Pedobiologia **52** (2008) 1—17

Available online at www.sciencedirect.com

ScienceDirect

Pedo biologia

www.elsevier.de/pedobi

The excess-zero problem in soil animal count data and choice of appropriate models for statistical inference

Gudeta Sileshi[*]

World Agroforestry Centre (ICRAF), SADC-ICRAF Agroforestry Programme, Chitedze Agricultural Research Station, P.O. Box 30798, Lilongwe, Malawi

Received 22 September 2006; received in revised form 22 November 2007; accepted 23 November 2007

KEYWORDS
Information measures;
Negative binomial distribution;
Poisson;
Zero inflation

Summary

Recent studies show that soil animal count data are characterized by the presence of excess zeros and overdispersion, which violate the assumptions of standard statistical tests. Despite this, analyses have consisted of mainly non-parametric tests and log-normal least square regression (i.e. ANOVA). Failure to accommodate zero inflation in count data can result in biased estimation of ecological effects jeopardizing the integrity of the scientific inference. The objective of this study was to compare statistical models for the analysis of soil animal count data and suggest appropriate methods for estimating abundance. The log-normal regression model, linear mixed model (LMM), standard Poisson, Poisson with correction for overdispersion (PCO), negative binomial distribution (NBD), the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models were compared using 12 count data sets of earthworms, millipedes, centipedes, beetles, ants and termites from soils under the miombo woodland and agroforestry systems in eastern Zambia. The NBD with covariates gave a better description of the data in nine out of 12 cases than did the standard Poisson, ZIP and ZINB. The ZIP and ZINB models with covariates gave the best description of earthworm counts from the miombo and millipede counts from agroforestry, respectively. In all cases, the ZIP model was better than the standard Poisson model. The ZINB was inferior to the NBD except for earthworm counts from the miombo and millipede counts in agroforestry. Significance tests based on the PCO, ZIP, NBD and ZINB were more conservative than those based on the standard Poisson model. The 95% confidence intervals computed using the PCO, ZIP, NBD and ZINB were also wider than those computed using least squares, LMM and assuming Poisson distribution. It is concluded that for the comparison among habitat types, land-use categories or treatments, the NBD, ZIP and ZINB perform better than the log-normal and Poisson models. Considering the excess-zero problem and

[*]Tel.: +2651 794355, Mobile: +2659 642149; fax: +2651 707323.
E-mail address: sgwelde@yahoo.com.

0031-4056/$ - see front matter © 2008 Elsevier GmbH. All rights reserved.
doi:10.1016/j.pedobi.2007.11.003

significant deviation of soil animal counts from the assumptions of normality and homoscedcity, the log-normal regression model is inappropriate. Therefore, routine application of the log-normal regression model and non-parametric tests for analysis of soil animal count data with many zeros should be discouraged.

© 2008 Elsevier GmbH. All rights reserved.

Introduction

Soil animals are of immediate concern in activities such as agriculture, forestry and environmental monitoring. However, the complexity and diversity of soil animals and the habitats in which they live pose unique challenges to those seeking to quantify the effects of land-use and management practices on individual taxa or assemblages (Lavelle et al., 2003; Susilo et al., 2004). The method described by Anderson and Ingram (1993) has been widely used for sampling soil animals. One of the problems with data collected using this design has been the strong spatial aggregation of soil animals in the field (Lavelle et al., 2003; Jones et al., 2005; Sileshi and Mafongoya, 2007). For example, due to the patchy distribution of colonies and individuals of termites and ants within habitats, most small- to medium-sized soil samples will contain relatively low number of individuals or none at all, while very few samples may have extremely large numbers if a nest or foraging party is encountered (Jones et al., 2005; Kilpeläilen et al., 2005; Sileshi and Mafongoya, 2007). Density estimates can therefore have high variance, making it difficult to demonstrate statistically significant differences among sites, land-use practices or treatments even if effects are relatively large (Lavelle et al., 2003; Susilo et al., 2004; Jones et al., 2005; Sileshi and Mafongoya, 2007).

Recent studies have revealed that soil animal counts exhibit two features: a substantial proportion of the values are zero and the remainder has a skewed distribution (Sileshi and Mafongoya, 2006a; Sileshi and Mafongoya, 2007). When the frequency of zeros is so large that the data do not readily fit standard distributions, the data set is referred to as zero inflated (Lambert, 1992; Martin et al., 2005). Statisticians make distinctions between structural zeros, which are inevitable, and sampling zeros, which occur by chance. Structural zeros consist of a large number of true zero observations caused by the real ecological effect 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). Sampling zeros often referred to as 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.

Zero inflation, a special case of overdispersion, creates problems with making sound statistical inference by violating basic assumptions implicit in standard distributions (Martin et al., 2005; Sileshi, 2006). If not properly modelled, overdispersion can lead to underestimation of the standard errors of regression parameters, confidence intervals that are too narrow, and *P*-values that are too small. This can result in biased estimation of ecological effects and jeopardize the integrity of the scientific inferences (Lambert, 1992; Martin et al., 2005; Sileshi, 2006). Standard statistical texts rarely discuss this problem and it is only recently that software for modelling count processes that accommodate excess zeros has emerged.

The most common analyses used for soil animals consisted of either non-parametric tests (Swift and Bignell, 2001; Jabin et al., 2004) or log-normal least squares regression (e.g. ANOVA), both of which do not deal overdispersion. The log-normal regression is generally inappropriate for modelling a discrete process. When testing for habitat, land-use or treatment effects, the distributional assumptions made about the response variable can have a critical impact on the conclusions drawn. Often data do not support only one model as clearly best for analysis (Dayton, 2003; Johnson and Omland, 2004). This raises the issue of comparing models to assess which ones are adequate for the data and which one could be chosen as the basis for interpretation, prediction, or other subsequent use. Therefore, the objective of this study was to compare statistical models for the analysis of soil animal count data and suggest appropriate methods for estimating abundance.

Materials and methods

Sources of data

The data used in this study were collected from the miombo woodland and agroforestry systems in eastern Zambia. These were reported elsewhere (Sileshi and Mafongoya, 2006a, b, 2007). The data collected from the miombo were used to investigate the effect of forest fire on soil animal communities (Sileshi and Mafongoya, 2006a), while those collected from the agroforestry systems were used to quantify temporal variations in macrofauna in relation to different land-use categories (Sileshi and Mafongoya, 2006b). The data from the miombo represent observational studies, where as those from the agroforestry are typical of experimental studies.

In the miombo, three sampling sites about 2 km apart from each other were located in the Msekera area where patches of the secondary miombo were affected by forest fires in July–September 2003 and 2004. At each site, soil samples were collected from forest patches that were burnt and adjacent patches that were not affected by fire (hereafter referred to as unburnt). This was done four times between December 2003 and November 2004 to coincide with contrasting periods in the climatic cycle of the study area; (1) in December – the beginning of the rainy season, (2) in February – mid-rainy season, (3) in July – mid-dry season, and (4) in November – end of the dry season. A total of 16, 26, 18 and 28 samples were collected from the same patches in December, February, July and November, respectively. Half of the samples were from the burnt patches and the other half from unburnt patches in adjacent areas (Sileshi and Mafongoya, 2006a). The sample size was not equal across months due to logistic constraints related to collection and processing soil samples.

