Ecology, 93(12), 2012, pp. 2526–2532

Reports

Incorporating the intraspecific occupancy–abundance relationship into zero-inflated models

Adam N. H. Smith,[1,3] Marti J. Anderson,[1] and Russell B. Millar[2]

[1]*New Zealand Institute for Advanced Study, Massey University Albany, Private Bag 102904, Albany, Auckland 0745 New Zealand*
[2]*Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142 New Zealand*

Abstract. Zero-inflated versions of standard distributions for count data are often required in order to account for excess zeros when modeling the abundance of organisms. Such distributions typically have as parameters λ, the mean of the count distribution, and π, the probability of an excess zero. Implementations of zero-inflated models in ecology typically model λ using a set of predictor variables, and π is fit either as a constant or with its own separate model. Neither of these approaches makes use of any relationship that might exist between π and λ. However, for many species, the rate of occupancy is closely and positively related to its average abundance. Here, this relationship was incorporated into the model for zero inflation by functionally linking π to λ, and was demonstrated in a study of snapper (*Pagrus auratus*) in and around a marine reserve. This approach has several potential practical advantages, including better computational performance and more straightforward model interpretation. It is concluded that, where appropriate, directly linking π to λ can produce more ecologically accurate and parsimonious statistical models of species abundance data.

Key words: *Bayesian modeling; count data; marine reserves; mixed-effects ANOVA; negative binomial distribution; New Zealand; occupancy–abundance relationship; overdispersion;* Pagrus auratus; *snapper; zero inflation.*

Introduction

Ecological studies often seek to quantify the abundance of organisms in order to explain and predict patterns observed in nature. Data from such studies are typically in the form of counts of individuals taken from some standardized sampling unit, such as quadrats or timed searches. When modeling count data, a standard option for the distribution of errors is the Poisson. This distribution has a single parameter, the mean, which also equals the variance (McCullagh and Nelder 1989). However, ecological data sets often have a variance that is greater than the mean, a condition known as overdispersion (Clapham 1936, Bliss and Fisher 1953, White and Bennetts 1996). There are two properties associated with overdispersion that are commonly found in ecological data. The first, broadly termed "contagion," is where individuals are more aggregated than would be expected if they occurred independently (Neyman 1939). The second is through an excess of zeros, termed "zero inflation," where a data set contains more zeros than would be expected from the Poisson distribution with which it is modeled. Both contagion and excess zeros may increase the variance relative to the mean, and therefore contribute to overdispersion. These properties increasingly are being incorporated into statistical models to produce more accurate inferences in ecology (e.g., Ver Hoef and Boveng 2007, Wenger and Freeman 2008).

There are a number of ways that zero-inflated counts can be modeled, including the use of the negative binomial distribution (Warton 2005) and/or explicit zero inflation (see reviews in Lambert 1992, Heilbron 1994, Welsh et al. 1996, Cunningham and Lindenmayer 2005, Martin et al. 2005). The most common way to model zero inflation explicitly is to use a mixture of two statistical distributions. The zero-inflated mixture model can be thought of as a two-step process: a Bernoulli distribution first determines whether the count is to be an excess zero (with probability π) and, if not, another statistical distribution, Φ, then generates the count (Ghosh et al. 2006). The general zero-inflated random variable can be written as follows:

$$Y \sim \begin{cases} 0 & \text{with probability } \pi \\ \Phi(\boldsymbol{\theta}) & \text{with probability } 1-\pi \end{cases} \quad (1)$$

where π is the probability of an excess zero and $\boldsymbol{\theta}$ is a vector of parameters for the count distribution, Φ. If Φ is a Poisson distribution, $\boldsymbol{\theta}$ consists of parameters that determine the value of λ, the mean count conditional on an excess zero not occurring. If Φ is a negative binomial distribution, $\boldsymbol{\theta}$ also includes a dispersion parameter, δ. Under this model, the mean of Y is $\mu = (1-\pi)\lambda$. Note that a zero can arise under the mixture model either as an excess zero (under the Bernoulli) or directly under the chosen count distribution, Φ. An alternative approach, termed the "conditional" or "hurdle" model, requires

Manuscript submitted 20 March 2012; revised 5 July 2012; accepted 9 July 2012. Corresponding Editor: B. D. Inouye.
[3] E-mail: anhsmith@gmail.com

the statistical distribution Φ to have a minimum value of one, ensuring that only the Bernoulli distribution can produce zeros (Welsh et al. 1996, Cunningham and Lindenmayer 2005).

