Bulletin of Entomological Research (2006) **96**, 479–488 DOI: 10.1079/BER2006449

Selecting the right statistical model for analysis of insect count data by using information theoretic measures

G. Sileshi*

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

Abstract

Researchers and regulatory agencies often make statistical inferences from insect count data using modelling approaches that assume homogeneous variance. Such models do not allow for formal appraisal of variability which in its different forms is the subject of interest in ecology. Therefore, the objectives of this paper were to (i) compare models suitable for handling variance heterogeneity and (ii) select optimal models to ensure valid statistical inferences from insect count data. The log-normal, standard Poisson, Poisson corrected for overdispersion, zero-inflated Poisson, the negative binomial distribution and zero-inflated negative binomial models were compared using six count datasets on foliage-dwelling insects and five families of soil-dwelling insects. Akaike's and Schwarz Bayesian information criteria were used for comparing the various models. Over 50% of the counts were zeros even in locally abundant species such as *Ootheca bennigseni* Weise, *Mesoplatys ochroptera* Stål and *Diaecoderus* spp. The Poisson model after correction for overdispersion and the standard negative binomial distribution model provided better description of the probability distribution of seven out of the 11 insects than the log-normal, standard Poisson, zero-inflated Poisson or zero-inflated negative binomial models. It is concluded that excess zeros and variance heterogeneity are common data phenomena in insect counts. If not properly modelled, these properties can invalidate the normal distribution assumptions resulting in biased estimation of ecological effects and jeopardizing the integrity of the scientific inferences. Therefore, it is recommended that statistical models appropriate for handling these data properties be selected using objective criteria to ensure efficient statistical inference.

Keywords: information criterion, model uncertainty, overdispersion, zero-inflation

Introduction

There are many kinds and levels of decisions that can be made based on insect count data. Count data are widely used by scientists and regulatory agencies to evaluate the status of ecosystems and endangered species, the impact of potentially toxic chemicals (Kennedy *et al*., 2001) and other projects including genetically modified crops (Perry *et al*., 2003). However, statistical inference from count data poses several challenges. As in most ecological count datasets (Fletcher *et al*., 2005; Martin *et al*., 2005; Warton, 2005), insect counts often exhibit two features: a substantial proportion of the values are zero, and the remainder has a skewed distribution. Count data also show heterogeneity of variances among observational groups or populations (Taylor, 1961). If the sampling variance exceeds the

*Address for correspondence: P.O. Box X389, Cross Roads, Lilongwe, Malawi
Fax: 00265 1707323
E-mail: sgwelde@yahoo.com

theoretical variance, the situation is called overdispersion (Mullahy, 1997).

A large proportion of entomological research consists of observational studies where the goal is to explain a pattern with a series of explanatory variables. To do so, entomologists have long relied on statistical models assuming that errors are independently and identically distributed normal random variables. Such models derive their validity from the randomization underlying designed experiments (Stephens *et al*., 2005). Although departures from this ideal situation are common in insect count data, in practice, researchers use these models widely even where the underlying assumption of homogeneity is guaranteed to be false. Field entomologists also rarely formally appraise variance heterogeneity in relation to covariate information. This is in sharp contrast to the importance of variability, which in its different forms, is considered as the subject of interest in ecology. Heterogeneous variances can readily be interpreted biologically since they may be related to their covariates, and their statistical significance can be tested.

If not properly modelled, the presence of excess zeros and variance heterogeneity can invalidate the distributional assumptions of the analyses resulting in biased estimation of ecological effects and jeopardizing the integrity of the scientific inferences (Mullahy, 1997; Fletcher *et al*., 2005; Martin *et al*., 2005). Distribution models used for description of count data range from single parameter models such as the Poisson to complicated models like the Pascal Type H (Wilson *et al*., 1983). However, few such as the Poisson and negative binomial distribution are used widely. Often data do not support only one model as clearly best for analysis (Chatfield, 1995; Burnham & Anderson, 2002). Therefore, there is always uncertainty about the operating model that has given rise to the observations because only a sample from the population is observed (Zucchini, 2000). This raises the issue of comparing models to assess which of the models are adequate for the data and which one could be chosen as the basis for interpretation, prediction, or other subsequent use. Currently, there are two basic approaches to model selection: the classical generalized likelihood ratio test used for comparing nested models and the new approach based on information theoretic measures. The likelihood ratio test is inherently inconsistent and favours larger models (Dayton, 2003; Johnson & Omland, 2004). Unlike likelihood ratio tests, information theoretic measures are more consistent and can be used in comparison of nested as well as non-nested models (Kuha, 2004). Information criteria penalize for the addition of parameters, and thus select a model that fits well but has a minimum number of parameters to ensure simplicity and parsimony (Johnson & Omland, 2004; Kuha, 2004; Stephens *et al*., 2005).

A variety of penalized information criteria, obtained from different theoretical starting points, have been proposed in the literature (Zucchini, 2000; Dayton, 2003; Kuha, 2004). However, the Schwarz Bayesian information criterion (BIC) (Schwarz, 1978; Wasserman, 2000) and Akaike's information criterion (AIC) (Akaike, 1973) were considered here because of their popularity (Dayton, 2003; Kuha, 2004). AIC has its foundation in Kullback–Leibler information discrepancy (Burnham & Anderson, 2002). From a Bayesian perspective the BIC is analogous to AIC in that its intent is to assess models in terms of their fit and complexity. Unlike AIC, the derivation of BIC rests on several stringent assumptions that are seldom satisfied with empirical data (Wasserman, 2000; Zucchini, 2000; Kuha, 2004). BIC is consistent (Zucchini, 2000) and better in extrapolation. The aim of BIC is to identify the model with the highest probabilities of being the true model for the data, assuming that one of the models under consideration is true. Unlike BIC, AIC is defined without reference to a 'true model'. Instead, AIC uses expected prediction of future data as the key criterion of the adequacy of a model. Although models and statistical software suitable for analysis of count data exist, researchers still rely heavily on modelling approaches that assume homogeneous variance. Therefore, the objectives of this paper were to (i) compare models suitable for handling variance heterogeneity and (ii) select optimal models to ensure valid statistical inferences from insect count data.