In the agroforestry practices, a total of 356 soil samples were collected from maize grown using leguminous agroforestry species and continuous monoculture maize in December 2003, February 2004, July 2004 and February 2005 at Msekera and Kalunga sites. A stratified-random sampling procedure was followed when sampling the agroforestry according to tree species, which differed in the quality and quantity of their organic inputs (Sileshi and Mafongoya, 2007). Five treatments were compared in the agroforestry system: maize monoculture, maize grown after pure species fallows of four legume species, namely, *Gliricidia sepium*, *Acacia anguistissima*, *Leucaena collinsi* and *Calliandra calothyrsus* (Sileshi and Mafongoya, 2006b). The treatments were replicated three times.

In the miombo woodlands and agroforestry systems, samples were collected using a soil monolith (25 cm × 25 cm and 25 cm depth) placed over a randomly selected spot (Anderson and Ingram, 1993; Swift and Bignell, 2001), and driven into the soil to ground level using a metallic mallet. Three samples were taken from each treatment replicated three times making a total of nine per treatment. From each soil monolith, macrofauna were hand-sorted to a family or order level and numbers recorded.

The statistical models

The first method involved ordinary least squares (OLS) regression (i.e. ANOVA). The probabilistic model underlying OLS regression models assumes that transformed data follow an approximate log-normal (Gaussian) distribution (i.e. the model errors are independently and identically distributed normal random variates). However, soil animal count data often depart from this ideal situation (Jabin et al., 2004; Sileshi and Mafongoya, 2007). Therefore, the data were transformed as log (count+1). These data were explicitly tested for normality and homogeneity of variance using the Shapiro-Wilk statistic (W) and Levene's test, respectively, before conducting analysis. The UNIVARIATE procedure of the SAS system (SAS, 2003) was used to test normality, while PROC GLM was used to test homogeneity of variance and other analyses. The SAS codes used to generate tests of homogeneity of variance (HOVTEST) and one- and two-way ANOVA are presented in Appendix 1A. Log-transformed counts of earthworms (tearthw) from the miombo (data = miombo) were used to illustrate the procedures.

The second method involved linear mixed modelling (LMM) of the count data using the MIXED procedure of SAS. LMMs extend the OLS regression model by providing a more flexible specification of the covariance matrix of the error, and allow for both correlation and heterogeneous variances. However, as in the OLS regression model, it is assumed that the data are normally distributed. The MIXED procedure fits the covariance structure one selects for the data using the method of restricted/residual maximum likelihood (REML) or maximum likelihood (ML). In the present study, the ML estimation method was used. Using the ML estimation method, the SAS procedures presented in Appendix 1B fit the full range of LMMs (null model, single-variable and two-variable models) to the log-transformed counts of earthworms (i.e. tearthw) data set from the miombo.

The third method was based on a Poisson generalized linear regression model (Lawless, 1987; Cameron and Trivedi, 1998). The Poisson distribution on abundance is a natural choice because it arises under the assumption that animals are distributed randomly in space. Poisson regression involves explicitly modelling the distribution of counts assuming that the variance (σ^2) is proportional to the mean (λ), say $\sigma^2 = \phi E(y) = \phi\lambda$ where ϕ is a dispersion parameter and $E(y)$ is the expectation of counts (Cameron and Trivedi, 1998). The variance equals the mean when $\phi = 1$, while $\phi > 1$ indicates overdispersion in the Poisson model. Parameters of the standard Poisson model were estimated for counts of the various soil animals using the GENMOD procedure of SAS. The GENMOD procedure fits generalized linear models (GLMs) using ML method. GLMs are an extension of traditional linear models. However, they allow the mean (λ or μ) of a population to depend on a linear predictor through a nonlinear link function and permit the response probability distribution to be any member of the exponential family of distributions. In the GLMs used here, the animal counts (Y_i) varying over sampling units ($i = 1, 2, \ldots, n$) were assumed to have a specified distribution (in this case Poisson) whose parameters depend on a vector of linear predictors (X_i such as treatment, time, site) according to a log-linear function: $\log \mu_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$, where α and β_i are regression coefficients to be estimated. This can be fitted sequentially starting from a null model (assuming no covariate effect) through single-variable models to n-variable main effects models. For example, the GENMOD procedures in Appendix 1C fit the full range of Poisson models to the untransformed earthworm counts (earthwo) from the miombo.

Although Poisson regression is the recommended approach for analyzing count data, it often does not fit overdispersed data very well. In modelling overdispersed count data, quasi-likelihood adjustments are normally made when a reasonable lack of fit to the standard Poisson is found (McCullagh and Nelder, 1989). This is based on a single variance inflation factor estimated by ML or, optionally, by the residual deviance or Pearson's chi-square divided by the associated degrees of freedom (SAS, 2003). The introduction of the variance inflation factor, however, does not introduce a new probability distribution. It adjusts the standard errors and provides wider confidence intervals and *P*-values larger than what is obtained under the standard Poisson model. In the present study, this method is termed Poisson with correction for overdispersion (PCO). SAS implements this by introducing an option SCALE = D or SCALE = P in the model statement of the GENMOD procedure for the Poisson. The rest of the GENMOD syntax will be the same as in the standard Poisson example above. For the earthworm count data used in the previous example, the GENMOD procedure in Appendix 1D estimated parameters of PCO models.

The PCO produces an appropriate inference only if overdispersion is modest (Cox, 1983). For heavily overdispersed count data, the negative binomial distribution (NBD) is more appropriate (Lawless, 1987). The NBD is characterized by the dispersion parameter k and the mean μ, and its variance is equal to μ+$k\mu^2$. According to Johnson and Kotz (1969), the NBD is a mixture of Poisson distributions such that the expected values of the Poisson distribution vary according to a gamma (Type III) distribution. It has been shown that the limiting distribution of the NBD, as the dispersion parameter (k) approaches zero, is the Poisson. When k is an integer, the NBD becomes the Pascal distribution, and the geometric distribution corresponds to $k = 1$. The log series distribution occurs when zeros are missing and as $k \to \infty$. As in the Poisson model, parameters of the NBD were estimated using the GENMOD procedure. The only difference is that, in this case DIST = NB is used instead of DIST = POISSON. The rest of the GENMOD model statement syntax will be the same. For the earthworm count data used in the Poisson example, the null and full NBD models may be fitted using the SAS codes in Appendix 1E.

For modelling count data with excess zeros, zero-inflated Poisson (ZIP) models have been proposed (Lambert, 1992). ZIP models apply when a large proportion of the sampling units have zero counts, and for the remainder the Poisson parameter takes the fixed value λ (Lambert, 1992). In some cases, the ZIP regression is often inadequate. To remedy this, zero-inflated negative binomial (ZINB) models have been adopted for ecological count data (Martin et al., 2005; Warton, 2005). The basic idea is to mix a distribution degenerate at zero with NBD. In this study, the ZIP and ZINB models were fitted to the various soil invertebrate data using the nonlinear mixed effects models (NLMIXED) procedure of SAS. NLMIXED belongs to a class of models called generalized linear mixed models (GLMMs). For distributions from the exponential family, GLMMs extend GLMs by including random effects in the linear predictor (McCulloch and Searle, 2001). This allows modeling the process of change within individuals in clustered data. This procedure also produces parameter estimates of the standard Poisson and NBD distributions. Details of the SAS codes used for estimating parameters of the standard Poisson, NBD, ZIP and ZINB models are given in Appendices 1F, 1G, 1H and 1I.

