ENVIRONMETRICS
Environmetrics 2005; **16**: 275–289
Published online 2 February 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.702

Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data

David I. Warton*,†

Department of Biological Sciences, Division of Environmental and Life Sciences, Macquarie University NSW 2109, Australia

SUMMARY

An important step in studying the ecology of a species is choosing a statistical model of abundance; however, there has been little general consideration of which statistical model to use. In particular, abundance data have many zeros (often 50–80 per cent of all values), and zero-inflated count distributions are often used to specifically model the high frequency of zeros in abundance data. However, in such cases it is often taken for granted that a zero-inflated model is required, and the goodness-of-fit to count distributions with and without zero inflation is not often compared for abundance data.

In this article, the goodness-of-fit was compared for several marginal models of abundance in 20 multivariate datasets (a total of 1672 variables across all datasets) from different sources. Multivariate abundance data are quite commonly collected in applied ecology, and the properties of these data may differ from abundances collected in autecological studies. Goodness-of-fit was assessed using AIC values, graphs of observed vs expected proportion of zeros in a dataset, and graphs of the sample mean–variance relationship.

The negative binomial model was the best fitting of the count distributions, without zero-inflation. The high frequency of zeros was well described by the systematic component of the model (i.e. at some places predicted abundance was high, while at others it was zero) and so it was rarely necessary to modify the random component of the model (i.e. fitting a zero-inflated distribution). A Gaussian model based on transformed abundances fitted data surprisingly well, and rescaled per cent cover was usually poorly fitted by a count distribution. In conclusion, results suggest that the high frequency of zeros commonly seen in multivariate abundance data is best considered to come from distributions where mean abundance is often very low (hence there are many zeros), as opposed to claiming that there are an unusually high number of zeros compared to common parametric distributions.

KEY WORDS: zero-inflated Poisson; Akaike's information criterion; negative binomial

1. INTRODUCTION

A fundamental method for studying the ecology of a species is to investigate what environmental variables are associated with its abundance. Regression methods are appropriate for such research,

*Correspondence to: D. I. Warton, Department of Statistics, School of Mathematics, University of New South Wales, NSW 2052 Australia.
†E-mail: David.Warton@unsw.edu.au

Received 1 December 2003
Accepted 12 July 2004

and three different methods are commonly used, when the response variable is the abundance in a taxon:

- least squares on transformed data (Attrill *et al.*, 1999, for example)
- log-linear models (Catling *et al.*, 2000, for example)
- models using zero-inflated count distributions (Welsh *et al.*, 1996).

Should choice of method depend on data properties, or is there a single method that could generally be used? These questions were addressed in this article, using multivariate versions of abundance data, referred to here as multivariate abundance data. The focus in this article is on multivariate abundance data because this type of data is frequently used in ecology, but it has not yet been established what properties these data generally have, and hence what statistical models for these data are generally appropriate.

Multivariate abundance data are multivariate in the sense that they are separately recorded for many taxa. This type of data is published in hundreds of studies every year, most of which have the aim of understanding what environmental variables are associated with the abundances in the community of taxa. Multivariate generalizations of the above methods are likely to be appropriate methods for analysis, because the types of problems of interest are usually multivariate generalizations of the problems for which least squares, log-linear models, and zero-inflated distributions are usually used. However, such methods are not currently used, although recent work (Warton and Hudson, 2004) demonstrated that test statistics based on least squares methods were at least as powerful as current methods, across 20 multivariate abundance datasets. If log-linear models or models based on zero-inflated models consistently fit multivariate abundance data more closely than least squares models, then this suggests that these methodologies would generally be more appropriate for this type of data.

Empirical comparisons of the three regression methods described above are rarely undertaken, although Welsh *et al.* (1996) did discuss their relative merits. Welsh *et al.* (1996) and others (Biondini *et al.*, 1988; Clarke and Green, 1988; Guisan and Zimmerman, 2000) argue that models based on least squares would provide a poor fit to counted abundances, particularly when counts are rare. However, compared to alternative methods, least squares is simpler to generalise to complex designs (such as models including random effects or spatial autocorrelation), and so the method remains potentially useful unless its fit to abundance data is clearly poor.

Fitting zero-inflated count distributions is of particular interest, because it is widely reported that an important property of abundance data is the high frequency of zeros (Clarke and Green, 1988, for example). Welsh *et al.* (1996) suggested a method of fitting zero-inflated count distributions that is simpler to fit and interpret than the more standard mixture model approach of Lambert (1992). The approach of Welsh *et al.* (1996) has subsequently been used to fit zero-inflated distributions to abundances in a range of different contexts (Lindenmayer *et al.*, 1999; Welsh *et al.*, 2000; Dobbie and Welsh, 2001; Pearce and Ferrier, 2001).

There is very little literature comparing the fits to abundance data of zero-inflated distributions and other count distributions. The only example I found was from a study of weevil counts on chestnuts (Desouhant *et al.*, 1998), which found that a Poisson model was clearly inadequate, there was no evidence against a negative binomial model in most cases, and no evidence against a zero-inflated Poisson model in nearly all cases. Also, the mixture model Welsh *et al.* (1996) fitted according to the approach of Lambert (1992) can be used to infer that, for their Leadbeater's possum abundances, there were significantly more zeros than expected under a Poisson or negative binomial distribution. Other studies that have fitted zero-inflated count distributions to abundance data do not present any evidence to justify this choice of model.

The properties of multivariate abundances are likely to differ from abundances collected in autecological (single species) studies. A particular property that may differ is the frequency of zeros. Sampling in an autecological study will undoubtedly be targeted at locations where the taxon of interest might or does occur. In contrast, sampling of many taxa simultaneously is not usually limited to locations where all taxa might occur, and hence multivariate abundance data might be expected to contain more frequent zeros.

This work extends Warton and Hudson (2004), where 20 multivariate abundance datasets drawn from the ecology literature were used to compare statistics testing for the effect of one or several factors. The use of these same 20 datasets in this study ensures that results can be generalized across multivariate abundance data. Given that there is little literature comparing the fit of different models to abundances, results of this study are also of interest for model choice in studies where abundance of a single taxon is sampled. However, caution should be exercised in interpreting results in this context, given the differences in properties between abundances from community and autecological studies.