Materials and methods

Sources of data

Six datasets collected by the author as part of various studies in eastern Zambia were re-analysed using models with different assumptions about mean-variance relationships. Detailed descriptions of the study site and management of experiments will be found in earlier reports by the author (Sileshi *et al*., 2001, 2002, 2006; Sileshi & Mafongoya, 2003, 2006a,b). The first dataset consisted of counts of *Ootheca bennigseni* Weise (Chrysomelidae: Coleoptera), a serious pest of many legumes in southern Africa (Sileshi & Kenis, 2003; Sileshi *et al*., 2006). Beetles were monitored on beans and cowpeas in experimental fields and two nearby private farmers' fields at Msekera in February 2003. One farm had only cowpeas and the numbers of beetles were recorded on 60 randomly selected cowpea plants. The other farm had both bean and cowpea fields. Here beetle counts were recorded on 60 plants each of bean and cowpea. On the research farm beetle counts were recorded on 120 plants each of bean and cowpea.

Datasets 2 and 3 consisted of counts of *Diaecoderus* spp. (Curculionidae: Coleoptera) that attack maize and the leguminous species *Sesbania sesban* (L.) Merrill, pigeon pea (*Cajanus cajan* (L.) Millsp.), *Tephrosia vogelii* Hook f. and *Crotalaria* spp. used in agroforestry (Sileshi & Mafongoya, 2003). Snout beetles were monitored during the 2000 rainy season in replicated trials consisting of fallows of these leguminous species. The beetles were counted on ten randomly selected 3–5-month-old plants. The beetles were also monitored on maize planted after clearing the legume fallows. Adult beetles were counted in February 2002 and 2003 on ten randomly selected maize plants in agroforestry practices (four pure-species and six mixed-species fallows), a traditional mixed vegetation fallow and monoculture maize grown with and without fertilizer. The treatments were replicated four times and arranged in a randomized complete blocks design (Sileshi & Mafongoya, 2003).

Dataset 4 consisted of counts of adult *Mesoplatys ochroptera* Stål (Chrysomelidae: Coleoptera). Counts were obtained using two-stage stratified random sampling conducted fortnightly from November 1997 to March 1998 in four fields of one-year-old *Sesbania sesban* fallows at Msekera (Sileshi *et al*., 2002). The fields were divided into two positions, central and peripheral, each consisting of 50 trees. Fifteen trees were randomly selected from each position, and the foliage canopy of each tree was divided horizontally into the lower, middle and upper strata based on the number of nodes. Two shoots were randomly selected from each stratum and the number of adults, egg masses and larvae were counted on each shoot (Sileshi *et al*., 2002).

Datasets 5 and 6 consisted of counts of *M. ochroptera* egg masses per plant and the predatory bug *Deraeocoris ostentans* Stål (Miridae: Heteroptera) from studies described in detail by Sileshi *et al*. (2001). Counts of both egg masses and *D. ostentans* that prey on the eggs were obtained by sampling less than one-year old plants at Msekera Research Station.

Dataset 7 consisted of a multivariate count dataset on soil dwelling Carabidae, Staphylinidae, Curculionidae, Tenebrionidae and Scarabaeidae in the miombo woodland and agroforestry systems at Msekera. The study areas, treatment, experimental design and management of the experiments have been described in detail by Sileshi & Mafongoya (2006a,b). 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.

Comparing frequency distributions

Information on the observed and expected number of individuals occurring within a series of sampling units was summarized in frequency distributions. The expected frequencies were computed assuming the Poisson and negative binomial distribution (NBD) models. Using the GENMOD procedure of SAS (SAS Institute, 2003) a maximum likelihood estimate of the dispersion parameter (k) of the NBD was obtained with and without covariate information. The expected probabilities were then calculated by substituting the sample mean for μ and k (with and without covariates separately) into the probability functions of the Poisson and NBD (equations 4 and 18 in Davis, 1994). Histograms were generated using the observed and expected frequencies for the Poisson and NBD models, and the percentage of excess zeros were computed relative to the expected frequency of zeros in each model.

Selection of statistical models

The first modelling approach considered here applied ANOVA to transformed insect counts. The probabilistic model assumed the underlying errors of the transformed count data are all uncorrelated with homogeneous variance, and hence followed an approximate log-normal distribution (Perry *et al*., 2003). The count data were transformed to natural logarithms, i.e. $\ln(y+1)$. However, the assumption of equality of variance in the log-transformed data was explicitly tested using Levene's tests via the GLM procedure of the SAS system. Normality was tested using the UNIVARIATE procedure of SAS. Standard and linear mixed model (LMM) ANOVA were then done on the transformed values using the ANOVA and MIXED procedures of the SAS system, respectively. The LMM was chosen because it extends the ANOVA model by providing a more flexible specification of the covariance matrix of the error, and allows for both correlation and heterogeneous variances via a restricted maximum likelihood (REML) methodology.

The second approach involved explicitly modelling the distribution of counts assuming that the variance (σ^2) is proportional to the mean (μ), say $\sigma^2 = \phi E(y) = \phi\mu$ where ϕ is a dispersion parameter and $E(y)$ is the expectation of counts. The variance equals the mean (i.e. Poisson assumption) when $\phi = 1$, while $\phi > 1$ indicates overdispersion. The Poisson and negative binomial distribution models were chosen to allow for many zero values of y and for the dependence of variance upon mean abundance for count data, which is often expressed through Taylor's power law (Taylor 1961; Perry *et al*., 2003). Such relationships between the variance and mean have already been demonstrated for *M. ochroptera* and *O. bennigseni* (Sileshi *et al*., 2002, 2006; Sileshi & Kenis, 2003) and *Diaecoderus* sp. (Sileshi & Mafongoya, 2003). However, this relationship is not known for the five beetle families. Therefore, the variance-mean relationship for these beetles was examined using Taylor's power law.