Comparing frequency distributions

The frequency distributions of the observed and expected (assuming the Poisson and NBD) number of individuals were compared. ML estimates of the dispersion parameter (k) were obtained for models with all covariates (full model) and without covariates (null model) using the GENMOD procedure. These were used to generate the expected frequencies under the NBD assumption. Expected frequencies were calculated by substituting the sample mean for μ and k (with and without covariates separately) into the probability functions of the Poisson and NBD (Eqs. (4) and (18) in Davis, 1994). Histograms were then generated for the percentage of 0, 1, …, n soil animal counts in the data.

Goodness of fit and criteria for model selection

Two statistics that are helpful in assessing the goodness of fit of a given generalized linear model are the scaled deviance (D) and Pearson's chi-square statistic (SAS, 2003). Adequacy of the Poisson and NBD regression models for the various soil animal count data was first checked using the ratio of the deviance to its associated degrees of freedom ($D/DF = \varphi$). If the regression model is adequate, the expected value of φ will be close to unity. Otherwise, the validity of the model could be doubtful.

A likelihood-ratio test may be conducted to compare the Poisson to the NBD since the Poisson is nested within the NBD. A more appropriate approach for comparing non-nested models is the use of information measures such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC) (Johnson and Omland, 2004; Kuha, 2004). AIC and BIC have some optimal properties providing certain justification for choosing them out of the entire range of model selection criteria (Dayton, 2003; Johnson and Omland, 2004). AIC and BIC apply to nested and non-nested models alike (Golden, 2000; Dayton, 2003; Kuha, 2004; Posada and Buckley, 2004; Sober, 2004). It must be noted that the AIC and BIC arise from a more general form of information criterion (IC): $\mathrm{IC} = -2\mathrm{ll}+\theta c$. If $c = 0$, IC is equal to the classical likelihood-ratio statistic. If $c = 1$, IC is equal to the GLIM goodness-of-fit procedure based on plotting the deviance against degrees of freedom (Smith and Speigelhalter, 1980). If $c = 2$, IC is identical to AIC, and if $c = \log N$, IC is equal to BIC (Atkinson, 1981). It is also known that AIC ($c = 2$) is asymptotically equivalent to a cross-validation criterion (Stone, 1977).

For any specified model in this study, AIC and BIC were computed as $\mathrm{AIC} = -2\mathrm{ll}+2\theta$ and $\mathrm{BIC} = -2\mathrm{ll}+\theta(\log(n))$, respectively. Here ll is the log-likelihood for the model, θ is the number of independent parameters that are estimated in fitting the model and n is the sample size (Dayton, 2003). Since AIC does not depend directly on sample size it lacks certain properties of asymptotic consistency. However, in finite samples, adjusted versions of AIC such as the second-order Akaike information criterion ($\mathrm{AIC_c}$) (Hurvich and Tsai, 1989) perform much better (Johnson and Omland, 2004). Hence, $\mathrm{AIC_c}$ may be computed as follows:

$$\mathrm{AIC_c} = -2\mathrm{ll} + 2\theta + \frac{2\theta(\theta+1)}{n-\theta-1}.$$

Using $\mathrm{AIC_c}$ and BIC as a guide, first the Poisson, NBD, ZIP and ZINB models were compared separately. All soil animal data from both the miombo and agroforestry were used for this comparison. It must be noted that the log-likelihoods for Poisson and NBD (estimated using the GENMOD procedure) are not comparable with those of the ZIP and ZINB (estimated using the NLMIXED procedure) although the other coefficients are the same. Therefore, SAS codes of the NLMIXED procedure were written to obtain comparable $\mathrm{AIC_c}$ for the Poisson, NBD, ZIP and ZINB (Appendices 1C, 1E, 1F and 1G). It must also be noted that comparison of $\mathrm{AIC_c}$ for the LMM and PCO with those of the other models is not straightforward as the likelihoods estimated by the GENMOD and MIXED procedures differ.

Comparison of subset models (nested in each distribution model) was also made using BIC and variants of the AIC. To demonstrate this, only earthworm counts were used. Akaike weights ($\mathrm{AIC_w}$) were computed from $\mathrm{AIC_c}$, as these have the advantage of being easy to interpret than $\mathrm{AIC_c}$. The $\Delta\mathrm{AIC_c}$, which is a measure of each model relative to the best model was first calculated as $\Delta\mathrm{AIC}_{ci} = \mathrm{AIC}_{ci}–\mathrm{AIC}_{c\,\mathrm{min}}$, where AIC_{ci} is the $\mathrm{AIC_c}$ value for model i, and $\mathrm{AIC}_{c\,\mathrm{min}}$ is the $\mathrm{AIC_c}$ value of the "best" model. $\mathrm{AIC_w}$ was then calculated as follows:

$$\mathrm{AIC_w} = \frac{\exp(-0.5\Delta\mathrm{AIC_c})}{\sum\exp(-0.5\Delta\mathrm{AIC_c})}.$$

$\mathrm{AIC_w}$ indicates the probability that the model is the best among the whole set of candidate models. Therefore, it provides a measure of the strength of evidence for each model. To demonstrate the application of AIC and BIC for nested model selection, only the earthworm data is presented here. To demonstrate the impact of model choice on statistical significance of covariates and parameter estimates, soil animal data from the miombo were used.

Results

Frequency of zeros

Figures 1 and 2 present the frequency distribution of animals in soils under the miombo woodland and agroforestry systems, respectively. Zeros constituted 71% and 74% of the total earthworm counts in the data from the miombo and agroforestry, respectively. The respective figures for the millipede data from the miombo and agroforestry were 43% and 74%, while those for centipedes were 61% and 91%. The observed frequency of zeros in earthworm, millipede and centipede counts was closest to that expected under the NBD than the Poisson (Figure 1). Zeroes also constituted 17% and 41% of beetle counts in soils under the miombo and agroforestry, respectively. The observed frequency of zeros in beetle counts was larger than what is expected under Poisson distribution assumption. However, it was smaller than what is expected assuming NBD especially with covariate effects (Figure 2). Zeros represented 36% and 65% of the total ant counts in the miombo and agroforestry, while those for termites were 32% and 54%, respectively. Ant and termite counts had more frequency of zeros than that expected assuming either NBD or Poisson distributions (Figures 1 and 2). It must be noted that some artifacts may occur in the termite and ant count data as a result of clustering of bins at highest densities. The expected frequency of zeros in all soil animal count data sets was higher under the NBD with all covariates than the same model without covariate structure (Figures 1 and 2).

Model fit and adequacy

Shapiro-Wilks and Levene's tests indicated significant departure from normality and homogeneity of variance in most data sets (Table 1). The φ values for the Poisson model were larger than unity (data not shown) indicating overdispersion relative to the Poisson assumption. However, φ values were closer to unity under the NBD indicating a better fit for most soil animals. According to the AIC (Table 2) and BIC (data not shown), the NBD with covariates provided a better description of the data in 10 out of the 12 cases than did the standard Poisson, ZIP and ZINB. The full NBD model was best in describing earthworm and termite counts from agroforestry, centipede, beetle and ant counts from both miombo and agroforestry. The ZIP and ZINB models with covariates gave the best description of earthworm counts from the miombo and millipede counts from agroforestry, respectively. In all cases, the ZIP model was better than the standard Poisson model. The ZINB was inferior to the NBD in most cases except for earthworm counts from the miombo and millipede counts in agroforestry. None of the data sets were adequately described by the standard Poisson model. In all cases, the full models (with all covariates) performed better than the null models for all soil animal count data sets except for termites in the miombo (Table 2). Since the AIC_c for the LMM and PCO are not comparable with those of the other models, they are not presented here. The BIC selected the same model as the AIC. Therefore, comparisons using BIC are not presented here.

