data are often collected in long-term ecological studies such as those of *M. ochroptera*. In such studies, observations have some type of clustering, with observations within clusters tending to be correlated. Such data often tend to show clumping at zero (Hall, 2000; Tooze et al., 2002). As illustrated here (Tables 1–3) and elsewhere (Gray, 2005; Hall, 2000; Sileshi, 2008) the excess zeros can be accommodated by zero-inflated models and covariates. Using the modelling framework proposed here, one can also explain the source of zeros. For example, the dramatic reduction in the $\pi$ values from the ZIP to

**Table 5**
Comparison of occupancy–abundance models of *Mesoplatys ochroptera*, *Diaecoderus* sp., *Heteropsylla cubana*, *Gonocephalum simplex* and *Ootheca bennigseni* using regression slopes, coefficient of determination ($R^2$), predicted residual sum of squares (PRESS) and Akaike weights (AICw).

| Species | Model | Intercept[a] | Slope[a] | $R^2$ | PRESS | AICw |
|---|---|---|---|---|---|---|
| *M. ochroptera* | Wilson and Room | −0.09 (−0.15, −0.03) | 0.72 (0.64, 0.80) | 0.773 | 1.36 | 0.000 |
| | Poisson | −0.01 (−0.06, 0.04) | 0.67 (0.60, 0.74) | 0.806 | 1.16 | 0.012 |
| | ZIP | −0.01 (−0.06, 0.04) | 1.32 (1.19, 1.45) | 0.806 | 1.16 | 0.017 |
| | NBD | −0.01 (−0.05, 0.03) | 0.70 (0.63, 0.77) | 0.812 | 1.13 | 0.050 |
| | He and Gaston | −0.03 (−0.07, 0. 01) | 0.80 (0.73, 0.87) | 0.818 | 1.09 | 0.410 |
| | **ZINB** | −0.01 (−0.05, 0.03) | **1.06 (0.96, 1.16)** | **0.826** | **1.02** | **0.511** |
| *Diecoderus* sp. | Wilson and Room | −0.05 (−0.20, 0.10) | 0.70 (0.53, 0.87) | 0.733 | 0.26 | 0.002 |
| | Poisson | 0.02 (−0.09, 0.13) | 0.70 (0.56, 0.84) | 0.795 | 0.20 | 0.067 |
| | ZIP | 0.02 (−0.09, 0.13) | 1.06 (0.84, 1.28) | 0.795 | 0.20 | 0.067 |
| | He and Gaston | 0.01 (−0.09, 0.11) | 0.78 (0.63, 0.93) | 0.816 | 0.18 | 0.268 |
| | NBD | −0.01 (−0.12, 0.10) | 0.87 (0.71, 1.03) | 0.816 | 0.18 | 0.274 |
| | **ZINB** | −0.02 (−0.13, 0.09) | **0.94 (0.76, 1.12)** | **0.818** | **0.18** | **0.321** |
| *H. cubana* | Wilson and Room | −0.26 (−0.78, 0.26) | 0.82 (0.29, 1.35) | 0.164 | 2.01 | 0.001 |
| | Poisson | −0.26 (−0.70, 0.18) | 0.84 (0.38, 1.30) | 0.213 | 1.75 | 0.006 |
| | ZIP | −0.27 (−0.71, 0.17) | 1.52 (0.68, 2.36) | 0.213 | 1.75 | 0.006 |
| | NBD | −0.26 (−0.65, 0.13) | 0.85 (0.44, 1.26) | 0.263 | 1.59 | 0.027 |
| | ZINB | −0.26 (−0.64, 0.12) | 0.86 (0.45, 1.27) | 0.273 | 1.56 | 0.302 |
| | **He and Gaston** | −0.17 (−0.43, 0.09) | **0.86 (0.55, 1.17)** | **0.388** | **1.25** | **0.659** |
| *G. simplex* | ZIP | 0.08 (0.04, 0.12) | 1.24 (0.99, 1.49) | 0.712 | 0.31 | 0.044 |
| | Poisson | 0.09 (0.05, 0.13) | 0.62 (0.50, 0.74) | 0.713 | 0.31 | 0.044 |
| | Wilson and Room | −0.02 (−0.08, 0.04) | 0.64 (0.52, 0.76) | 0.735 | 0.29 | 0.060 |
| | ZINB | 0.08 (0.04, 0.12) | 0.72 (0.58, 0.86) | 0.726 | 0.28 | 0.126 |
| | NBD | 0.07 (0.03, 0.11) | 0.76 (0.62, 0.90) | 0.725 | 0.28 | 0.177 |
| | **He and Gaston** | 0.03 (−0.02, 0.08) | 0.77 (0.60, 0.88) | **0.746** | **0.26** | **0.550** |
| *O. bennigseni* | Wilson and Room | −0.11 (−0.23, 0.01) | 0.88 (0.73, 1.03) | 0.909 | 0.17 | 0.000 |
| | ZINB | −0.03 (−0.11, 0.05) | 1.26 (1.11, 1.41) | 0.954 | 0.08 | 0.101 |
| | NBD | −0.02 (−0.09, 0.05) | **1.03 (0.92, 1.14)** | 0.958 | 0.08 | 0.130 |
| | He and Gaston | −0.02 (−0.09, 0.05) | 0.94 (0.84, 1.04) | 0.961 | 0.07 | 0.221 |
| | Poisson | −0.01 (−0.08, 0.06) | 0.89 (0.80, 0.98) | 0.962 | 0.07 | 0.254 |
| | **ZIP** | −0.01 (0.08, 0.06) | 0.89 (0.80, 0.98) | **0.962** | **0.06** | **0.293** |