In this paper, two main questions were addressed, relating to the goodness-of-fit to abundance data for different parametric models:

- Can abundance data be transformed such that least squares models fit data as well as count distributions?
- Are zero-inflated models usually required to account for frequent zeros in abundance data?

Goodness-of-fit is compared in two different ways:

- Model selection approach—Akaike's information criterion (Akaike, 1972), abbreviated AIC, was used to compare the different models ($AIC = -2L + 2q$, where L is the log-likelihood and q is the number of parameters in the model). The best-fitting models have the lowest AIC values.
- Graphical approach—for mean-variance relationships and proportion of zeros in a variable, fitted and observed values were compared graphically. If fitted values are consistently far from observed values, the model is poor.

2. METHODS

All work was carried out using Matlab version 6.5 (The Mathworks Inc., 1984–2003).

2.1. Datasets

Table 1 describes dataset properties. All of the available data were used in this study, whereas only a fraction of some datasets was used in Warton and Hudson (2004). All variables with less than two non-zero abundances were excluded—if all except one abundance are zero, then clearly there is negligible information from which different parametric models for abundance could be compared.

Most of the datasets arose from observational studies or field experiments. Three studies were conducted under controlled experimental conditions, as indicated in Table 1. Datasets were selected according to two criteria: availability, and to represent a range of different data types (in terms of type of study organism, level to which it was classified, type of abundance measure, source).

For all datasets, observations were classified into several groups, based on experimental manipulations or environmental variables measured with the observations. Models for abundance assumed that the location parameter (and zero inflation parameter, where applicable) was different for each group of observations, but that any nuisance parameters were constant for abundances in a given variable. This

Table 1. Reference datasets, and their properties. All datasets had size $N \times p$, where N is the number of observations and p is the number of variables (taxa). '#Groups' is the number of groups of replicates into which observations were categorized

Source	Organism	Observation	Abundance	N	p	#Groups
N. Andrews	Invertebrates	Plant	Count	360	17	36
A. Pik	Invertebrates	Area of litter	Count	49	24	10
Moulton (1982)	Arthropods	Soil core	Count	24	62	12
Moulton (1982)	Arthropods	Soil core	Count	16	28	8
I. Lunt[§]	Plants	Plot	Count	100	57	10
A. Pik	Invertebrates	Pitfall	Count	134	21	15
Warwick *et al.* (1990b)	Macrofauna	Sediment core	Count	24	46	6
Gray *et al.* (1990)	Macrofauna	Sediment grab	Count	39	137	4
Warwick *et al.* (1988)[‡]	Copepods	Sediment core	Count	16	44	4
Gee *et al.* (1985)[‡]	Nematodes	Sediment core	Count	12	39	3
Warwick *et al.* (1990a)	Copepods	Sediment core	Count	16	11	8
van den Brink *et al.* (1996)[‡]	Invertebrates	Sediment, water	Count	132	94	55
van der Aart and Smeenk-Enserink (1970)	Spiders	Pitfall	Count	28	12	6
B. Rice	Plants	Plot	% cover	50	147	5
J. Overton	Plants	Plot	% cover	38	139	6
B. Rice	Plants	Plot	% cover	46	147	6
B. Rice	Plants	Plot	% cover	39	293	9
Clements (1980)	Plants	Plot	% cover	48	231	16
van Dobben *et al.* (1999)	Plants	Plot	% cover	32	77	16
Pearson and Blackstock (1984)	Macrofauna		Biomass	12	46	4

[§]This dataset was collected by P. Foreman, M. Titcumb and I. Lunt, who would like to acknowledge funding from Parks Victoria, the Johnstone Centre and Charles Sturt University.
[‡]Conducted under controlled experimental conditions.

type of model is commonly used in hypothesis testing, and in the least-squares case this leads to the ANOVA model.

Some of the datasets considered did not contain counted data, but instead were per cent cover data, or in one case, biomass. In such cases each variable was rescaled so that the minimum non-zero value was 1. Such rescaled data might behave like counted data, if per cent cover and biomass are approximately constant across individuals in a taxon.

2.2. *Model estimation*

This section describes the models fitted to abundances and the methods of parameter estimation used. The vector $\mathbf{y}$ represents abundance in a given taxon, although it should be noted that each dataset contained many such $\mathbf{y}$s, for which separate parameter estimates and AIC values were obtained. The vector $\mathbf{y}$ contains N values, the ith value denoted y_i, and is a realisation of the random variables $\mathbf{Y}$.

Transformation/least squares: A simple approach to analysis is to transform data to shorten the right tail of the distribution, and use ordinary least squares methods. This method is common in practice for univariate abundances, with the most common transformations being $\mathbf{z} = \log(\mathbf{y} + 1)$ or $\mathbf{y}^{1/4}$ (Clarke and Green, 1988). This approach has the advantages that estimation is straightforward, that the properties of estimators are well understood (Miller, 1986), and that the method is familiar to practitioners.

We consider only the case where $\mathbf{z} = \log(\mathbf{y} + 1)$. A particularly useful property of this transformation is that it assumes a quadratic mean–variance relationship, as do the other models considered here. The probabilistic model underlying least squares assumes transformed abundances $\mathbf{Z}$ are Gaussian, and given the relationship between density functions of $f_Y(y) = f_Z(z)/(y+1)$, the log-likelihood of abundances is

$$L(\mu, \sigma; \mathbf{y}) = -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N}\left\{\frac{1}{2\sigma^2}(z_i - \mu_i)^2 + z_i\right\}$$

We denote this model LS. The maximum likelihood estimates are $\hat{\sigma}^2 = \sum_{i=1}^{N}(z_i - \mu_i)^2/N$ and $\hat{\mu}_i$ is the sample mean of $\mathbf{z}$ for the group to which the ith observation belongs. The maximized likelihood is

$$L(\hat{\mu}, \hat{\sigma}; \mathbf{y}) = -\frac{N}{2}\log(2\pi\hat{\sigma}^2) - \frac{N}{2} - \sum_{i=1}^{N} z_i$$