The Poisson distribution was considered because it arises under the assumption that insects are distributed randomly in space and the variance equals the mean. However, insect count data often exhibit overdispersion, with a variance larger than the mean (Taylor, 1961). A reasonable criterion for detecting overdispersion is that the deviance should be at least twice the number of degrees of freedom, but the actual presence of overdispersion should then be checked by some appropriate modelling procedure (Lindsey, 1999). Therefore, three modelling approaches were considered when the Poisson model indicated overdispersion. The first method involved introducing covariates thought to influence abundance using a generalized linear regression model (GLM) with a logarithmic link (McCullagh & Nelder, 1989). The second method based on a quasi-likelihood approach accounted for overdispersion via introduction of a dispersion parameter (ϕ) into the relationship between the variance and mean as $\sigma^2 = \phi\mu$. The dispersion parameter was estimated as a ratio of the deviance to its associated degrees of freedom (McCullagh & Nelder, 1989). The third method involved fitting a zero-inflated Poisson (ZIP) to the count datasets.

To account for overdispersion relative to the Poisson distribution, the NBD model was considered. One important characteristic of the NBD is that it naturally accounts for overdispersion because its variance is always greater than the variance of a Poisson distribution with the same mean. The NBD can be derived from the Poisson when the mean parameter is not identical for all members of the population, but itself is distributed with gamma distribution. Specifically, if λ has a gamma distribution then y has a negative binomial distribution with mean parameters μ and dispersion parameter k (White & Bennetts, 1996). A zero-inflated negative binomial (ZINB) model was also fitted to the data.

Under both the Poisson and NBD models, the insect counts varying over sampling units were assumed to depend on a vector of explanatory variables (X_i, i.e. time, treatment) according to the log-linear function:

$$Log(\mu) = a + b_1\ X_1 + b_2 X_2 + \cdots + b_n\ X_n \qquad (1)$$

where a is the intercept and b_i is a parameter to be estimated for the i^{th} covariate. Parameters of the standard Poisson and NBD were estimated using the GENMOD procedure, while ZIP and ZINB were fitted using the NLMIXED procedures of SAS. The GENMOD procedure fits GLMs that allow the mean to depend on linear predictors through a non-linear link function via maximum likelihood, and it allows the response probability distribution to be any member of the exponential family. The NLMIXED procedure

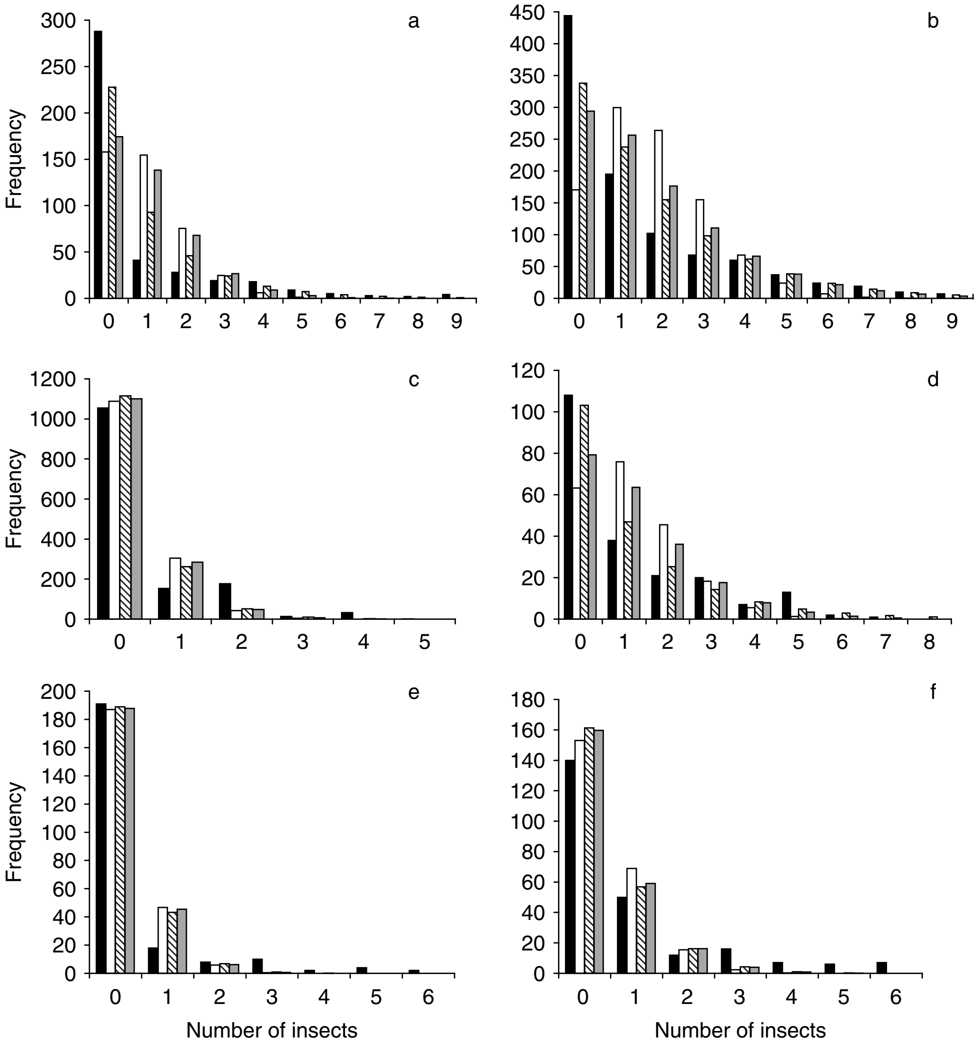

Fig. 1. Histogram of the observed and expected frequencies of *Ootheca bennigseni* on beans and cowpeas (a), *Diaecoderus* sp. on maize (b), *Mesoplatus ochroptera* adults (c) and trees (d), and egg masses (e), and the predatory bug *Deraeocoris ostentans* (f) under the Poisson and negative binomial model with (NBDcov) and without covariate information (NBDnocov). ■, Observed; □, expected Poisson; ▧, expected NBDcov; ▒, expected NBDnocov.