Figures in bold face indicate the best model for each species.

[a] Figures in parentheses are the lower and upper 95% confidence limits of the intercepts and slopes.

ZINB (with the same covariates) in the case of *H. cubana*, *Diaecoderus* sp. and *G. simplex* clearly indicates that most of the zeros were due to contagion that could not be accounted for by the ZIP models. Similarly, the significant reduction in $\pi$ values in the best ZINB model compared to the null ZINB model in all datasets indicates that some zeros are related to environmental correlates of abundance. This indicates that the parameters (e.g. $\pi$, $k$) used in occupancy–abundance models can be significantly influenced by the distributional assumption and environmental correlates used in the modelling.

The analysis has also demonstrated that only a small portion of the variation in these parameters has been explained by the covariates used even in the best model selected by AIC. This highlights the need for considering more complex cases of hierarchical designs and additional explanatory variables to improve accuracy of abundance parameters. For example, in Zambia (Sileshi et al., 2002), stratification of trees into canopies explained 6.5% of the deviances in *M. ochroptera* abundance. In most datasets, the first ranked occupancy–abundance model was also not very different from the second or third-ranked model in its likelihood of being the best. This indicates the relatively high amount of uncertainty regarding the best model. When no single model is clearly the best, we cannot base predictions on the model ranked in first place. Another subset of predictors could perform as well as the one chosen, or if certain parameters are included the rank of the selected model may change.

Except for *O. bennigseni*, the occupancy predicted by the standard Poisson showed poorer fit to the observed occupancy compared with those from the ZIP, NBD and ZINB models. This may be due to temporal and spatial variations inducing heterogeneity that could not be adequately accounted by the Poisson models. The fundamental problem is that a Poisson density predicts the probability of zeros to be considerably less than is actually observed in a sample. Predictions of occupancy from the NBD also showed poorer fit to the observed. Just like the Poisson, the standard NBD does not allow for additional zeros than predicted by the NBD. Another problem with the NBD has been the variability in $k$ and its doubtful relationship with contagion (Taylor et al., 1979). Besides varying with environmental correlates of abundance and zero-inflation (Tables 2 and 3), $k$ also depends on sample size (Lloyd-Smith, 2007). Small sample sizes from the NBD may lead to systematic underestimation of the mean and variance and overestimation of $k$. This is because small samples are less likely to include values from the right-hand tail of the NBD, without which datasets appears more homogeneous (Lloyd-Smith, 2007). If these issues are not considered carefully when estimating $k$, the NBD occupancy–abundance relationship may be distorted.