Ecologists often wish to use a statistical model to make inferences with respect to the mean of some measure of abundance. This is usually a density in the form of counts of individuals per sampling unit that has been standardized (temporally and/or spatially) across the full study design. For analyzing such data, generalized linear models (GLMs, McCullagh and Nelder 1989) provide a convenient framework, allowing some nonnormal error structures to be modeled while incorporating relevant predictor variables to explain variation in the mean. In a standard GLM for count data, a log link function is used to map the mean (λ) onto a linear predictor η_λ, where η_λ is a linear combination of k predictor variables, $X_1, X_2, \ldots, X_k$, giving

$$\log(\lambda) = \eta_\lambda = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad (2)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ are coefficients to be estimated. In the case of a Poisson model, the vector $\boldsymbol{\beta}$ corresponds to the vector of parameters $\boldsymbol{\theta}$ in Eq. 1.

We will now describe three models for zero inflation that are commonly used in conjunction with a GLM that predicts the conditional mean of the count values; a fourth model is then described in the next section. The first model simply allows for no zero inflation, as follows:

$$\pi = 0. \qquad \text{(Model 1)}$$

Secondly, π may be modeled as a constant value for all observations, giving

$$\pi = \alpha. \qquad \text{(Model 2)}$$

Thirdly, π may be modeled using its own separate linear predictor, η_π, typically with a logit link function, giving

$$\log\Big(\pi/(1-\pi)\Big) = \eta_\pi = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \ldots + \alpha_\ell Z_\ell \qquad \text{(Model 3)}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_\ell)$ is a vector of coefficients and $Z_1, Z_2, \ldots, Z_\ell$ are a set of ℓ predictor variables, which could be equivalent to some or all of the predictors $X_1, X_2, \ldots, X_k$ in Eq. 2. Modeling λ and π using two separate linear predictors can result in a heavily parameterized model. However, this approach may well be appropriate in situations where the ecological drivers giving rise to excess zeros are fundamentally different from those that are important in predicting patterns of relative abundance. Importantly, none of these commonly implemented approaches formally makes use of any potential relationship between the patterns of excess zeros and the mean count of abundance.

There is substantial empirical evidence for the existence of a strong relationship between presence (or occurrence) and abundance for many species. Specifically, the proportion of sites occupied by a species is positively related to its average abundance at a range of spatial or temporal scales (Brown 1984). Borregaard and Rahbek (2010: Fig. 1) show a useful schematic diagram of this intraspecific occupancy–abundance relationship, along with a review of proposed mechanisms. A relationship between occupancy rates and mean abundance is implicit to standard count distributions such as the Poisson and the negative binomial (Wright 1991, Hartley 1998), yet the maintenance and utility of this relationship in the context of zero-inflated models has not been fully explored. Recently, two articles (Nielsen et al. 2005, Sileshi et al. 2009) have examined the relationship between occupancy and abundance using parameters derived from zero-inflated models with separate linear predictors for π. The results were mixed, indicating that the relationship varies among taxa. For all five species of insect examined, parameters for occupancy and abundance were strongly positively related (Sileshi et al. 2009), while for moose, there was a weak positive relationship and, for bracken fern, there was no apparent relationship (Nielsen et al. 2005). Although zero-inflated models have been used in the study of occupancy–abundance relationships, this relationship, itself, has not yet been used explicitly and directly to improve zero-inflated models. A model with a direct link between zero inflation and abundance is presented in the next section.