A problem with using the above expression in AIC calculations is that continuous distributions are being fitted to discrete data, but the log-likelihoods of discrete and continuous distributions are not comparable. In particular, the density of a continuous distribution can exceed one (hence L can be positive), whereas this is not the case for discrete distributions. To compare the LS model with those based on discrete distributions, we discretized the Gaussian distribution for AIC calculations:

$$L(\hat{\mu}, \hat{\sigma}; \mathbf{y}) = \sum_{i=1}^{N}\log\left\{\Phi\left[\frac{\log(y_i + 1.5) - \hat{\mu}_i}{\hat{\sigma}}\right] - \Phi\left[\frac{\log(y_i + 0.5) - \hat{\mu}_i}{\hat{\sigma}}\right]\right\}$$

where $\Phi(c)$ is the lower tail probability at c from the standard normal distribution. When $y_i = 0$, the second term in the above expression was set to 0. In most cases, the function $\Phi(c)$ was estimated using the 'normcdf' function in the Matlab toolbox Statpack 2.1 (the Mathworks Inc., 1993–1997). However, this function sets $\Phi(c)$ to one when c is large (> 8.3), which leads to computational problems when y_i is large (taking the logarithm of 0), and so in such cases $\Phi(c)$ was estimated using the first four terms of an asymptotic series for $\Phi(c)$:

$$\Phi(c) \approx 1 - \frac{1}{\sqrt{2\pi}}e^{\frac{-c^2}{2}}\left(c^{-1} - c^{-3} + 3c^{-5} - 15c^{-7}\right)$$

(from Acton, 1970, p. 16). This is a good approximation—the error (as a percentage of $\Phi(c)$) is about 0.02% when $c = 5$, and this error decreases as c increases.

Log-linear model: Counted data are traditionally analyzed using log-linear models (McCullagh and Nelder, 1989, chap. 6). Because abundance data are typically overdispersed compared to the Poisson distribution, negative binomial models for overdispersion were considered (with two forms of overdispersion, mean–variance relationship either $V(\mu) = \mu + \phi\mu^2$ or $V(\mu) = \phi\mu$).

Five different cases were considered.

P: Poisson model. The log-likelihood expression is

$$L(\mu; \mathbf{y}) = \sum_{i=1}^{N}(y_i \log\mu_i - \mu_i - \log y_i!)$$

$\hat{\mu}_i$ is the sample mean of the observations of $\mathbf{y}$ which belong to the same group as the ith observation. When $\hat{\mu}_i = 0$, the ith component of L was set to zero for P and all other count models.

$\mathbf{NB_{LR}}$: Negative binomial model, $V(\mu) = \mu + \phi\mu^2$, with a maximum likelihood estimate for ϕ.

$$L(\mu, \phi; \mathbf{y}) = \sum_{i=1}^{N} \left\{ y_i \log(\phi\mu_i) - \left(y_i + \frac{1}{\phi}\right)\log(1 + \phi\mu_i) + \log\Gamma\left(y_i + \frac{1}{\phi}\right) - \log\Gamma\left(\frac{1}{\phi}\right) - \log y_i! \right\}$$

$\hat{\mu}_i$ is the sample mean of the observations of $\mathbf{y}$ which belong to the same group as the ith observation. Only overdispersion compared to the Poisson ($\hat{\phi} \geq 0$) was of interest, so $\hat{\phi}$ was found by maximizing $L(\hat{\mu}, \phi; \mathbf{y})$ over the non-negative domain, using an optimisation routine in Matlab 6 (the 'fminbnd' function). The limit of L as $\phi \to 0$ is the log-likelihood of the Poisson distribution, so $L(\mu, 0; \mathbf{y})$ was taken to be the log-likelihood expression for P.

$\mathbf{NB_{mom}}$: Negative binomial model, $V(\mu) = \mu + \phi\mu^2$, method of moments estimator used for ϕ. The log-likelihood expression is as for $\mathrm{NB_{LR}}$, but (following Lawless, 1987) $\tilde{\phi}$ is the estimator of ϕ that satisfies

$$N - p = \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\mu_i(1 + \phi\mu_i)}$$

Again only $\tilde{\phi} \geq 0$ was considered. The right-hand side of this equation is a decreasing function of ϕ for $\phi \geq 0$, so there is a unique solution when $\sum_{i=1}^{N}\{(y_i - \mu_i)^2/\mu_i\} > N - p$. If this condition was not satisfied, $\tilde{\phi}$ was set to 0. As previously, $\hat{\mu}_i$ is the sample mean of the observations of $\mathbf{y}$ which belong to the same group as the ith observation.

$\mathbf{NB_{\varphi\mu,LR}}$: Negative binomial model, $V(\mu) = \phi\mu$, with a maximum likelihood estimate for ϕ.

$$L(\mu, \phi; \mathbf{y}) = \sum_{i=1}^{N} \left\{ y_i \log\left(1 - \frac{1}{\phi}\right) - \frac{\mu_i}{\phi - 1}\log(\phi) + \log\Gamma\left(y_i + \frac{\mu_i}{\phi - 1}\right) - \log\Gamma\left(\frac{\mu_i}{\phi - 1}\right) - \log y_i! \right\}$$

$\hat{\mu}_i$ is the sample mean of the observations of $\mathbf{y}$ which belong to the same group as the ith observation. Only overdispersion compared to the Poisson was of interest, so $\hat{\phi}$ was found by maximizing $L(\hat{\mu}, \phi; \mathbf{y})$ over the domain $\phi \geq 1$, using an optimisation routine in Matlab 6 (the 'fminbnd' function).

The limit of L as $\phi \to 1$ is the log-likelihood of the Poisson distribution, so $L(\mu, 1; \mathbf{y})$ was taken to be the log-likelihood expression for P.