Wilson and Room's model was generally the poorest predictor of occupancy relative to the observed. The poor performance of this model (and He and Gaston's model to some extent) may be attributed partly to the drawbacks of TPL. From Fig. 1 it can be seen that TPL systematically underestimates the variance at all mean densities. It has also been repeatedly pointed out that TPL has the unrealistic characteristic of predicting $s^2 < m$ when it is extrapolated to low densities (Lepš, 1993; Routledge and Swartz, 1991; Yamamura, 2000). With a decrease in $m$ (particularly when $m < 1$), the predicted $s^2$ becomes lower than the theoretical minimum (Lepš, 1993). This is especially common where each sampling unit either contains singletons or is empty. Thus TPL is bound to give a bad relative approximation of $\sigma^2$ for small $\mu$ unless $\beta = 1$, or $\beta < 1$ and $\sigma^2/\mu$ is exceedingly large. Furthermore, if for small $\mu$, the sporadically occurring individuals are likely to show up as singletons, then $\sigma^2/\mu \to 1$ as $\mu \to 0$. In this instance, TPL can accurately predict the variance for small $\mu$ only if it predicts a Poisson relationship with $\sigma^2 = \mu$ (Routledge and Swartz, 1991). TPL would not be expected to hold at densities below $\mu = \alpha^{(1/(1/1-\beta))}$, since whatever are the behavioural and demographic processes which underpin Eq. (3) will no longer be operating in the same way at densities below that for which the population is distributed randomly (Perry and Woiwod, 1992; Yamamura, 2000).

Other statistical problems in fitting TPL include random variation in $m$, which may cause serious bias in $\alpha$ and $\beta$ estimates. This is not easily solved, because it is worse for small densities, when $\sigma^2 = \mu$ (Perry and Woiwod, 1992). In small sample estimation of TPL parameters there are also other problems (Clark and Perry, 1994) including (1) exclusion of samples for which $m = s^2 = 0$; (2) exclusion of samples for which $s^2 = 0$, but $m > 0$; (3) restrictions on the maximum and minimum variance expressible in a sample; (4) underestimation of $\log(s^2)$ for skew distributions; and (5) the limited set of possible values of $m$ and $s^2$. In the present analysis, the major problem was exclusion of samples for which $m = s^2 = 0$ and $s^2 = 0$ but $m > 0$. For example, in the case of *M. ochroptera* 59 mean and variance pairs (41% of the points) were removed. Similarly, 42.9% and 30.0% of the mean and variance pairs were removed from the *O. bennigseni* and *G. simplex* datasets. For the same reason, Taylor and Woiwod (1982) omitted over 60% of the 1080 species available for analysis. Taylor and Woiwod (1982) warned that this problem introduces artefacts which diminish the primary regression coefficient ($\beta$) and raise the intercept ($\log \alpha$). As can be seen in Fig. 2, if there are many zeroes in the data, the calculated variance will be an underestimate of the true value, sometimes a bad one. This makes Wilson and Room's and He and Gaston's models inadequate for rare species and zero-inflated counts.

The ZINB allowed for more zeros, and hence the occupancy it predicted agreed with the observed more closely than the standard Poisson, ZIP and NBD. However, the occupancy predicted by all models appears to increase faster than the observed as density increased. This is probably because the rate of increase in population size may be different from occupancy rates. This may also be due to the fact that the behavioural and demographic processes which underpin occupancy rates at the lower density ranges are different from those at higher densities. Therefore, further studies and modelling efforts are needed to unravel the underlying cause of mismatch between the model predictions and reality at high densities.