fits models in which both fixed and random effects enter non-linearly. This procedure enables one to specify a conditional distribution for the data (given the random effects) having either a standard form such as the binomial, Poisson or a general distribution that one codes using SAS programming statements (SAS Institute, 2003). Goodness-of-fit and model selection were based on the AIC and BIC computed as:

$$AIC = -2ll + 2\theta \tag{2}$$

$$BIC = -2ll + \theta(\ln(n)) \tag{3}$$

where ***ll*** is the log-likelihood, θ is the number of parameters in 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 (Dayton, 2003). However, in finite samples, adjusted versions of AIC such as the second-order Akaike information criterion (AIC_c) (Hurvich & Tsai, 1989) can behave much better in this respect (Johnson & Omland, 2004). Hence, in this study the second-order Akaike information criterion (AIC_c) correcting for small sample size (Hurvich & Tsai, 1989) was used. AIC_c was computed as:

$$AIC_c = -2ll + 2\theta + \frac{2\theta(\theta+1)}{n-\theta-1} \tag{4}$$

To obtain the quasi-likelihood AIC_c ($QAIC_c$), a dispersion parameter (ϕ) was introduced into equation 3 as:

$$QAIC_c = -\left(\frac{2ll}{\phi}\right) + 2\theta + \frac{2\theta(\theta+1)}{n-\theta-1} \tag{5}$$

where ϕ is the ratio of the deviance to its degrees of freedom, *ll*, θ and *n* are defined as in equation 2. Smaller AIC_c or $QAIC_c$ values indicated a more parsimonious model (Johnson & Omland, 2004).

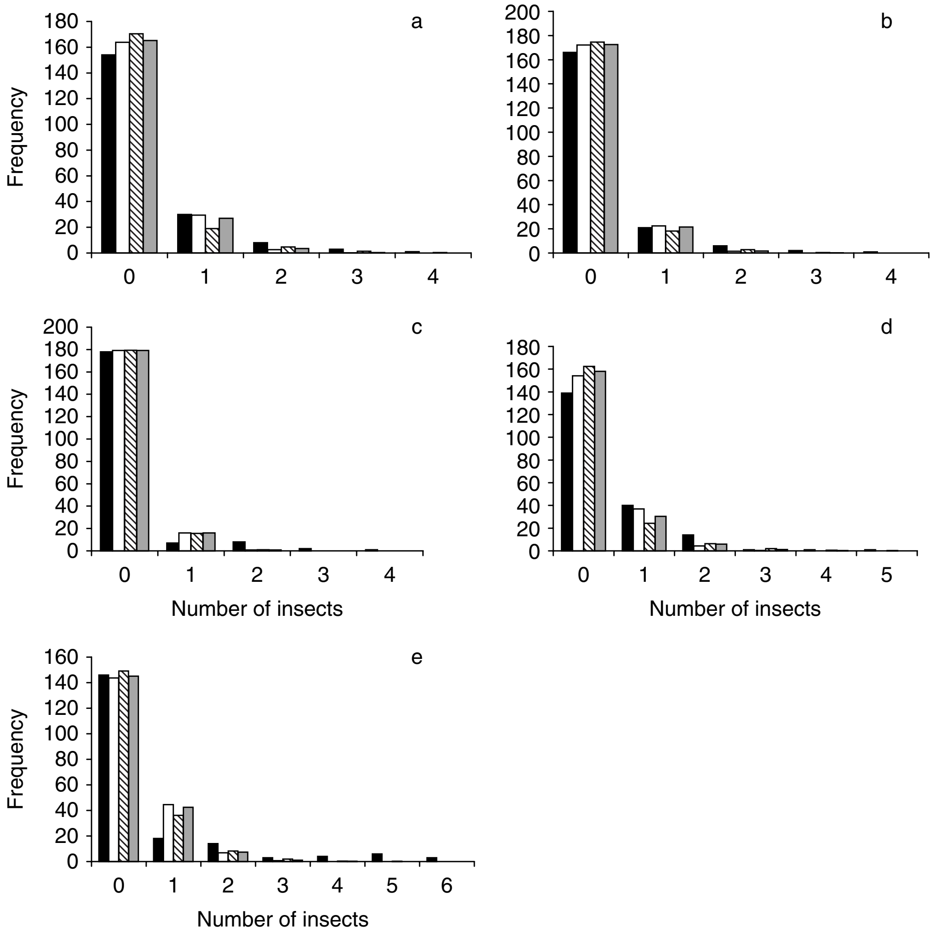

Fig. 2. Histogram of the observed and expected frequencies of (a) Carabidae, (b) Staphylinidae, (c) Curculionidae, (d) Tenebrionidae and (e) Scarabaeidae under the Poisson and negative binomial model with (NBDcov) and without covariate information (NBDnocov). ■, Observed; □, expected Poisson; ▨, expected NBCcov; ▩, expected NBDnocov.

Results

Frequency of zeros, normality and variance heterogeneity

The observed and expected frequencies of zeros in the various insect count datasets are presented in figs 1 and 2. Over 68% of the sampling units had zero counts of *O. bennigseni*. The frequency of zeros was 45 and 39% more than that expected under the Poisson and NBD models without covariate information, respectively. Inclusion of covariate information under the NBD assumption reduced the percentage of excess zeros to about 21% of that expected (fig. 1). *Diaecoderus* counts had 41–62% and 33–34% more zeros than expected under the Poisson and NBD models, respectively. However, when the NBD model was extended by inclusion of covariate information, the frequency of excess zeros dropped to 24% more than expected. Some 58–80% of the sampling units had zero counts of *M. ochroptera* adults, egg masses, and *D. ostentans*. Similarly, 70–91% of the sampling units had zero counts of the five beetle families. However, the frequency of zeros was less than that expected under the Poisson and NBD models for *M. ochroptera*, *D. ostentans* (fig. 1) and the five beetle families (fig. 2). The observed frequency distribution of all insects was more right-skewed and platykurtic than that expected under the Poisson and NBD (figs 1 and 2).

Tests of normality (Shapiro-Wilk statistic) and homogeneity of variance (Levene's test) indicated departure of the log-transformed data from normality and homogeneity of variance (table 1) across fixed effects in almost all insect species and families studied. Strong and positive relationships were found between the variance and mean abundance of the five soil-dwelling beetle families (fig. 3). Taylor's power law explained over 86% of the variation in the variance to mean relationship.