$\mathbf{NB_{\varphi\mu,mom}}$: Negative binomial model, $V(\mu) = \phi\mu$, method of moments estimator used for ϕ. The log-likelihood expression is as for $\mathrm{NB}_{\phi\mu,\mathrm{LR}}$, but (following McCullagh and Nelder, 1989, pp. 199–200) the estimator of ϕ is

$$\tilde{\phi} = \max\left(1, \frac{1}{N - p}\sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\mu_i}\right)$$

The minimum possible value of $\tilde{\phi}$ was set to 1 because only overdispersion was of interest. As previously, $\hat{\mu}_i$ is the sample mean of the observations of $\mathbf{y}$ which belong to the same group as the ith observation.

Zero-inflated distributions: If the frequency of zeros in data is substantially higher than would be expected under a Poisson or negative binomial model, a model with extra zeros could be used for abundances:

$$\mathrm{P}(Y_i = x) = \begin{cases} = 1 - \pi_i & x = 0 \\ = \pi_i f(x; \mu_i, \phi) & x > 0 \end{cases} \quad (1)$$

where $f(x; \mu_i, \phi)$ is a probability distribution separately fitted to non-zero values. A truncated Poisson or negative binomial distribution has been suggested for $f(x; \mu_i, \phi)$ when modelling the abundance of Leadbeater's possum (Welsh *et al.*, 1996).

Equation (1) could be considered as a mixture model, but it is a different type of mixture model to the one proposed for fitting zero-inflated distributions by Lambert (1992). The model due to Lambert (1992) mixes a distribution degenerate at zero with a count distribution, so the mixing proportion represents the proportion of extra zeros in the distribution. The above model, however, is a mixture of a distribution that is degenerate at zero and a count distribution that is *truncated at zero*. Formulating the model in this way ensures that $\hat{\pi}_i$ and $\hat{\mu}_i$ are orthogonal, and can be estimated separately (Welsh *et al.*, 1996). A particular advantage of this approach is in interpretation—the practitioner can separately assess the effects of presence/absence (π_i) and the magnitude of abundances (μ_i). Distinguishing changes in abundance from changes in presence/absence is desirable in many contexts, as these have different implications in terms of biology and conservation.

Two cases were considered:

1. **ZIP**: Zero-inflated Poisson distribution. In this case $f(x; \mu_i)$ has a truncated Poisson distribution, i.e.

$$f(x; \mu_i) = \frac{1}{1 - e^{-\mu_i}} e^{-\mu_i} \frac{x^{\mu_i}}{x!}$$

and so

$$L(\mu, \pi; \mathbf{y}) = \sum_{i=1}^{N} \mathrm{I}(y_i = 0) \log(1 - \pi_i) + \sum_{i=1}^{N} \mathrm{I}(y_i > 0) \{\log \pi_i + y_i \log \mu_i - \log(e^{\mu_i} - 1) - \log y_i!\}$$

This simplifies to a presence/absence and non-zero component, $L(\mu, \pi; \mathbf{y}) = L_\pi(\pi; \mathbf{y}) + L_\mu(\mu; \mathbf{y})$:

$$L_\pi(\pi; \mathbf{y}) = \sum_{i=1}^{N} \{\mathrm{I}(y_i = 0) \log(1 - \pi_i) + \mathrm{I}(y_i > 0) \log(\pi_i)\}$$

$$L_\mu(\mu; \mathbf{y}) = \sum_{i=1}^{N} \mathrm{I}(y_i > 0) \{y_i \log \mu_i - \log(e^{\mu_i} - 1) - \log y_i!\}$$

The presence/absence component is a standard logistic regression, and in this case $\hat{\pi}_i$ is the sample proportion of non-zero values of $\mathbf{y}$ in the group to which the ith observation belongs. When $\hat{\pi}_i = 0$ or 1, the ith component of L_π was set to zero for all zero-inflated distributions.

Welsh *et al.* (1996) give an algorithm for maximum likelihood estimation of L_μ, but in the present context, $\hat{\mu}_i$ is the value of μ_i satisfying

$$\bar{\mu}_i = \frac{\mu_i}{1 - e^{-\mu_i}} \tag{2}$$

where $\bar{\mu}_i$ is the sample mean of the non-zero observations of $\mathbf{y}$ which belong to the same group as the ith observation. Equation (2) does not have a closed form solution, but since the right-hand side is an increasing function spanning $(1, \infty)$ for positive μ_i, it has a unique solution in this range. The Newton–Raphson algorithm was used to find this solution iteratively.

As μ_i approaches zero, the right-hand side of equation (2) approaches 1. So when all non-zero observations in a group were 1, $\hat{\mu}_i = 0$.

2. **ZINB**: Zero-inflated negative binomial. Two parameterizations of the negative binomial distribution have been described in the above, which lead to two different values of the log-likelihood if the nuisance parameter (ϕ) is held constant for all observations of a variable. For brevity, the log-likelihood expressions below are based only on the negative binomial parameterization for which $V(\mu) = \mu + \phi\mu^2$, and results are only presented for this parameterization (as it generally had higher log-likelihood). In this case $f(x; \mu, \phi)$ has a truncated negative binomial distribution, i.e.

$$f(x; \mu, \phi) = \frac{1}{1 - p^{1/\phi}} \frac{\Gamma(x + 1/\phi)}{\Gamma(1/\phi)x!} (1 - p)^x p^{1/\phi}$$

where $p = 1/(1 + \phi\mu)$. Note that the limit as $\phi \to 0$ is the zero-inflated Poisson distribution.

As previously, $L(\mu, \phi, \pi; \mathbf{y}) = L_\pi(\pi; \mathbf{y}) + L_\mu(\mu, \phi; \mathbf{y})$, although on this occasion

$$L_\mu(\mu, \phi; \mathbf{y}) = \sum_{i=1}^{N} \mathrm{I}(y_i > 0) \left\{ y_i \log(1 - p) - \log\left(p^{-\frac{1}{\phi}} - 1\right) + \log\Gamma\left(y_i + \frac{1}{\phi}\right) - \log\Gamma\left(\frac{1}{\phi}\right) - \log y_i! \right\}$$

The maximum likelihood estimates $\hat{\mu}_i$ and $\hat{\phi}$ were found iteratively, using the current estimate of ϕ to find the future estimates of μ_i, and vice versa. The initial estimate of ϕ was 0.