The general conclusion from this analysis is that standard count models such as Poisson and NBD fail to account for zero-inflation and hence their respective occupancy–abundance models are inadequate for ecological count datasets with many zeros. Abundance–variance–occupancy models may also be inadequate for such datasets as TPL parameters may be seriously biased in the presence of zero-inflation. We have demonstrated that parameters used in occupancy–abundance models depend on the distributional assumption and predictors of abundance. Appropriate stratification of the habitat and use of site-specific or time-specific covariates that adequately captures information on the heterogeneity in counts and zero-inflation can improve parameter estimates. We recommend the use of hierarchical designs for sampling rare species and analytical tools such as the one proposed here for parameter estimation. The methods we propose have not been tested on count data that are not zero-inflated. Therefore, conclusions cannot be made about their general applicability to all species. Further studies are needed to test their performance over a range of species.

## Acknowledgements

The authors gratefully acknowledge financial support from the Canadian International Development Agency (CIDA), Swedish International Development Agency (Sida) and Irish Aid. The authors would also like to thank the anonymous reviewers for their useful comments that substantially improved this paper.

## Appendix A.

Table A.1.

**Table A.1**
Candidate models of *M. ochroptera* abundance assuming standard and zero-inflated Poisson and negative binomial distributions.

| Distribution | Candidate model | Model structure |
|---|---|---|
| Standard Poisson or NBD | 1. Null | $\alpha_0 + u$ |
| | 2. Year (Y) | $\alpha_0 + \alpha_1(\text{Y}) + u$ |
| | 3. Date (D) | $\alpha_0 + \alpha_2(\text{D}) + u$ |
| | 4. Site (S) | $\alpha_0 + \alpha_3(\text{S}) + u$ |
| | 5. Treatment (T) | $\alpha_0 + \alpha_4(\text{T}) + u$ |
| | 6. Year + date | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_2(\text{D}) + u$ |
| | 7. Year + site | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_3(\text{S}) + u$ |
| | 8. Year + treatment | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_4(\text{T}) + u$ |
| | 9. Date + site | $\alpha_0 + \alpha_2(\text{D}) + \alpha_3(\text{S}) + u$ |
| | 10. Date + treatment | $\alpha_0 + \alpha_2(\text{D}) + \alpha_4(\text{T}) + u$ |
| | 11. Site + treatment | $\alpha_0 + \alpha_3(\text{S}) + \alpha_4(\text{T}) + u$ |
| | 12. Year + date + site | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_2(\text{D}) + \alpha_3(\text{S}) + u$ |
| | 13. Year + date + treatment | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_2(\text{D}) + \alpha_4(\text{T}) + u$ |
| | 14. Year + site + treatment | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_3(\text{S}) + \alpha_4(\text{T}) + u$ |
| | 15. Date + site + treatment | $\alpha_0 + \alpha_2(\text{D}) + \alpha_3(\text{S}) + \alpha_4(\text{T}) + u$ |
| | 16. Year + date + site + treatment | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_2(\text{D}) + \alpha_3(\text{S}) + \alpha_4(\text{T}) + u$ |
| Zero-inflated Poisson or NBD | 1. Null | $\alpha_0 + u;\ \beta_0$ |
| | 2. Year | $\alpha_0 + \alpha_1(\text{Y}) + u;\ \beta_0 + \beta_1(\text{Y})$ |
| | 3. Date | $\alpha_0 + \alpha_2(\text{D}) + u;\ \beta_0 + \beta_2(\text{D})$ |
| | 4. Site | $\alpha_0 + \alpha_3(\text{S}) + u;\ \beta_0 + \beta_3(\text{S})$ |
| | 5. Treatment | $\alpha_0 + \alpha_4(\text{T}) + u;\ \beta_0 + \beta_4(\text{T})$ |
| | 6. Year + date | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_2(\text{D}) + u;\ \beta_0 + \beta_1(\text{Y}) + \beta_2(\text{D})$ |
| | 7. Year + site | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_3(\text{S}) + u;\ \beta_0 + \beta_1(\text{Y}) + \beta_3(\text{S})$ |
| | 8. Year + treatment | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_4(\text{T}) + u;\ \beta_0 + \beta_1(\text{Y}) + \beta_4(\text{T})$ |
| | 9. Date + site | $\alpha_0 + \alpha_2(\text{D}) + \alpha_3(\text{S}) + u;\ \beta_0 + \beta_2(\text{D}) + \beta_3(\text{S})$ |
| | 10. Date + treatment | $\alpha_0 + \alpha_2(\text{D}) + \alpha_4(\text{T}) + u;\ \beta_0 + \beta_2(\text{D}) + \beta_4(\text{T})$ |
| | 11. Site + treatment | $\alpha_0 + \alpha_3(\text{S}) + \alpha_4(\text{T}) + u;\ \beta_0 + \beta_3(\text{S}) + \beta_4(\text{T})$ |
| | 12. Year + date + site | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_2(\text{D}) + \alpha_3(\text{S}) + u;\ \beta_0 + \beta_1(\text{Y}) + \beta_2(\text{D}) + \beta_3(\text{S})$ |
| | 13. Year + date + treatment | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_2(\text{D}) + \alpha_4(\text{T}) + u;\ \beta_0 + \beta_1(\text{Y}) + \beta_2(\text{D}) + \beta_4(\text{T})$ |
| | 14. Year + site + treatment | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_3(\text{S}) + \alpha_4(\text{T}) + u;\ \beta_0 + \beta_1(\text{Y}) + \beta_3(\text{S}) + \beta_4(\text{T})$ |
| | 15. Date + site + treatment | $\alpha_0 + \alpha_2(\text{D}) + \alpha_3(\text{S}) + \alpha_4(\text{T}) + u;\ \beta_0 + \beta_2(\text{D}) + \beta_3(\text{S}) + \beta_4(\text{T})$ |
| | 16. Year + date + site + treatment | $\alpha_0 + \alpha_1(\text{Y}) + \alpha_2(\text{D}) + \alpha_3(\text{S}) + \alpha_4(\text{T}) + u;\ \beta_0 + \beta_1(\text{Y}) + \beta_2(\text{D}) + \beta_3(\text{S}) + \beta_4(\text{T})$ |