Statistical models

The Poisson model after correction for overdispersion and the standard NBD model provided better descriptions of the probability distribution of seven out of the 11 cases. The Poisson corrected for overdispersion described counts of *Diaecoderus* sp. on legumes, adults and egg mass of *M. ochroptera* and *D. ostentans* on *S. sesban* better than all other models. The standard NBD model described counts of *O. bennigseni* and *Diaecoderus* sp. on legumes, and Scarabaeidae in soil samples more parsimoniously than any

Table 1. Test for homogeneity of variance in the log-transformed [*ln*(count + 1)] data and statistical significance of fixed effects under the log-normal (standard ANOVA and linear mixed model (LMM)), Poisson and negative binomial distribution (NBD) assumptions.

Insect species/family	Sample size	Fixed effects (DF)	Levene's *F*-test	ANOVA		Poisson	NBD
				Standard	LMM		
Ootheca bennigseni	420	Farm (2)	94.9***	<0.001	<0.001	<0.001	<0.001
	420	Crop (1)	35.3***	<0.001	0.0012	<0.001	<0.001
Diaecoderus (maize)	990	Year (1)	12.9***	<0.001	<0.001	<0.001	<0.001
	990	Practice (12)	4.5***	<0.001	<0.001	<0.001	<0.001
Diaecoderus (legumes)	210	Legumes (6)	8.3***	<0.001	<0.001	<0.001	<0.001
Mesoplatys ochroptera adults	1440	Date (7)	14.2***	<0.001	<0.001	<0.001	<0.001
	1440	Position (1)	0.2ns	0.848	0.849	0.751	0.966
	1440	Stratum (2)	54.6***	<0.001	<0.001	<0.001	<0.001
M. ochroptera egg mass	240	Date (1)	19.2***	<0.001	<0.001	<0.001	<0.001
D. ostentans	240	Date (1)	10.6***	<0.001	<0.001	<0.001	<0.001
	240	Egg mass (1)	1.3ns	0.010	0.813	0.774	0.876
Carabidae	196	Land-use (6)	3.8***	0.008	0.006	<0.001	0.002
	196	Month (2)	7.4***	0.002	<0.001	<0.001	<0.001
Staphylinidae	196	Land-use (6)	2.0ns	<0.001	<0.001	<0.001	0.001
	196	Month (2)	11.7***	<0.001	<0.001	<0.001	<0.001
Curculionidae	196	Land-use (6)	7.2***	<0.001	<0.001	<0.001	<0.001
	196	Month (2)	0.2ns	0.807	0.782	<0.001	0.802
Tenebrionidae	196	Land-use (6)	3.9***	0.002	0.003	<0.001	0.001
	196	Month (2)	2.4ns	0.034	0.045	0.017	0.025
Scarabaeidae	196	Land-use (6)	11.4***	<0.001	<0.001	<0.001	<0.001
	196	Month (2)	2.6ns	0.004	0.017	<0.001	0.004

*, ** and *** indicate significance at the 5, 1 and 0.1% levels; ns, not significant.

other model (table 2). The log-normal model was best only for description of carabaid and staphylinid count datasets. According to the AIC, the ZINB was the best for description of counts of Curculionidae and Tenebrionidae, while the log-normal model was preferred by the BIC (table 2). In nine out of the 11 insects, the AIC and BIC selected the same model as the best. In terms of overall performance, the standard NBD was the best, while the quasi-likelihood method and log-normal model stood second and third, respectively. According to both Akaike's and Bayesian information criteria, the standard Poisson, ZINB and ZIP stood fourth, fifth and sixth, respectively, when ranked across datasets.

Table 1 shows the exact probabilities of significance of covariate effects under the assumptions of log-normal, Poisson and NBD models. The *P* values for covariate effects differed under the assumptions of the various models. The most striking difference between the various models was in the standard errors of parameter estimates and the 95% confidence intervals (table 3). A subset of the data is presented in table 3 to demonstrate the differences between the models in parameter estimates. The lower 95% confidence limit for the log-normal was generally higher than all the other models. Correcting for overdispersion in the Poisson increased the standard errors and widened the 95% confidence intervals of population densities (table 3). Generally, the NBD had wider 95% confidence intervals compared with the log-normal and Poisson in all insect species and families (data not shown).

Discussion

For all of the insects studied, the log-transformed data significantly deviated from the assumptions of normality and homogeneity of variance as expected. From fig. 3 and earlier studies (Sileshi *et al.*, 2002, 2006; Sileshi & Kenis, 2003; Sileshi & Mafongoya, 2003), it is clear that the variance of the counts is proportional to the mean. Over 50% of the counts were zeros even in locally abundant species such as *O. bennigseni*, *M. ochroptera* and *Diaecoderus* spp. This is in agreement with the growing body of literature (Welsh *et al.*, 1996; Martin *et al.*, 2005; Warton, 2005) demonstrating that excess zeros are practical phenomena in count data. The study has also demonstrated that the excess zeros and variance heterogeneity can be accommodated by adjusting for overdispersion or inclusion of covariate information (figs 1 and 2). The presence of excess zeros and overdispersion in a data set may arise as a result of patchiness of the environment, inherent heterogeneity of the species concerned, imperfect detection of a species or may even be a sign of inadequacy of the model used (MacKenzie *et al.*, 2002; Fletcher *et al.*, 2005; Warton, 2005). Zero-inflation is often the result 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). False zeros occur when the species under study is present at the time of sampling, but the observer does not detect it because the species is cryptic or secretive (MacKenzie *et al.*, 2002). Overdispersion may also arise due to habitat heterogeneity or biological phenomena such as aggregation (Mullahy, 1997; Fletcher *et al.*, 2005). Lack of independence, arising as a result of the innate behaviour of the insects being studied, may also lead to overdispersion in counts. For example, *M. ochroptera* frequently mate throughout their adult life and a pair behaves almost as an individual (Sileshi *et al.*, 2002). Lack of independence as in *M. ochroptera* and *O. bennigseni* (Sileshi *et al.*, 2002; Sileshi & Kenis, 2003) and spatial variation induced by habitat