Statistical significance of covariates

Table 3 shows the difference among distribution models in the statistical significance (*P*-values) of treatment and time (month) effects. In this table nested models, i.e. (1) models with the intercept and main effect of treatment alone, (2) models with the intercept and main effect of time and (3) models with the intercept and the main effects of both treatment and time are also compared. Examination of the *P*-values suggests that our conclusions about the effect of treatments and time on animal counts are greatly affected by the choice of the model. For example, the Poisson model indicated highly significant ($P<0.0001$) treatment and time effects on all animals except centipedes even where the log-normal regression, LMM and NBD showed slight or no effects (Table 3). Significance tests based on the PCO and NBD were more conservative than those based on the standard Poisson model. Under each distribution model, the *P*-values for the main effects of treatment and time differed depending on whether they were analyzed separately or together (Table 3).

Using BIC and AIC, Table 4 compares the whole range of nested models (the null model, single-variable models, two-variable main effects models) for the earthworm count data sets. The ZIP model was the best for the data from the miombo, while the NBD was the best for the counts from agroforestry. For the earthworm count data from the miombo, the model with both main effects (treatment and time) was the best ($\text{AIC}_w = 1.0$). The null model and single-variable models had less than 1% likelihood. For the data from agroforestry systems, the best model which consisted of site had 38% likelihood, while models where site and month or site and treatment each had 24% likelihood. The full model has only 14% likelihood of being the correct model.

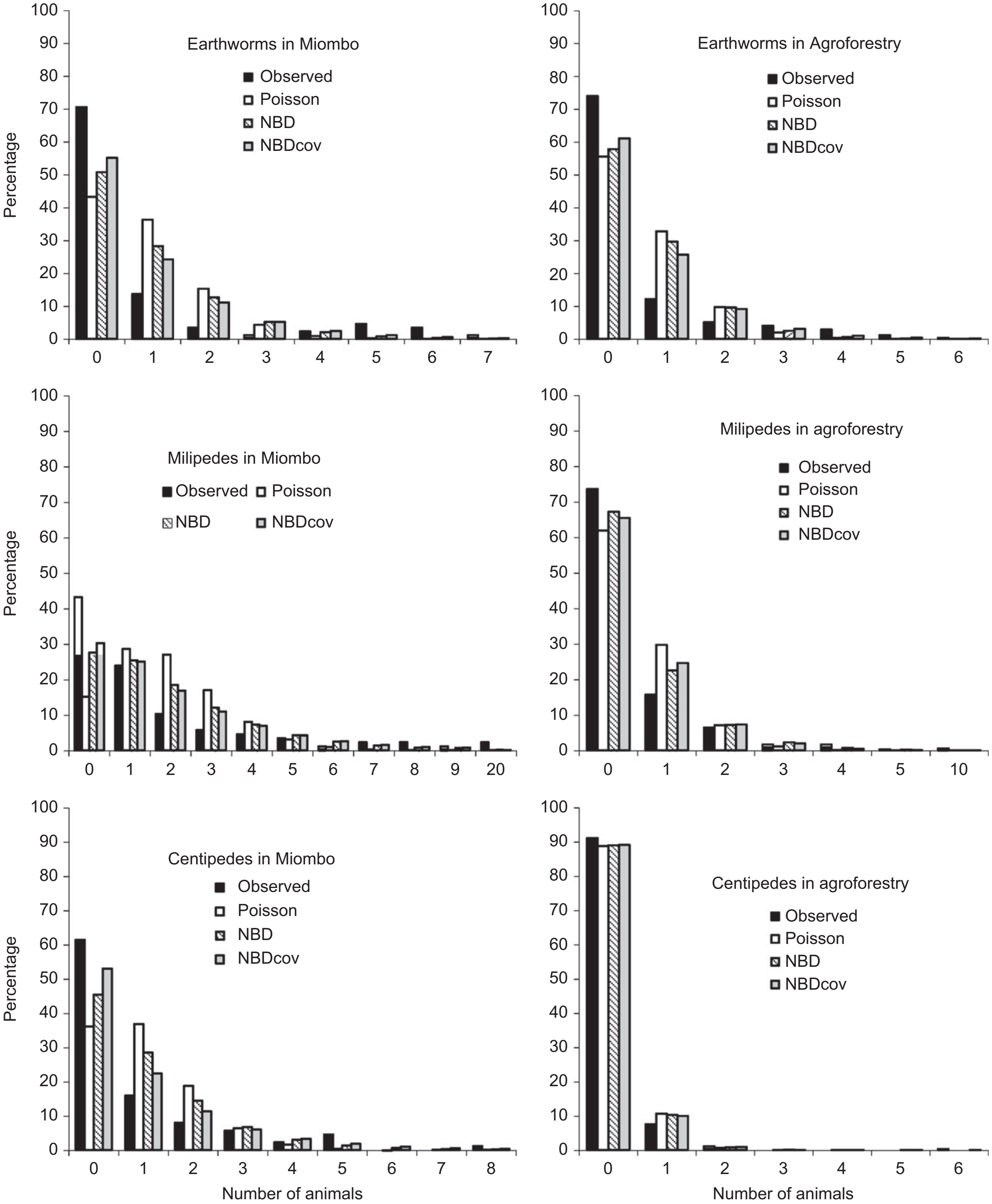

Figure 1. Frequency distribution of observed and expected counts of earthworms, millipedes and centipedes in soil monoliths taken from the miombo woodland and agroforestry plots in eastern Zambia. Histograms represent the percentage of 0, 1, ..., *n* soil animal counts in the data. NBDcov represents the negative binomial distribution model with all covariates while NBD represents the negative binomial distribution without covariates.

The most striking effect of model choices was on the standard errors of parameter estimates and the 95% confidence intervals (Table 5). The 95% confidence intervals computed using the PCO, ZIP, NBD and ZINB were also wider than those obtained using the OLS, LMM and standard Poisson.