Given ϕ, the maximum likelihood estimate of μ_i satisfies the following:

$$\bar{\mu}_i = \frac{\mu_i}{1 - (1 + \phi\mu_i)^{-1/\phi}}$$

where $\bar{\mu}_i$ is the mean of the non-zero observations of $\mathbf{y}$ in the group to which the ith observation belongs. The solution to this equation has the same properties as described for the zero-inflated Poisson case, equation (2). In particular, as previously, as μ_i approaches zero, the right-hand side approaches 1, so when all non-zero observations were 1, $\hat{\mu}_i$ was set to 0.

Because the score equation for $\hat{\phi}$ does not have a simple form, $\hat{\phi}$ was estimated given current estimates of μ_i using an optimization function on Matlab (the 'fminbnd' function).

For some taxa, there was only one or no non-zero abundances in every group, which is insufficient for estimation of parameters. In this case $\hat{\phi}$ was set to 0, i.e. the ZIP procedure was used for estimation of μ_i.

3. RESULTS

3.1. AIC comparison

Across the 20 datasets considered here, AIC values were calculated for a total of 1672 variables. These values were averaged across datasets to provide a summary comparing all models (Table 2), and averaged within a dataset in Figure 1. Because AIC values are interpreted relative to each other, and the original scale has no intrinsic meaning, they were rescaled on Figure 1 so that the smallest AIC in each dataset was 0.

It is clear that some variables exhibited strong overdispersion compared to the Poisson distribution such that, overall, the AIC values were much higher for the Poisson and zero-inflated Poisson models than for other models (Table 2). However, in many cases overdispersion was not observed, so that although the sum of AIC values was about three times as large for the Poisson model as for the negative binomial, the AIC for the Poisson model was smaller than that of the negative binomial for half of all variables (892 of the 1672, 53 per cent).

Of the two methods of incorporating overdispersion using a negative binomial model, using the mean–variance function $V(\mu) = \mu + \phi\mu^2$ usually led to slightly lower values of AIC. Across datasets, the AIC was slightly smaller when $V(\mu) = \mu + \phi\mu^2$ than when $V(\mu) = \phi\mu$ for all of the 20 datasets considered and for 75 per cent of variables.

When the zero-inflated negative binomial distribution could be fitted to abundance data, it usually had a poorer fit than the negative binomial distribution. In many cases ZINB could not be fitted to data in the first place. For 672 variables, there were so few non-zero values that parameters of ZINB could not be estimated. Of the 1000 variables for which it could be fitted, the zero-inflated negative binomial model fitted data better than the negative binomial for only 112 (11 per cent) variables, and for only 7 per cent of variables with counted abundances. The sum of all AIC values was smaller for the negative binomial model without zero inflation (Table 2). In fact, the sum of AIC values in each of the 20 datasets was smaller for the negative binomial without zero-inflation (Figure 1).

Although transformed least squares was not the best fitting model for data, it fitted data reasonably well. Surprisingly, transformed least squares appeared to fit data about as well as the zero-inflated negative binomial model, based on AIC values (Table 2, Figure 1). The AIC for transformed least squares was not as small as for the negative binomial model, overall, although it was smaller for 20 per cent of the variables considered here.

Estimating the nuisance parameter in the negative binomial model by moments had little effect on log-likelihood values in most instances.

Table 2. Mean AIC value across all 1672 variables in the 20 datasets

Model	Mean AIC
LS	120
P	298
NB_{LR}	105
NB_{mom}	109
$NB_{\phi\mu,LR}$	109
$NB_{\phi\mu,mom}$	114
ZIP	221
ZINB	121

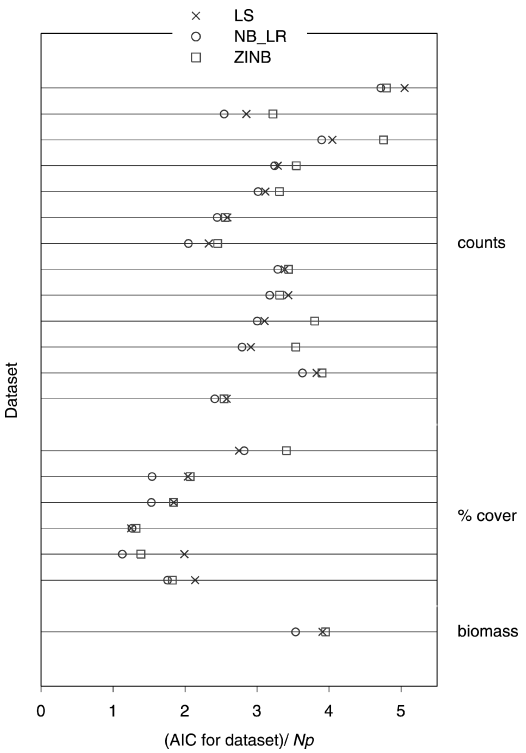

Figure 1. Comparison of AIC for $\log(\mathbf{y}+1)$ transformed least squares (LS), negative binomial (NB_LR) and zero-inflated negative binomial (ZINB) models. Values plotted are the average AIC across all variables in a dataset, but rescaled so that the smallest of these is set to zero. Datasets are ordered as in Table 1, and grouped by type of abundance measure (counts, per cent cover or biomass)

Burnham and Anderson (1998, p. 51) recommend using a small-sample correction term for AIC when there are a small number of observations. Use of this correction increases the penalty of including extra parameters in the model. The AIC values were recalculated using this correction term, and similar results obtained, except that the zero-inflated negative binomial model usually had noticeably higher AIC than the transformed least squares model. As the LS and $\mathrm{NB_{LR}}$ models contained the same number of parameters, their relative differences in AIC are not affected by choice of AIC formula.