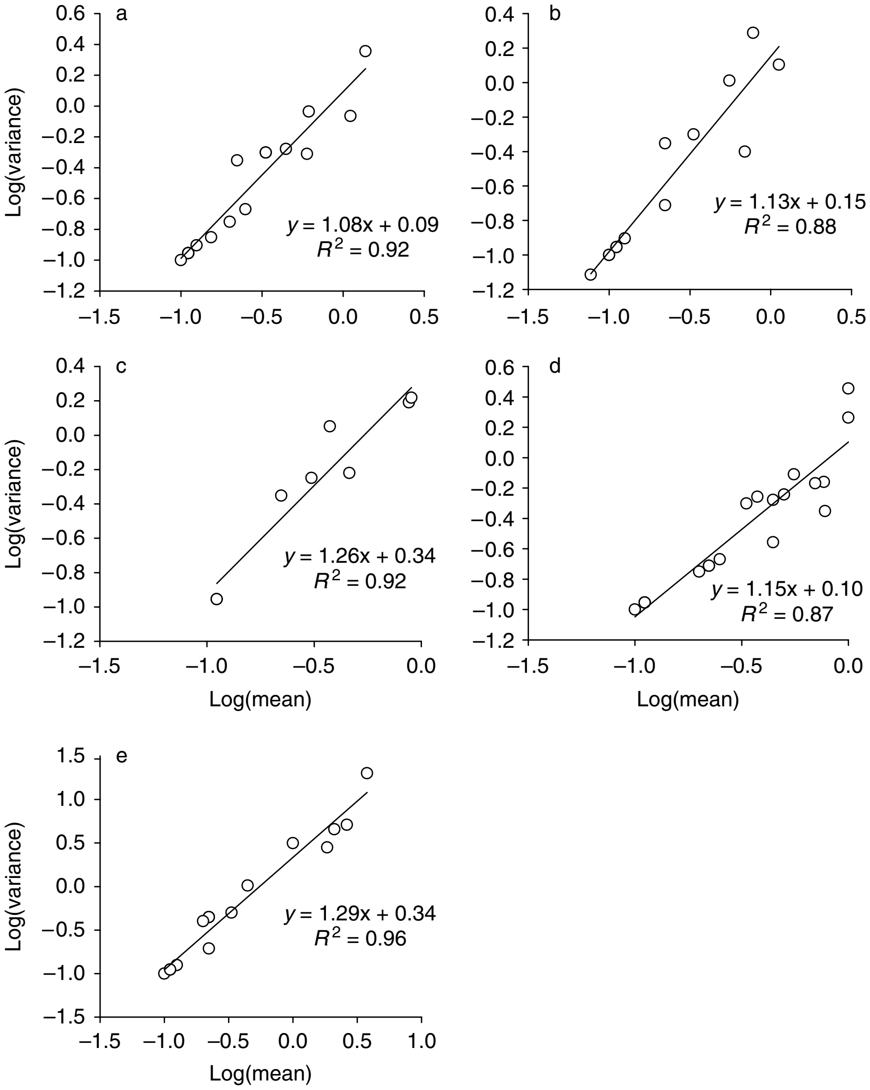

Fig. 3. Relationship between variance and mean abundance of (a) Carabidae, (b) Staphylinidae, (c) Curculionidae, (d) Tenebrionidae and (e) Scarabaeidae in the soil under miombo woodland and agroforestry species.

heterogeneity in counts of *Diaecoderus* (Sileshi & Mafongoya, 2003) could have led to the observed overdispersion. The high frequency of zeros observed in the beetle counts from soils may be due to the multivariate nature of the count (Li *et al.*, 1999; Warton, 2005). Sampling of many taxa simultaneously is not usually limited to habitats where all taxa might occur, hence multivariate abundance data are naturally expected to contain more frequent zeros (Warton, 2005).

For seven out of the 11 insects, the quasi-likelihood method and NBD model yielded better statistical inference than one based on the log-normal, standard Poisson, ZIP or ZINB. The log-normal and Poisson models had narrower 95% confidence intervals of population densities, and hence their estimates may not be consistent when counts show overdispersion. Poisson estimates are consistent when the variance is proportional (not just equal) to the mean. Thus, Poisson standard errors tend to be conservative in the presence of overdispersion. If the Poisson is the chosen model and if we are sure that the lack of fit is not due to poor specification of the systematic part of the model, the standard errors need to be corrected via the quasi-likelihood method (McCullagh & Nelder, 1989). For practical reasons, fitting the NBD may be more preferable to the log-normal ANOVA or the more complicated route of fitting zero-inflated models. The primary advantage of the NBD is that it does not assume homogeneous variances but actually modelled heterogeneity in variance reflected by the dispersion parameter (*k*). This parameter provided additional insight not explicit in the log-normal ANOVA model, contributing more information about the data being analysed (White & Bennetts, 1996). ANOVA tests whether there are differences in means. Of equal biological interest is whether two populations are distributed similarly even if the means do not differ. Ignoring overdispersion in the analysis would lead to underestimation of standard errors, and

Table 2. Second-order Akaike information criteria (AIC$_c$) and Bayesian information criterion (BIC) for comparing the global model (containing all covariates) under the log-normal (mixed ANOVA), Poisson and negative binomial model assumptions for foliar and soil-dwelling insects.