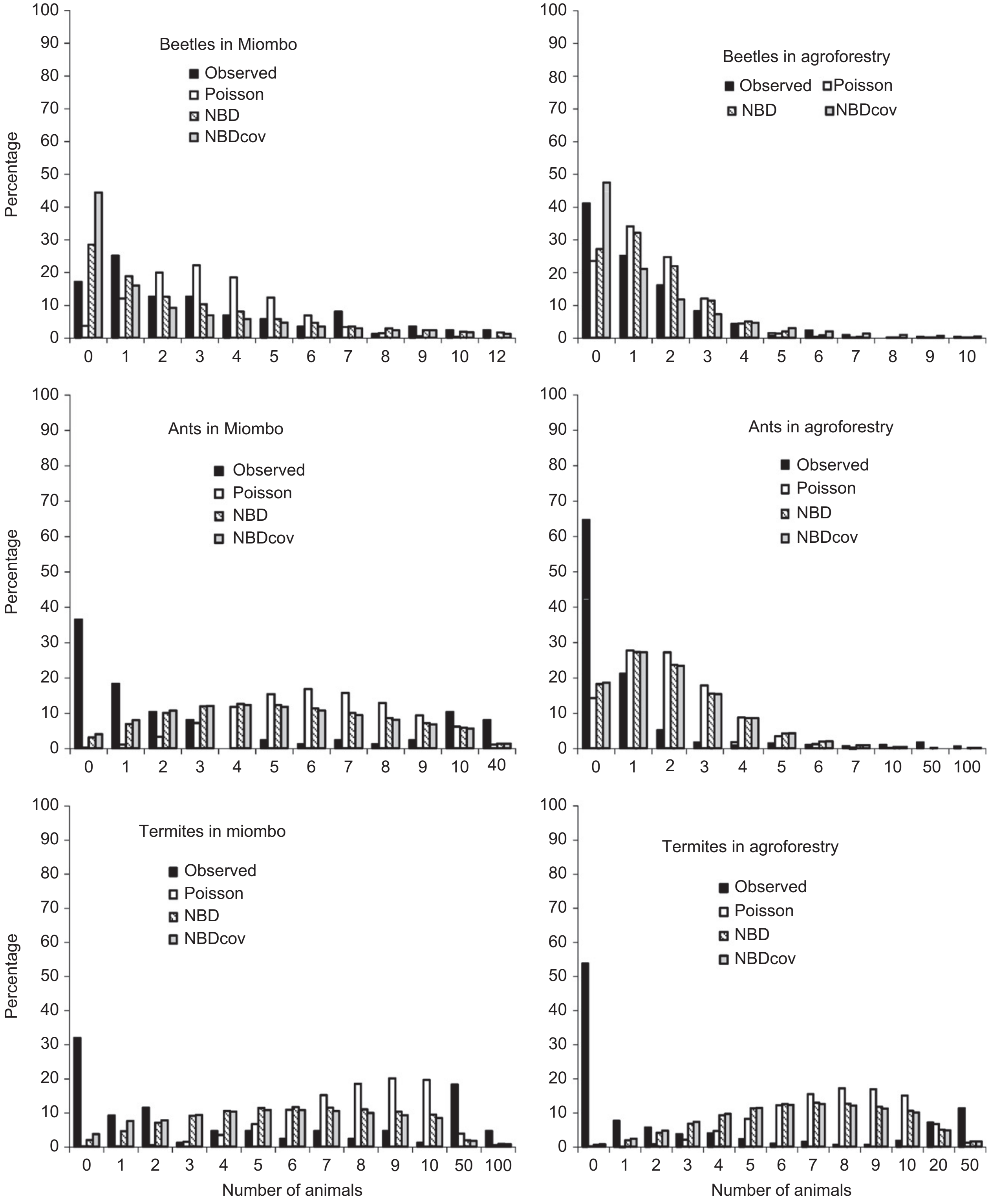

Figure 2. Frequency distribution of observed and expected counts of beetles, ants and termites in soil monoliths taken from the miombo woodland and agroforestry plots in eastern Zambia. Histograms represent the percentage of 0, 1, ..., *n* soil animal counts in the data. NBDcov represents the negative binomial distribution model with all covariates while NBD represents the negative binomial distribution without covariates.

Table 1. Shapiro-Wilk test of normality and Levene's test of homogeneity of variance of log-transformed soil invertebrate count data under the miombo woodland and agroforestry species

Invertebrate groups	Test of normality (Shapiro-Wilk)[a]		Levene's test of homogeneity of variance		
	Miombo	Agroforestry	Sources	Miombo	Agroforestry
Earthworms	0.62***	0.60***	Treatment	15.3***	0.9ns
			Month	15.8***	17.2***
			Site	NA	35.5***
Beetles	0.95***	0.86****	Treatment	0.3ns	2.2*
			Month	1.9ns	2.6*
			Site	NA	15.5***
Ants	0.85***	0.62***	Treatment	4.0*	0.7ns
			Month	1.3ns	0.7ns
			Site	NA	0.04ns
Termites	0.90***	0.76***	Treatment	0.02ns	1.2ns
			Month	2.8*	7.4***
			Site	NA	9.5***
Centipedes	0.72***	0.32***	Treatment	0.3ns	0.9ns
			Month	7.9***	3.9**
			Site	NA	4.8*
Millipedes	0.84***	0.60***	Treatment	0.9ns	2.3*
			Month	2.4ns	4.1**
			Site	NA	2.9ns

NA = not applicable.*, ** and *** indicate significance at the 5%, 1% and 0.1% levels, respectively.

[a] Shapiro-Wilk W statistic and its significance.

Table 2. Comparison of the linear mixed-model (LMM), standard Poisson, Poisson corrected for overdispersion (PCO), the negative binomial distribution (NBD), zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) null models (without covariate structure) and full models (with all covariates) for different soil invertebrate groups using Akaike information criteria (AIC_c)

Land-use	Animal group	Null model				Full models			
		Poisson	ZIP	ZINB	NBD	Poisson	ZIP	ZINB	NBD
Miombo	Earthworms	291	216	209	207	197	**172**	174	176
	Millipedes	475	405	622	324	420	382	321	**317**
	Centipedes	312	256	438	241	282	241	228	**227**
	Beetles	509	484	416	414	429	427	396	**390**
	Ants	1658	1564	463	460	1419	1142	460	**453**
	Termites	2068	1943	529	**527**	3869	3263	560	553
Agroforestry	Earthworms	898	719	705	706	819	688	683	**682**
	Millipedes	716	681	649	647	723	652	**615**	630
	Centipedes	288	272	265	263	262	255	250	**249**
	Beetles	1403	1303	1184	1182	1260	1207	1132	**1123**
	Ants	3455	2495	1031	1030	3388	2413	1036	**1031**
	Termites	10853	6486	1706	1704	10061	6345	1698	**1698**

For each soil invertebrate group (within a row), the best model is indicated by bold AIC_c scores. Models within a column should not be compared.

Since this pattern was consistently observed in the data sets from both the miombo and agroforestry, for brevity only the miombo data set was presented here.

Discussion

For most of the taxa studied, the data sets significantly deviated from the assumptions of the

Table 3. Impact of choice of non-nested and nested model choice on statistical significance (*P*-value) of treatment and time (month) effects on abundance of soil invertebrate groups in the miombo woodland

Invertebrate groups	Nested model	Variables in model	OLS	LMM	Poisson	PCO	ZIP[a]	NBD	ZINB[a]
Earthworms	1	Treatment	0.0004	0.0008	<0.0001	<0.0001	0.0434	0.0005	0.3902
	2	Month	0.0002	0.0002	<0.0001	<0.0001	0.0009	<0.0001	0.4128
	3	Treatment	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
	3	Month	<0.0001	<0.0001	<0.0001	<0.0001	0.0029	<0.0001	<0.0001
Millipedes	1	Treatment	0.0550	0.1024	<0.0001	0.0381	0.0720	0.0574	0.2023
	2	Month	0.0665	0.0396	<0.0001	0.0051	<0.0001	0.0164	0.0019
	3	Treatment	0.0447	0.0388	<0.0001	0.0234	0.3530	0.1020	0.8440
	3	Month	0.0555	0.0457	<0.0001	0.0033	<0.0001	0.0252	0.0006
Centipedes	1	Treatment	0.5516	0.9980	0.9962	0.9976	0.0958	0.9980	0.2820
	2	Month	<0.0001	<0.0001	<0.0001	<0.0001	0.2049	<0.0001	0.8299
	3	Treatment	0.3717	0.3577	0.8789	0.9000	0.0801	0.2951	0.5364
	3	Month	<0.0001	<0.0001	<0.0001	<0.0001	0.1346	<0.0001	0.7675
Beetles	1	Treatment	<0.0001	0.0003	<0.0001	<0.0001	<0.0001	<0.0001	0.0013
	2	Month	0.0053	0.0045	<0.0001	0.0011	0.0011	0.0015	0.0003
	3	Treatment	<0.0001	<0.0001	<0.001	<0.0001	<0.0001	<0.0001	0.0014
	3	Month	0.0013	0.0009	<0.0001	0.0002	<0.0001	0.0006	0.0002
Ants	1	Treatment	0.0130	0.0192	<0.0001	0.0008	<0.0001	0.0036	0.0124
	2	Month	0.0119	0.2929	<0.0001	0.0456	<0.0001	0.0685	0.0471
	3	Treatment	0.0124	0.0101	<0.0001	0.0007	<0.0001	0.2153	0.0217
	3	Month	0.1075	0.0921	<0.0001	0.0333	<0.0001	0.0157	0.0946
Termites	1	Treatment	0.3916	0.2698	<0.0001	0.1604	<0.0001	0.2425	0.4123
	2	Month	0.0009	0.0601	<0.0001	0.0035	<0.0001	0.0169	0.0005
	3	Treatment	0.6387	0.6287	<0.0001	0.1603	<0.0001	0.8343	0.0006
	3	Month	0.0008	0.0005	<0.0001	0.0037	<0.0001	<0.0001	<0.0001