3.2. Graphical diagnostics

Extra zeros: For a negative binomial distribution parameterized so that $V(\mu) = \mu + \phi\mu^2$,

$$\mathrm{P}(Y_i = 0) = (1 + \phi\mu_i)^{-1/\phi}$$

and so the expected number of zeros in a variable, denoted $\mathrm{E}(n_0)$, is

$$\mathrm{E}(n_0) = \sum_{i=1}^{N} (1 + \phi\mu_i)^{-1/\phi}$$

The parameter estimates obtained under ZINB were used to compare the number of zeros in each variable, n_0, with the expected number under this negative binomial model for abundances, $\mathrm{E}(n_0)$. By using parameter estimates from ZINB, only non-zero abundances were used in estimation of $\mathrm{E}(n_0)$.

If abundances tend to have unusually high numbers of zeros, compared to a negative binomial distribution, then we would expect $n_0 > \mathrm{E}(n_0)$ in most cases. However, this was not observed (Figure 2). In relatively few instances did a negative binomial model substantially underestimate the observed number of zeros (values toward the top-left of plots in Figure 2).

For datasets of counted abundances, there was generally quite close agreement between the observed number of zeros and that expected under a negative binomial model. Points on Figure 2 were

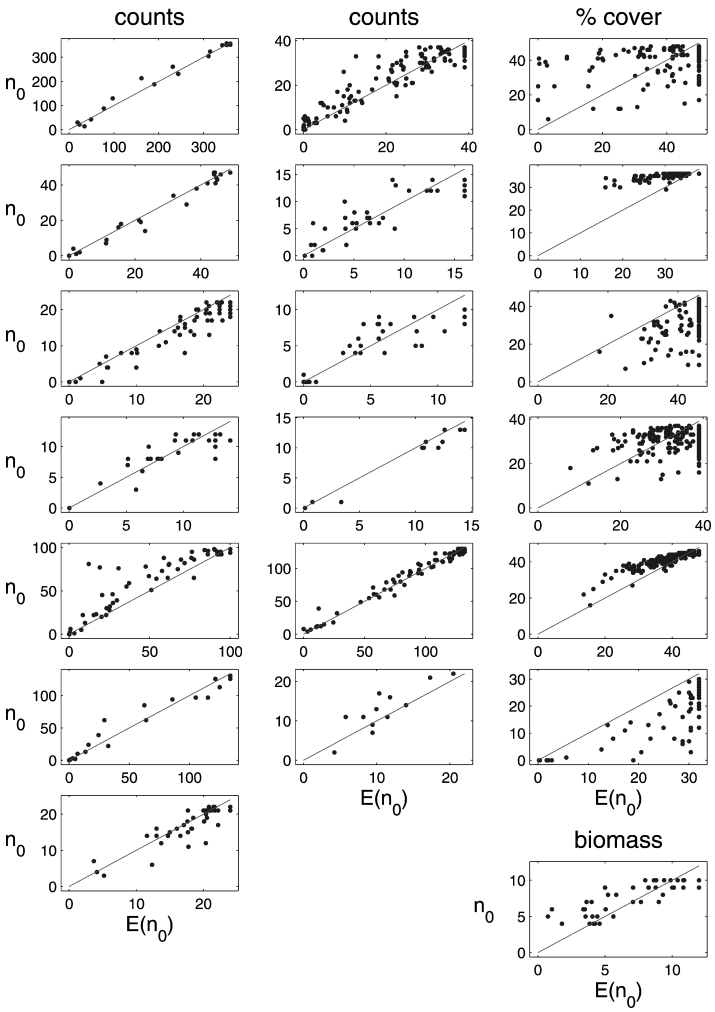

Figure 2. Observed vs expected number of zeros in a variable, n_0 vs $\mathrm{E}(n_0)$. Only non-zero values were used to estimate $\mathrm{E}(n_0)$, under a negative binomial model estimated as in ZINB. Datasets are ordered as in Table 1 when reading down the page, and abundance was measured as counts (first two columns), per cent cover (last column) or biomass (bottom right)

usually close to the one-to-one line, with more scatter in small datasets where sample variation is greater on the proportion scale.

For datasets where abundance was measured as per cent cover, number of zeros in a variable was usually poorly predicted (Figure 2, last column). This suggests that a negative binomial model for rescaled per cent cover was not useful. Even in these cases, there was no consistent trend for underestimation of the number of zeros, suggesting that there is no general pattern of extra zeros in abundance data.

Mean–variance relationship: The assumed mean–variance relationship plays an important role in modelling, so plots of the sample variance vs the sample mean were used to assess how well the mean–variance relationship was modelled by the negative binomial and transform least squares models (Figure 3). Obviously, it was not feasible to consider all 1672 variables—only one variable from each dataset was plotted, and only variables with at least three non-zero means were plotted (given that little can be seen of the mean–variance relationship if most groups have zero mean and variance).