Insect species	Criterion	Mixed ANOVA	Poisson	Poisson		Negative binomial	
				Quasi-likelihood	ZIP	NBD	ZINB
Ootheca bennigseni	AICc	564.4	177.8	158.8	1111.3	**87.2**	1050.8
	BIC	568.4	184.1	165.5	1147.2	**93.9**	1074.9
Diaecoderus sp. (maize)	AICc	2062.6	915.4	417.0	3912.0	**259.4**	3532.4
	BIC	2067.5	953.4	455.8	3955.9	**298.2**	3561.7
Diaecoderus sp. (legumes)	AICc	381.1	318.1	**196.5**	650.4	279.3	639.4
	BIC	384.4	320.0	**198.5**	679.7	281.3	659.0
Mesoplatys ochroptera adult	AICc	1881.8	2043.7	**1652.0**	2926.6	1768.2	2795.2
	BIC	1887.0	2151.1	**1760.3**	2973.9	1876.5	2826.8
M. ochroptera egg mass	AICc	347.7	241.3	**174.8**	696.2	204.0	661.4
	BIC	351.1	242.3	**176.1**	726.7	205.3	681.9
D. ostentans	AICc	436.1	439.6	**228.8**	696.2	336.0	661.4
	BIC	439.6	442.0	**231.4**	726.7	338.6	681.9
Carabidae	AICc	**158.6**	271.2	342.7	294.1	273.9	287.0
	BIC	**161.8**	267.6	338.4	322.6	269.6	306.2
Staphylinidae	AICc	**113.2**	222.8	333.3	239.9	224.3	232.9
	BIC	**116.4**	219.2	329.0	268.5	220.0	252.1
Curculionidae	AICc	99.8	186.2	293.1	98.8	179.9	**96.7**
	BIC	**102.9**	182.6	288.8	127.3	175.6	115.9
Tenebrionidae	AICc	215.6	329.4	347.5	326.6	330.9	**201.0**
	BIC	**218.8**	325.8	343.2	355.1	326.6	220.3
Scarabaeidae	AICc	268.8	191.4	173.3	455.6	**155.9**	419.8
	BIC	271.9	187.8	169.0	484.1	**151.6**	439.1

Bold-faced entries in a row indicate the best model according to AICc and BIC for each species or family of insects. NBD, negative binomial distribution; ZINB, zero-inflated negative binomial; Zip, zero-inflated Poisson.

Table 3. The 95% confidence intervals of mean densities of *Ootheca bennigseni* on food legumes in three farms and *Diaecoderus* spp. on maize grown in agroforestry practices during 2002 and 2003 under various model assumptions.

Insect species	Fixed effects	Log-normal	Poisson		NBD
			Before scaling	After scaling	
O. bennigseni	Farm 1 (Mr Banda)	2.1–2.5	1.8–2.4	1.8–2.5	1.7–2.6
	Farm 2 (Mr Nyirenda)	2.3–3.0	2.2–3.0	2.2–3.1	2.0–3.5
	Farm 3 (Research station)	1.0–1.1	0.01–0.06	0.01–0.06	0.01–0.06
	Bean	1.6–1.8	0.3–0.5	0.3–0.6	0.3–0.6
	Cowpea	1.9–2.1	1.3–1.6	1.2–1.7	1.1–1.8
Diaecoderus spp.	Year 2002	2.4–2.8	2.3–2.6	2.2–2.7	2.2–2.8
	Year 2003	1.5–1.7	1.0–1.2	1.0–1.3	1.0–1.3
	Cajanus cajan (Cc)	1.4–1.9	0.9–1.3	0.8–1.6	0.8–1.5
	Crotalaria grahamiana (Cg)	1.3–1.8	0.7–1.1	0.6–1.3	0.6–1.3
	Cc + Cg	1.5–2.1	0.9–1.4	0.8–1.6	0.8–1.6
	Fertilized monoculture maize	2.0–2.6	2.0–2.6	1.8–2.9	1.7–3.1
	Unfertilized monoculture maize	1.3–1.8	0.7–1.2	0.6–1.4	0.7–1.3
	Traditional grass fallow	1.3–1.7	0.6–1.0	0.5–1.1	0.5–1.1
	Sesbania sesban (Ss)	2.0–2.7	1.7–2.3	1.5–2.6	1.4–2.7
	Tephrosia vogelii (Tv)	2.3–3.0	2.2–2.9	2.0–3.1	1.9–3.3
	Ss + Cc	2.2–2.9	2.2–2.9	2.0–3.2	1.9–3.4
	Ss + Cg	2.4–3.2	2.5–3.3	2.3–3.6	2.1–3.8
	Ss + Tv	1.9–2.6	1.6–2.3	1.5–2.6	1.4–2.7
	Tv + Cc	2.5–3.4	2.5–3.3	2.3–3.6	2.1–3.9
	Tv + Cg	1.5–2.0	0.9–1.4	0.8–1.7	0.8–1.6

consequent overstatement of significance in hypothesis testing. Comparison of *P*-values (table 1) and 95% confidence intervals of population densities (table 3) suggest that statements of statistical significance about fixed effects and parameter estimates are greatly influenced by the choice of the model. Therefore, use of inappropriate models can result in biased estimation of effects and erroneous predictions and conclusions regarding ecological processes.

It is recommended that researchers select models appropriate for handling variance heterogeneity rather than attempting to homogenize variances using transformations or resorting to non-parametric methods. Information criteria provided an objective way of determining which model among a set of models is most appropriate for the data at hand. The primary disadvantage of testing various models is that some models are computationally demanding especially for complex experimental designs. However, software and computer programs that can handle these calculations with relative ease are appearing.

Acknowledgements

The author would like to thank all those who helped in the data collection during the various studies and two anonymous reviewers for their critical comments and suggestions that substantially improved the final manuscript. Financial support for data collection came from the Canadian International Development Agency (CIDA), Swedish International Development Agency (SIDA) through the Zambezi Basin Agroforestry Project of the World Agroforestry Centre (ICRAF) and the Deutscher Akademischer Austauschdienst (DAAD) through the International Centre of Insect Physiology and Ecology (ICIPE).

References

Akaike, H. (1973) Information theory as an extension of the maximum likelihood principle. pp. 267–281 *in* Petrov, B.N. & Csaki, F. (*Eds*) *Second International Symposium on Information Theory*. Akademiai Kiado, Budapest.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach*. 2nd edn. New York, Springer-Verlag.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A* **158**, 419–466.

Davis, P.M. (1994) Statistics for describing populations. pp. 33–54 *in* Pedigo, L.P. & Buntin, G.D. (*Eds*) *Handbook of sampling methods for arthropods in agriculture*. Boca Raton, Florida, CRS Press Inc.

Dayton, C.M. (2003) Model comparison using information measures. *Journal of Modern Applied Statistical Methods* **2**, 281–292.

Fletcher, D., MacKenzie, D. & Villouta, E. (2005) Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics* **12**, 45–54.

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

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution* **19**, 101–108.

Kennedy, P.J., Conrad, K.F., Perry, J.N., Powell, D., Aegerter, J., Todd, A.D., Walters, K.F.A. & Powell, W. (2001) Comparison of two field-scale approaches for the study of effects of insecticides on polyphagous predators in cereals. *Applied Soil Ecology* **17**, 253–266.