The non-nested models are ordinary least square (OLS) regression, linear mixed model (LMM), standard Poisson, Poisson corrected for overdispersion (PCO) and negative binomial distribution (NBD). Under each of these are three models: (1) models with the intercept and main effect of treatment, (2) models with the intercept and main effect of time and (3) model with the intercept and the main effects of both treatment and time.

[a] ZIP and ZINB models produce *P* of covariates for both the zero inflated and count parts. Here *P* values of only the count part of the model are presented.

log-normal regression model, and the logarithmic transformation did not achieve the desired result (Table 1). The effectiveness of transformations to stabilize the variance in count data decreases with increase in the number of zeros. Especially, the logarithmic transformation is not applicable in such cases (Yamamura, 1999). Researchers often transform the data or use non-parametric tests to analyze count data. However, these procedures have their own limitations and they may have lower power. Until recently, non-parametric tests could be used only in one-way ANOVA. Non-parametric methods for multi-way ANOVA have become available after Brunner and Puri's (2001) work that laid the theoretical foundations for analyzing data originating in factorial designs.

Most of the data sets had more zeros than the Poisson or NBD distribution models can accommodate. However, the NBD with covariate information (e.g. treatment, time, site) provided a better description of most soil animal counts compared to the Poisson, ZIP or ZINB. Results of this study agree with the growing body of literature (Welsh et al., 1996; Martin et al., 2005; Warton, 2005; Sileshi, 2006) demonstrating that excess zeros are practical phenomena in count data. Zero inflation and overdispersion may be caused by patchiness of the environment, inherent heterogeneity of the soil animal concerned or imperfect detection of the animals (Martin et al., 2005; Warton, 2005; Sileshi, 2006). For example, the high frequency of zeros (41–91% of counts) observed in animal counts under agroforestry compared to the miombo soils (17–71%) may be due to the patchiness of agroforestry plots. Data may also be overdispersed when experimental conditions are not perfectly under control and thus the unknown parameters (μ_i) vary not only with measured covariates (such as land use

Table 4. The negative log-likelihoods (−2ll), the dispersion parameters (k), Byesian information criteria (BIC), second-order Akaike information criterion (AIC$_c$) and Akaike weights (AIC$_w$) of the range of ZIP models for earthworm counts from the miombo woodland and NBD models for earthworm counts in agroforestry (NBD)

Nested models	−2ll	Dispersion parameter (k)	BIC	AIC$_c$	AIC$_w$
Miombo					
Null (intercept only)	212	–	221	216	0
Treatment	196	–	214	204	0
Month	185	–	203	194	0
Month+treatment	159	–	186	172	1.0
Agroforestry					
Null (intercept only)	700	3.92	712	705	0
Treatment	700	3.90	718	706	0
Month+treatment	697	3.78	721	705	0
Month	698	3.80	715	704	0
Site+month+treatment	672	2.86	702	682	0.14
Site+month	673	2.88	697	681	0.24
Site+treatment	672	2.93	697	681	0.24
Site	674	2.94	692	680	0.38

A total of 88 and 356 sampling units (n) were used for the analysis.

Table 5. The 95% confidence intervals of mean densities estimated using various distribution assumptions for counts of soil invertebrates in the miombo woodland

Invertebrates	Treatment	OLS	LMM	Poisson	PCO	ZIP	NBD	ZINB
Earthworms	Burnt	0.0–0.5	−0.3–0.7	0.1–0.4	0.1–0.6	0.1–0.4	0.1–0.5	0.1–0.6
	Unburnt	0.8–2.1	0.9–1.9	1.1–1.8	1.0–2.0	0.8–2.3	0.8–2.5	0.8–2.5
Beetles	Burnt	0.3–3.7	1.0–3.0	1.6–2.5	1.4–2.9	1.4–2.7	1.5–2.7	1.4–2.8
	Unburnt	−0.9–10.1	3.6–5.6	1.5–5.3	3.7–5.8	3.7–5.7	3.6–5.9	3.6–5.9
Centipedes	Burnt	0.3–1.7	0.5–1.6	0.8–1.4	0.6–1.7	0.7–1.6	0.6–1.8	0.7–1.8
	Unburnt	0.5–1.5	0.5–1.6	0.8–1.4	0.6–1.6	0.5–1.6	0.6–1.8	0.6–1.8
Millipedes	Burnt	0.6–2.0	0.3–2.3	1.0–1.7	0.7–2.2	0.6–2.3	0.8–2.0	0.8–1.8
	Unburnt	1.3–3.7	1.5–3.4	2.1–3.0	1.7–3.5	1.5–3.8	1.6–3.8	1.5–3.6
Ants	Burnt	2.4–3.6	−1.1–7.1	2.5–3.6	1.6–5.9	2.7–5.2	1.8–5.3	1.8–5.2
	Unburnt	8.7–11.1	5.9–13.9	9.0–10.9	6.9–14.3	6.5–15.5	5.9–16.6	6.0–16.0
Termites	Burnt	5.6–18.3	6.5–17.5	10.9–13.1	8.0–17.8	8.9–19.8	7.0–20.3	7.3–21.0
	Unburnt	2.9–12.3	2.2–13.1	6.9–8.5	4.7–12.6	4.5–12.5	4.5–12.9	4.5–12.9

or treatment) but with latent and uncontrolled factors. This is often the case in observational studies such as the one conducted in the miombo woodland.

Another factor contributing to the excess-zero counts is the multivariate nature of the data (Warton, 2005). Such data are multivariate in the sense that they are separately recorded for many taxa from the same soil monolith. Sampling of many taxa simultaneously is not usually limited to locations where all taxa might occur, hence multivariate abundance data are naturally expected to contain more frequent zeros (Sileshi, 2006; Warton, 2005). Imperfect detection of soil animals could also lead to excess-zero counts. For example, a variable proportion of individuals ranging from 10% to 100% are found during soil processing, depending on the size, color and mobility of the animals (Lavelle et al., 2003).