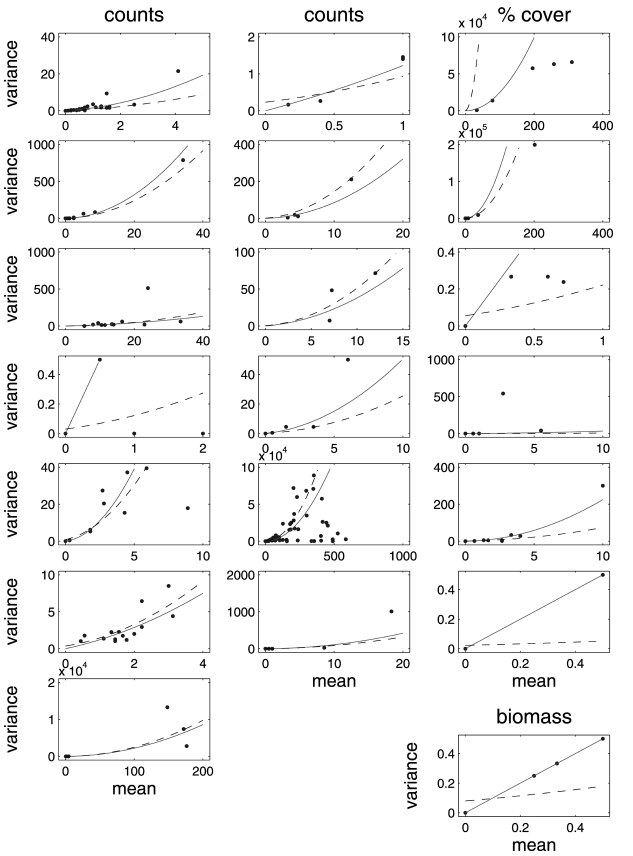

Figure 3. Sample mean–variance relationship for one randomly chosen variable from each dataset. Each point represents the sample mean and variance of a group, the lines are fitted mean–variance relationships under a negative binomial model (red, solid line) or a transformed least squares model (black, dashed line). One variable from each dataset is represented in a plot, the variable plotted being randomly chosen from all the variables in the dataset that have at least three non-zero means. Reading down the page, datasets are ordered as in Table 1

Under the transformed least squares model, a particular type of increasing quadratic trend was expected on mean–variance plots. The transformed least squares model assumes $\log(\mathbf{Y}+1)$ has the same variance σ^2 for all groups, and using standard results for lognormal variables (Johnson *et al.*, 1994), $\text{var}(Y_i) = (\mu_i+1)^2(e^{\sigma^2}-1) \approx (\mu_i+1)^2\sigma^2$.

Counted data displayed a roughly quadratic mean–variance relationship (Figure 3) that was reasonably well fitted by either assuming a negative binomial model or that $\log(\mathbf{Y}+1)$ is Gaussian. In fact, these two curves were almost indistinguishable on some of the plots. The apparent differences between the two fits were:

- transformed least squares underestimated the variance of larger means when there were many groups with zero variance
- for transformed least squares, the fitted variance approaches (approximately) $\hat{\sigma}^2$ as $\hat{\mu}_i$ approaches zero, whereas for the negative binomial the fitted variance (correctly) approaches zero.

Figure 3 suggests that the second of these points is a minor consideration. For most variables, the largest mean was two or greater (Figure 3), in which case the largest fitted variance under LS was at least $\hat{\sigma}^2(2+1)^2 = 9\hat{\sigma}^2$, so the difference between the limiting values $\hat{\sigma}^2$ and 0 was negligible compared to the range of variances across all groups. Consequently, we have the somewhat surprising result that despite assuming equal variance across groups, least squares can model the mean–variance relationship of rare taxa quite well (although if many groups have variance zero, least squares usually does relatively poorly).

Neither model appeared to adequately describe the mean–variance relationship of per cent cover data, for a couple of datasets in particular. In these datasets difficulties seemed to occur because even though non-zero abundances were often large (10–80 per cent), there were occasional, very small abundance values in most variables. This was due to the presence of saplings or small plants in the understorey in some quadrats, or to a branch hanging over the quadrats. This means that individual plant size varied considerably within a species, so the assumption that per cent cover is proportional to counted data was inappropriate. The relatively high number of groups with zero variance caused difficulties for least squares methods.

4. DISCUSSION

From these results, the general interpretation of multivariate abundance data is that, although there are usually many zeros in the typical dataset, this does not appear to be because abundances have extra zeros compared to a negative binomial distribution. Such abundances are more likely to arise from negative binomial distributions with small means than from zero-inflated negative binomial distributions. In many recent studies, abundances have been modelled using zero-inflated distributions (Desouhant *et al.*, 1998; Pearce and Ferrier, 2001; Lindenmayer *et al.*, 1999; Dobbie and Welsh, 2001; Welsh *et al.*, 2000), but of these studies only Desouhant *et al.* (1998) compared the goodness-of-fit with alternative count distributions. If one were to fit a zero-inflated model, it would be advisable to present quantitative evidence that the zero-inflation term was required. Based on the present results, it is likely that a term for extra zeros is not needed, and a simpler model will usually suffice.

Modelling abundances with a zero inflated distribution implies that replicate observations are a mixture of two different types—ones where organisms of a taxon do not occur, and ones where the organisms do occur (and follow a nominated distribution). This approach to modelling would only be required in cases where the two different types of observations cannot be distinguished (either a priori

or using environmental variables measured during sampling). Note that, in regression modelling, it is the conditional distribution of abundance (conditional on measured environmental variables) that is modelled to be zero inflated. Hence, for this model to be appropriate, observations with *the same values* of measured environmental variables must be a mixture of places where the organism occurs in low abundance ($\mu > 0$) and does not occur at all ($\mu = 0$). In this study it appeared that zero inflated distributions were not usually needed, which implies that observations in which a taxon does not occur could usually be distinguished from those where the taxon does occur (a priori or using environmental variables), i.e. most zeros could be attributed to the systematic component of the model, rather than taking the more complicated route and incorporating them into the random component of the model.

A secondary conclusion from this study is that transformed least squares, although a relatively crude approach, can fit abundance data reasonably well. The equal-variance property does not prohibit use of this model for rare taxa, although when many groups have zero variance, variance in other groups appears to be underestimated. It is recommended that the first choice model for abundance be negative binomial log-linear models, and least squares only be used in complex models where the negative binomial cannot be used easily, for example when accounting for spatial autocorrelation.

A Poisson model with overdispersion parameter is commonly fitted to counted data (McCullagh and Nelder, 1989); however, in this study, an alternative negative binomial usually fitted data better than a model with variance proportional to the mean. Both AIC values (Table 2) and mean–variance plots (Figure 3) suggested that a quadratic trend between variance and mean ($V(\mu) = \mu + \phi\mu^2$) was usually more consistent with data than a linear relationship ($V(\mu) = \phi\mu$). When there is a quadratic mean–variance relationship, assuming the variance is proportional to the mean will systematically underestimate variance for larger means, and overestimate the variance for intermediate means.