In the model structure, $\alpha_0$ is the intercept; $\alpha_1$, $\alpha_2$, $\alpha_3$, an $\alpha_4$ are the coefficients in the linear predictors of the Poisson or NBD mean. The u is the random effect (tree in the case of *M. ochroptera*). For the zero-inflated models, the terms after the semicolon (;) represent the zero part of the model, with $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, and $\beta_4$ being the coefficients of the Poisson or NBD mean.

## References

Agresti, A., Booth, J.G., Hobert, J.P., Caffo, B., 2000. Random-effects modeling of categorical response data. Sociol. Method. 30, 27–80.

Anderson, R.M., May, R.M., 1985. Helminth infection of humans: mathematical models, population dynamics and control. Adv. Parasit. 24, 1–101.

Brown, J.H., 1984. On the relationship between abundance and distribution of species. Am. Nat. 124, 255–279.

Clark, S.J., Perry, J.N., 1994. Small sample estimation for Taylor's power law. Environ. Ecol. Stat. 1, 287–302.

Cunningham, R.B., Lindenmayer, D.B., 2005. Modeling count data of rare species: some statistical issues. Ecology 86, 1135–1142.

Gagné, P., Dayton, C.M., 2002. Best regression model using information criteria. J. Mod. Appl. Stat. Methods 1, 479–488.

Gaston, K., Blackburn, T.M., Greenwood, J.J.D., Gregory, R., Quinn, R.M., Lawton, J.H., 2000. Abundance–occupancy relationships. J. Appl. Ecol. 37, 39–59.

Gaston, K.J., Borges, P.A.V., He, F., Gaspar, C., 2006. Abundance, spatial variance and occupancy: arthropod species distribution in the Azores. J. Anim. Ecol. 75, 646–656.

Gray, B.R., 2005. Selecting distributional assumption for modelling relative densities of benthic macroinvertebrates. Ecol. Model. 185, 1–12.

Hall, D.B., 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics 56, 1030–1039.

He, F., Gaston, K.J., 2003. Occupancy, spatial variance, and the abundance of species. Am. Nat. 162, 366–375.