At first guess, one might suppose the distributions of soil animals to follow the Poisson assumption. It is also tempting to believe that the Poisson model is better considering the significance of effects (Table 3). However, the ϕ values indicated poor fit of most data to the Poisson. This may be due to the temporal and spatial (treatment-specific) changes inducing heterogeneity that could not be adequately accounted by the Poisson model

(Sileshi and Mafongoya, 2007). The other fundamental problem is that the Poisson distribution is parameterized in terms of a single scalar (λ) so that all moments of y are a function of λ. In many applications a Poisson density predicts the probability of a zero count to be considerably less than is actually observed in a sample (i.e. excess-zero problem). That is why the ZIP model was superior to the standard Poisson model. A second and more obvious deficiency of the Poisson is that for count data the variance usually exceeds the mean (overdispersion), while the Poisson implies equality of the variance and the mean (Cameron and Trivedi, 1998). Poisson standard errors tend to be deflated in the presence of overdispersion and hence confidence intervals will be narrow (Table 5). Therefore, it is important to control overdispersion because large overdispersion can lead to grossly inflated statistics (Table 3) and deflated standard errors in the usual ML output (Table 5).

The NBD, ZIP and ZINB allowed for more zero counts and overdispersion than can be described by the Poisson. This was particularly so when covariate effects were included in the model (Figures 1 and 2). This agrees with Warton (2005) conclusion that the NBD fits multivariate counts very well because the high frequency of zeros can be well described by the systematic component of the model. Considering the significant deviation of the counts from the assumptions of normality and homogeneity of variance, the log-normal regression model was inappropriate. Therefore, it is concluded that for the comparison among habitat types, land-use categories or treatments, the NBD performs better than the log-normal and Poisson models. Unlike the log-normal model, the NBD does not assume homogeneous variances but actually accommodates spatial variances.

Some soil animal taxa in a community are rare, and counts of such taxa may contain more zeros than the Poisson and NBD models can accommodate. Yet such taxa will frequently be of ecological, conservation, or management interest. Recently, two-part conditional models have been used for handling zero-inflated data (Cunningham and Lindenmayer, 2005). These models are generally known as hurdle models. Future studies may explore the possibility of applying such methods for modelling count data of rare species. However, a model which is readily fitted and simple to interpret should be advocated.

In conclusion, routine application of non-parametric tests and log-normal regression models for analysis of soil animal count data with many zeros should be discouraged. Instead a statistical model appropriate for the observed data should be selected using objective criteria so that optimal inferences can be drawn about habitat, land-use or treatment effects. Information criteria such as AIC and BIC allow one to compare non-nested as well as subset models (Tables 2 and 5) by taking into account model uncertainty. This method performs better than the traditional null hypothesis test (e.g. Table 3) for interpreting effects within a specified model (Dayton, 2003).

Acknowledgments

I would like to thank the anonymous reviewers for rigorously reviewing this manuscript, and their useful comments that substantially improved this paper. Financial support for the studies came from the Canadian International Development Agency (CIDA) and Swedish International Development Agency (Sida) through the World Agroforestry Centre (ICRAF).

Appendix 1A. SAS codes used for conducting homogeneity of variance tests and ANOVA

```
Proc  glm data = miombo;
Class sample treat Month;
Model tearthw = Month/ss3;
Means Month/HOVTEST;/*HOVTES and one-way ANOVA for month*/
Run;
Proc  glm data = miombo;
Class sample treat Month;
Model tearthw = treat/ss3;
Means treat/HOVTEST;/*HOVTEST and one-way ANOVA for treatment*/
Run;
Proc  glm data = miombo;
```

```
Class sample treat Month;
Model tearthw = treat Month/ss3;/*Two-way ANOVA*/
Run;
```

Appendix 1B. SAS codes for fitting linear mixed models

```
Proc   mixed Method = ML data = miombo;
Class sample treat Month;
Model tearthw = ; /*Null model with no variables*/
Random sample;
Run;
Proc   mixed Method = ML data = miombo;
Class sample treat Month;
Model tearthw = treat;/*Treatment main effect model*/
Random sample;
Run;
Proc   mixed Method = ML data = miombo;
Class sample treat Month;
Model tearthw = month;/*Month main effect model*/
Random sample;
Run;
Proc   mixed Method = ML data = miombo;
Class sample treat Month;
Model tearthw = treat Month; /*Model with both main effects*/
Random sample;
Run;
```

Appendix 1C. SAS codes for fitting the standard Poisson model using the GENMOD procedure

```
Proc   genmod data = miombo; /*Null model using GENMOD*/
Class sample Month treat;
Model earthwo = /dist = poisson link = log;
Run;
Proc   genmod data = miombo;
Class sample Month treat;
Model earthwo = Month/dist = poisson link = log type3;/*Main effect of month*/
Run;
Proc   genmod data = miombo;
Class sample Month treat;
Model earthwo = treat/dist = poisson link = log type3;/*Treatment main effect*/
Run;
Proc   genmod data = miombo; /*Full model GENMOD*/
Class sample Month treat;
Model earthwo = Month treat/dist = poisson link = log type3;
Run;
```

Appendix 1D. SAS codes for fitting the PCO using the GENMOD procedure

```
Proc   genmod data = miombo;
Class sample Month treat;
Model earthwo = /dist = poisson link = log scale = D;/*Null model using GENMOD*/
```

```
Run;
Proc   genmod data = miombo;
Class  sample Month treat;
Model  earthwo = Month/dist = poisson link = log scale = D type3;/*Month effect*/
Run;
Proc   genmod data = miombo;
Class  sample Month treat;
Model  earthwo = treat/dist = poisson link = log scale = D type3;/*Treatment effect*/
Run;
Proc   genmod data = miombo;
Class  sample Month treat;
Model  earthwo = Month treat/dist = poisson link = log scale = D type3;/*All main ef-
       fects*/
Run;
```

Appendix 1E. SAS codes for fitting the NBD using the GENMOD procedure

```
Proc   genmod data = miombo; /*Null model using GENMOD*/
Class  sample Month treat;
Model  earthwo = /DIST = NB link = log;
Run;
Proc   genmod data = miombo;
Class  sample Month treat;
Model  earthwo = Month/dist = NB link = log type3;/*Month effect*/
Run;
Proc   genmod data = miombo;
Class  sample Month treat;
Model  earthwo = treat/dist = NB link = log type3;/*Treatment effect*/
Run;
Proc   genmod data = miombo; /*Full model GENMOD*/
Class  sample Month treat;
Model  earthwo = Month treat/dist = NB link = log type3;
Run;
```

Appendix 1F. SAS codes for fitting the standard Poisson model using the NLMIXED procedure

```
Proc   nlmixed data = miofaun; /*Null model using NLMIXED*/
parms  a0 = 0;
       eta = a0;
       lambda = exp(eta);
model  earthwo~poisson(lambda);
run;
proc   nlmixed data = miofaun; /*Full model NLMIXED*/
parms  a0 = 0 a1 = 0 a2 = 0;
       eta = a0+a1*Month+a2*treat;
       lambda = exp(eta);
model  earthwo~poisson(lambda);
run;
```

Appendix 1G. SAS codes for fitting the NBD using the NLMIXED procedure

```
PROC   NLMIXED DATA = miofaun; /*Null model/
PARMS  b0 = 0;
       eta = b0;
```

```
      mean = EXP(eta);
      loglike = earthwo*LOG(k*mean)-(earthwo+(1/k))*LOG(1+k*mean)+LGAMMA(earthwo+
      (1/k))-LGAMMA(1/k)-LGAMMA(earthwo+1);
MODEL earthwo~general(loglike);
      ESTIMATE ‘mean’ exp(b0);
run;
proc  nlmixed data = miofaun; /*Full model/
parameters a0 = 0 a1 = 0 a2 = 0;
      linpinfl = a0+a1*Month+a2*treat;
      lambda = exp(linpinfl);
      II = lgamma(earthwo+(1/k))-lgamma(earthwo+1)-lgamma(1/k) +earthwo*log(k*lambda)-
      (earthwo+(1/k))*log(1+k*lambda);
model earthwo~general(II);
run;
```