In summary, although one of the key properties of multivariate abundances is a high frequency of zeros (Clarke and Green, 1988, or Figure 2), our results suggest that special techniques are not generally necessary to account for the high frequency of zeros. The negative binomial was found to be a good model for the number of zeros in counted abundance datasets, suggesting that a good approach to analysing such data will often be to use negative binomial log-linear models. This is not the case when abundances are measured as per cent cover. Caution should be exercised when using least squares to model abundances; in particular, variances may be underestimated when many observations have an estimated mean of 0.

ACKNOWLEDGEMENTS

The author is grateful to all who contributed their data, and to those who commented on the manuscript—Mark Westoby, Malcolm Hudson, Barry Quinn and the anonymous referees.

REFERENCES

Acton FS. 1970. *Numerical Methods That Work*. Harper & Row: New York.

Akaike H. 1972. Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Petrov BN, Csaki F (eds). Akademiai Kiado: Budapest; 267–281.

Attrill MJ, Power M, Thomas RM. 1999. Modelling estuarine crustacea population fluctuations in response to physico-chemical trends. *Marine Ecology—Progress Series* **178**: 89–99.

Biondini ME, Mielke PW Jr, Berry KJ. 1988. Data-dependent permutation techniques for the analysis of ecological data. *Vegetation* **75**: 161–168.

Burnham KP, Anderson DR. 1998. *Model Selection and Inference—a Practical Information-theoretic Approach*. Springer-Verlag: New York.

Catling P, Burt R, Forrester R. 2000. Models of the distribution and abundance of ground-dwelling mammals in the eucalypt forests of north-eastern New South Wales in relation to habitat variables. *Wildlife Research* **27**: 639–654.

Clarke KR, Green RH. 1988. Statistical design and analysis for a 'biological effects' study. *Marine Ecology—Progress Series* **46**: 213–226.

Clements A. 1980. *The vegetation of bushland in the northern Sydney area*. M.Sc. thesis, Macquarie University, Australia.

Desouhant E, Debouzie D, Menu F. 1998. Oviposition pattern of phytophagous insects: on the importance of host population heterogeneity. *Oecologia* **114**: 382–388.

Dobbie MJ, Welsh AH. 2001. Modelling correlated zero-inflated count data. *Australian and New Zealand Journal of Statistics* **43**: 431–444.

Gee JM, Warwick RM, Schaanning M, Berge JA, Ambrose WG Jr. 1985. Effects of organic enrichment on meiofaunal abundance and community structure in sublittoral soft sediments. *Journal of Experimental Marine Biology and Ecology* **91**: 247–262.

Gray JS, Clarke KR, Warwick RM, Hobbs G. 1990. Detection of initial effects of pollution on marine benthos: an example from the Ekofisk and Eldfisk oilfields, North Sea. *Marine Ecology—Progress Series* **66**: 285–299.

Guisan A, Zimmerman NE. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* **135**: 147–186.

Johnson NL, Kotz S, Balakrishnan N. 1994. *Continuous Univariate Distributions, Volume 1* (2nd edn). John Wiley & Sons: New York.

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

Lawless JF. 1987. Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* **15**: 209–225.

Lindenmayer DB, Cunningham RB, Pope ML, Donnelly CF. 1999. The response of arboreal marsupials to landscape context: a large-scale fragmentation study. *Ecological Applications* **9**: 594–611.

McCullagh P, Nelder JA. 1989. *Generalized Linear Models* (2nd edn). Chapman & Hall: London.

Miller RG Jr. 1986. *Beyond ANOVA, Basics of Applied Statistics*. John Wiley & Sons: New York.

Moulton TP. 1982. *The effect of prescribed burning and simulated burning on soil and litter arthropods in open forest at Cordeaux, N.S.W., Australia*. Ph.D. thesis, Macquarie University, Australia.

Pearce J, Ferrier S. 2001. The practical value of modelling relative abundance of species for regional conservation planning: a case study. *Biological Conservation* **98**: 33–43.

Pearson TH, Blackstock J. 1984. Garroch Head sludge dumping ground survey, final report; Technical report, Dunstaffnage Marine Research Laboratory.

van den Brink PJ, van Wijngaarden RPA, Lucassen WGH, Brock TCM, Leeuwangh P. 1996. Effects of the insecticide Dursban 4E (active ingredient chlorpyrifos) in outdoor experimental ditches: II. Invertebrate community responses and recovery. *Environmental Toxicology and Chemistry* **15**: 1143–1153.

van der Aart PJM, Smeenk-Enserink N. 1970. Correlations between distribution of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. *Netherlands Journal of Zoology* **25**: 1–45.

van Dobben HF, ter Braak CJF, Dirkse GM. 1999. Undergrowth as a biomonitor for deposition of nitrogen and acidity in pine forest. *Forest Ecology and Management* **114**: 83–95.

Warton DI, Hudson HM. 2004. A MANOVA statistic is just as powerful as distance-based statistics, for multivariate abundances. *Ecology* **85**: 858–874.

Warwick RM, Carr MR, Clarke KR, Gee JM, Green RH. 1988. A mesocosm experiment on the effects of hydrocarbon and copper pollution on a sublittoral soft-sediment meiobenthic community. *Marine Ecology—Progress Series* **46**: 181–191.

Warwick RM, Clarke KR, Gee JM. 1990a. The effect of disturbance by soldier crabs, *Mictyris platycheles* H. Milne Edwards, on meiobenthic community structure. *Journal of Experimental Marine Biology and Ecology* **135**: 19–33.

Warwick RM, Platt HM, Clarke KR, Agard J, Gobin J. 1990b. Analysis of macrobenthic and meiobenthic community structure in relation to pollution and disturbance in Hamilton Harbour, Bermuda. *Journal of Experimental Marine Biology and Ecology* **138**: 119–142.

Welsh AH, Cunningham RB, Donnelly CF, Lindenmeyer DB. 1996. Modelling the abundance of rare species: statistical methods for counts with extra zeros. *Ecological Modelling* **88**: 297–308.

Welsh AH, Cunningham RB, Chambers RL. 2000. Methodology for estimating the abundance of rare animals: seabird nesting on North East Herald Cay. *Biometrics* **56**: 22–30.