He, F., Gaston, K.J., Wu, J., 2002. On species occupancy–abundance models. Ecoscience 9, 119–126.

Johnson, J.B., Omland, K.S., 2004. Model selection in ecology and evolution. Trends Ecol. Evol. 19, 101–108.

Lepš, J., 1993. Taylor's power law and the measurement of variation in the size of populations in space and time. Oikos 68, 349–356.

Lloyd-Smith, J.O., 2007. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PloS ONE 2, E180.

MacKenzie, D.T., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A., Langtimm, C.A., 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83, 2248–2255.

Martin, T.G., Wintel, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A., Possingham, H.P., 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol. Lett. 8, 1235–1246.

Mwalili, S.M., Lesaffre, E., Declerck, D., 2008. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. Stat. Method. Med. Res. 17, 123–139.

Perry, J.N., Woiwod, I.P., 1992. Fitting Taylor's power law. Oikos 65, 538–542.

Piñeiro, G., Perelman, S., Guerschman, J.P., Paruelo, J.M., 2008. How to evaluate models: observed vs predicted or predicted vs observed. Ecol. Model. 216, 316–322.

Pinheiro, J.C., Bates, D.M., 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. J. Comp. Graph. Stat. 4, 12–35.

Routledge, R.D., Swartz, T.B., 1991. Taylor's power law re-examined. Oikos 60, 107–112.

Saha, K., Paul, S., 2005. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. Biometrics 61, 179–185.

Sileshi, G., 2006. Selecting the right statistical model for analysis of insect count data by using information theoretic measures. Bull. Entomol. Res. 96, 479–488.

Sileshi, G., 2008. The excess-zero problem in soil animal count data and choice of models for statistical inference. Pedobiologia 52, 1–17.

Sileshi, G., Mafongoya, P.L., 2007. Quantity and quality of organic inputs from coppicing leguminous trees influence abundance of soil macrofauna in maize crops in eastern Zambia. Biol. Fertil. Soils 43, 333–340.

Sileshi, G., Baumgaertner, J., Sithanantham, S., Ogol, C.K.P.O., 2002. Spatial distribution and sampling plans for *Mesoplatys ochroptera* Stål (Coleoptera: Chrysomelidae) on sesbania. J. Econ. Entomol. 95, 499–506.

Sileshi, G., Girma, H., Mafongoya, P.L., 2006. Occupancy–abundance models for predicting densities of three leaf beetles damaging the multipurpose tree *Sesbania sesban* in eastern and southern Africa. Bull. Entomol. Res. 96, 61–69.

Taylor, L.R., 1961. Aggregation, variance and the mean. Nature 189, 732–735.

Taylor, L.R., Woiwod, I.P., 1982. Comparative synoptic dynamics. I. Relationship between inter- and intra-specific spatial and temporal variance/mean population parameters. J. Anim. Ecol. 51, 879–906.

Taylor, L.R., Woiwod, I.P., Perry, J.N., 1979. The negative binomial as a dynamic ecological model and the density-dependence of *k*. J. Anim. Ecol. 48, 289–304.

Tooze, J.A., Grunwald, G.K., Jones, R.H., 2002. Analysis of repeated measures data with clumping at zero. Stat. Methods Med. Res. 11, 341–355.

Warton, D.I., 2005. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16, 275–289.

Welsh, A.H., Cunningham, R.B., Donnelly, C.F., Lindenmayer, D.B., 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecol. Model. 88, 297–308.

Wilson, L.T., Room, P.M., 1983. Clumping patterns of fruit and arthropods in cotton, with implications for binomial sampling. Environ. Entomol. 12, 50–54.

Wilson, P.D., 2008. The pervasive influence of sampling and methodological artefacts on a macroecological pattern: the abundance–occupancy relationship. Global Ecol. Biogeogr. 17, 457–464.

Yamamura, K., 2000. Colony expansion model for describing the spatial distribution of populations. Popul. Ecol. 42, 161–169.