Appendix 1H. SAS codes for fitting the ZIP model

```
Proc  nlmixed data = miofaun; /*Null model*/
parameters a0 = 0 b0 = 0;
      linpinfl = a0;
      bpart = b0;
      lambda = exp(bpart);
      infprob = 1/(1+exp(linpinfl));
if    earthwo = 0 then ll = log(infprob+(1-infprob)*exp(-lambda));
else  ll = log((1-infprob))+earthwo*log(lambda)-lgamma(earthwo+1)-lambda;
model earthwo~general(ll);
run;
proc  nlmixed data = miofaun; /*Full model*/
parameters a0 = 0 a1 = 0 a2 = 0 b0 = 0 b1 = 0 b2 = 0;
      linpinfl = a0+a1*Month+a2*treat;
      bpart = b0+b1*Month+b2*treat;
      lambda = exp(bpart);
      infprob = 1/(1+exp(linpinfl));
if    earthwo = 0 then ll = log(infprob+(1-infprob)*exp(-lambda));
else  ll = log((1-infprob))+earthwo*log(lambda)-lgamma(earthwo+1)-lambda;
model earthwo~general(ll);
ESTIMATE ‘infprob’ 1/(1+exp(linpinfl));
run;
```

Appendix 1I. SAS codes for fitting the ZINB

```
PROC  NLMIXED data = miofaun; /*Null model*/
PARMS a0 = 0 b0 = 0;
      linpinfl = a0;
      infprob = 1/(1+exp(linpinfl));
      eta_nb = b0;
      lambda = exp(eta_nb);
      p0 = infprob+(1-infprob)*exp(-(earthwo+(1/k))*log(1+k*lambda));
      p_else = (1-infprob)*exp(lgamma(earthwo+(1/k))-lgamma(earthwo+1)-lgamma(1/
      k)+earthwo*log(k*lambda)-(earthwo+(1/k))*log(1+k*lambda));
if    earthwo = 0 then loglike = log(p0);
else  loglike = log(p_else);
model earthwo~general(loglike);
run;
```

```
proc   nlmixed data = miofaun; /*Full model*/
parms  a0 = 0 a1 = 0 a2 = 0 b0 = 0 b1 = 0 b2 = 0;
       linpinfl = a0+a1*Month+a2*treat;
       infprob = 1/(1+exp(linpinfl));
       eta_nb = b0+b1*Month+b2*treat;
       lambda = exp(eta_nb);
       p0 = infprob+(1-infprob)*exp(-(earthwo+(1/k))*log(1+k*lambda));
       p_else = (1-infprob)*exp(lgamma(earthwo+(1/k))-lgamma(earthwo+1)-lgamma
       (1/k)+earthwo*log(k*lambda)-(earthwo+(1/k))*log(1+k*lambda));
if     earthwo = 0 then loglike = log(p0); else loglike = log(p_else);
model  earthwo~general(loglike);
ESTIMATE ‘infprob’ 1/(1+exp(linpinfl));
run;
```

References

Anderson, J.M., Ingram, J.S.I., 1993. Tropical Soil Biology and Fertility. A Handbook of Methods, second ed. CAB International, Wallingford, 221pp.

Atkinson, A.C., 1981. Likelihood ratios, posterior odds and information criteria. J. Econometrics 16, 15–20.

Brunner, E., Puri, M.L., 2001. Nonparametric methods in factorial designs. Stat. Pap. 42, 1–52.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, New York.

Cox, D.R., 1983. Some remarks on overdispersion. Biometrika 70, 269–274.

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

Davis, P.M., 1994. Statistics for describing populations. In: Pedigo, L.P., Buntin, G.D. (Eds.), Handbook of Sampling Methods for Arthropods in Agriculture. CRS Press Inc., pp. 33–54.

Dayton, C.M., 2003. Model comparison using information measures. J. Mod. Appl. Stat. Method 2, 281–292.

Golden, R.M., 2000. Statistical tests for comparing possibly misspecified and non-nested models. J. Math. Psychol. 44, 153–170.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Jabin, M., Mohr, D., Kappes, H., Topp, W., 2004. Influence of deadwood on density of soil macro-arthropods in a managed oak–beech forest. For. Ecol. Manage. 194, 61–69.

Johnson, N.I., Kotz, S., 1969. Discrete Distributions. Houghton Mifflin Company, Boston.

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

Jones, D.T., Verkerk, R.H.J., Eggleton, P., 2005. Methods for sampling termites. In: Leather, S. (Ed.), Insect Sampling in Forest Ecosystems. Blackwell Publishing, Oxford, UK, pp. 221–253.

Kilpeläilen, J., Punttila, P., Sundström, L., Niemelä, P., Finér, L., 2005. Forest stand structure, site type and distribution of ant mounds in boreal forest in Finland in the 1950s. Ann. Zool. Fennici. 42, 243–258.

Kuha, J., 2004. AIC and BIC: comparison of assumptions and performance. Sociol. Method Res. 33, 188–229.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to random defects in manufacturing. Technometrics 34, 1–14.

Lavelle, P., Senapati, B., Barros, E., 2003. Soil macrofauna. In: Schroth, G., Sinclair, F.L. (Eds.), Trees, Crops and Soil Fertility: Concepts and Research Methods. CAB International, Wallingford, pp. 303–323.

Lawless, J.F., 1987. Negative binomial and mixed Poisson regression. Can. J. Stat. 15, 209–225.

Mackenzie, D.I., 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.

McCulloch, C.E., Searle, S.R., 2001. Generalized, Linear and Mixed Models. Wiley, New York.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman & Hall, Longo, 511pp.

Posada, D., Buckley, T.R., 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53, 793–808.

SAS Institute Inc, 2003. SAS/STAT, Release 9.1. SAS Institute Inc., Cary, NC.

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., Mafongoya, P.L., 2006a. The short-term impact of forest fire on soil invertebrates in the miombo. Bidiver. Conserv. 15, 3153–3160.

Sileshi, G., Mafongoya, P.L., 2006b. Variation in macrofaunal communities under contrasting land use systems in eastern Zambia. Appl. Soil Ecol. 31, 49–60.

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.

Smith, A.F., Spiegelhalter, D.J., 1980. Bayes factors and choice of criteria for linear models. J. R. Stat. Soc. B 42, 213–220.

Sober, E., 2004. The contest between parsimony and likelihood. Syst. Biol. 53, 644–653.

Stone, M., 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J. R. Stat. Soc. B 39, 44–47.

Susilo, F.X., Neutel, A.M., van Noordwijk, M., Hairiah, K., Brown, G., Swift, M., 2004. Soil biodiversity and food webs. In: van Noordwijk, M., Cadisch, G., Ong, C.K. (Eds.), Below-ground Interactions in Tropical Agroecosystems. CAB International, pp. 285–307.

Swift, M., Bignell, D., 2001. Standard Methods for Assessment of Soil Biodiversity and Land Use Practice – Lecture Note 6b. International Centre for Research in Agroforestry, Bogor, Indonesia.

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.

Yamamura, K., 1999. Transformation using (x+0.5) to stabilize the variance of populations. Res. Pop. Ecol. 41, 229–234.