Kuha, J. (2004) AIC and BIC: comparison of assumptions and performance. *Sociological Methods and Research* **33**, 188–229.

Li, C.S., Lu, J.C., Park, J., Kim, K., Brinkley, P.A. & Peterson, J.P. (1999) Multivariate zero-inflated Poisson models and their application. *Technometrics* **41**, 29–38.

Lindsey, J.K. (1999) On the use of corrections for overdispersion. *Journal of the Royal Statistical Society: Series C* **48**, 553–561.

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. *Ecology Letters* **8**, 1235–1246.

McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*. 2nd edn. London, Chapman and Hall.

Mullahy, J. (1997) Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics* **12**, 337–350.

Perry, J.N., Rothery, P., Clark, S.J., Heard, M.S. & Hawes, C. (2003) Design, analysis and statistical power of the farm-scale evaluation of genetically modified herbicide-tolerant crops. *Journal of Applied Ecology* **40**, 17–31.

SAS Institute (2003) SAS/STAT, Release 9.1, Cary, North Carolina, SAS Institute Inc.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Sileshi, G. & Kenis, M. (2003) Temporal and spatial distribution of *Ootheca bennigseni* Weise (Coleoptera: Chrysomelidae), a defoliator of food legumes and *Sesbania sesban* in southern Africa. pp. 68–69 *in* Stals, R. (*Ed.*) *Proceedings of the 14th Entomological Congress*, 6–9 July 2003, Pretoria, Entomological Society of Southern Africa.

Sileshi, G. & Mafongoya, P.L. (2003) Effect of rotational fallows on abundance of soil insects and weeds in maize crops in eastern Zambia. *Applied Soil Ecology* **23**, 211–222.

Sileshi, G. & Mafongoya, P.L. (2006a) Variation in macrofaunal communities under contrasting land use systems in eastern Zambia. *Applied Soil Ecology* **33**, 49–60.

Sileshi, G. & Mafongoya, P.L. (2006b) The short-term impact of forest fire on soil invertebrates in the miombo. *Biodiversity and Conservation* (in press).

Sileshi, G., Kenis, M., Ogol, C.K.P.O. & Sithanantham, S. (2001) Predators of *Mesoplatys ochroptera* Stål in sebania-planted fallows in eastern Zambia. *BioControl* **46**, 289–310.

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. *Journal of Economic Entomology* **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. *Bulletin of Entomological Research* **96**, 61–69.

Stephens, P.A., Buskirk, S.W., Hayward, G.D. & Del Rio, C.M. (2005) Information theory and hypothesis testing: a call for pluralism. *Journal of Applied Ecology* **42**, 4–12.

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

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.

Wasserman, L. (2000) Bayesian model selection and model averaging. *Journal of Mathematical Psychology* **44**, 92–107.

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. *Ecological Modelling* **88**, 297–308.

White, G.C. & Bennetts, R.E. (1996) Analysis of frequency count data using the negative binomial distribution. *Ecology* **77**, 2549–2557.

Wilson, L.T., Room, P.M. & Bourne, A.S. (1983) Dispersion of arthropods, flower buds and fruit in cotton fields: effects of population density and season on the fit of probability distributions. *Journal of the Australian Entomological Society* **22**, 129–134.

Zucchini, W. (2000) An introduction to model selection. *Journal of Mathematical Psychology* **44**, 41–61.

Appendix 1

SAS procedures used for fitting various probability distributions

```
/*The following MIXED procedure estimates the response with
covariates Farm and Crop assuming log-normal error distribution*/
Proc mixed data=Ootheca;
Class Farm Crop;
model tcount = Farm Crop; /*tcount is log10(count+1)*/
lsmeans Farm Crop/ adjust=Tukey CL;
Run;

/*The following GENMOD procedure estimates the response with
covariates Farm and Crop assuming standard Poisson error
distribution*/
Proc genmod data=Ootheca;
Class Farm Crop;
model Count = Farm  Crop/dist = poisson link = log type3; Run;

/*The following GENMOD procedure estimates the response with
covariates Farm and Crop and correcting for over-dispersion in the
Poisson distribution*/
Proc genmod data=Ootheca;
Class Farm Crop;
model Count = Farm  Crop/dist = poisson link = log dscale type3; Run;

/*The following NLMIXED procedure fits a zero-inflated Poisson
distribution*/
proc nlmixed data=Ootheca;
parameters a0=0 a1=0 a2=0 a3=0 b0=0 b1=0 b2=0;
linpinfl = a0 + a1*VAR1 + a2*VAR2 + a3*VAR1*VAR2;
infprob = 1/(1+exp(-linpinfl)); lambda = exp(b0 + b1*VAR1 +b2*VAR2);
if Count=0 then ll = log(infprob + (1-infprob)*exp(-lambda));
else ll = log((1-infprob)) + Count *log(lambda)-lgamma(Count+1)-lambda;
model Count ~ general(ll);
predict (1-infprob)*lambda out = PREDICTED_Count; run;

/*The following GENMOD procedure estimates the response with
covariates Farm and Crop assuming negative binomial error
distribution*/
Proc genmod data=Ootheca;
Class Farm Crop;
model Count = Farm  Crop/dist = nb link = log type3; Run;

/*The following NLMIXED procedure fits a zero-inflated negative
binomial distribution*/
proc nlmixed data=Ootheca;
parms a0=0 a1=0 b0=0 b1=0;
linpinfl = a0 + a1*G1;
psi = 1 / (1 + exp(linpinfl));
eta_nb = b0 + b1*G1; lambda = exp(eta_nb);
p0 = psi + (1-psi)*exp(-(Count+(1/k))*log(1+k*lambda));
p_else = (1-psi)* exp(lgamma(Count+(1/k))-lgamma(Count+1)-lgamma(1/k)+
Count*log(k*lambda)-(Count+(1/k))*log(1+k*lambda));
if Count=0 then loglike = log(p0);
else loglike = log(p_else);
model Count ~ general(loglike);
run;
```

(Accepted 5 June 2006)