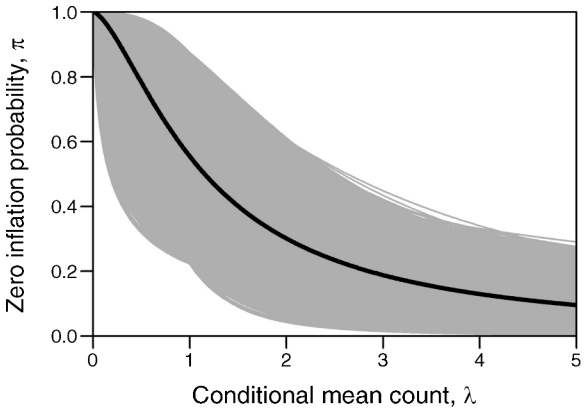

FIG. 1. The relationship between the conditional mean count (λ) of snapper (*Pagrus auratus*) per baited underwater video deployment and the probability of an excess zero (π) for legal-sized snapper from a marine reserve monitoring program in northeastern New Zealand. This relationship was estimated using a Bayesian zero-inflated model (Appendix A) where π and λ were linked explicitly as $\text{logit}(\pi) = \gamma_0 + \gamma_1 \log(\lambda)$. The black line shows this function using point estimates for the parameters of $\gamma_0 = 0.34$ and $\gamma_1 = -1.60$. The gray lines show this function using the paired values of these parameters under Markov chain Monte Carlo (MCMC) within their joint 95% credible bounds.

The Linked Model

Here, we suggest that an occupancy–abundance relationship can be incorporated directly into a zero-inflated statistical model by using a very simple

approach in which a single linear predictor is used to model both π and λ (Lambert 1992). The link between the two parameters may be made by employing a simple linear model:

$$\text{logit}(\pi) = \gamma_0 + \gamma_1 \times \eta_\lambda \qquad \text{(Model 4)}$$

where η_λ is the linear predictor for λ from Eq. 2 and γ_0 and γ_1 are parameters to be estimated. The parameter γ_1 will usually be negative, reflecting a decreasing rate of zero inflation (non-occupancy) with increasing λ. Note that π and λ are explicitly linked *within* the modeling process, rather than examining the potential relationship post hoc, as in previous studies (e.g., Nielsen et al. 2005, Sileshi et al. 2009).

The idea of linking π and λ in zero-inflated models was originally proposed by Lambert (1992) in the context of modeling defects in a manufacturing process. More recently, Liu and Chan (2010) developed a library of functions for R (R Development Core Team 2012), named COZIGAM, which can implement linked ("constrained") zero-inflated generalized additive models (GAMs). This general approach is not yet widespread, however. Furthermore, COZIGAM is restricted to the one-parameter exponential family of distributions (precluding, e.g., the negative binomial distribution) and can only implement models that contain at least one nonparametric term, such as a smoothing function. Here, we demonstrate that, within the Bayesian framework, the linked approach can be easily implemented in a fully parametric model, incorporating generalized linear mixed model structures with fixed and random factors and a nonstandard error distribution (i.e., the negative binomial). These sorts of features are commonly required in ecological models.

In the following section, we present analyses of counts of snapper (*Pagrus auratus*, Sparidae) from baited underwater video (BUV) deployments in northeastern New Zealand. We applied a set of potential models to allow for contagion and zero inflation in the four forms just described. The linked model was found to give the best fit to the data with a relatively modest number of parameters, and results obtained under this model are presented. We then discuss in greater detail the advantages of using a zero-inflated mixture model that directly links zero inflation with abundance in ecological studies.

Example

Background

This example uses data from a monitoring program of snapper (*Pagrus auratus*) at Te Whanganui-A-Hei (Hahei) marine reserve, a no-take marine park covering 9 km^2, established in 1993 in northeastern New Zealand (Willis et al. 2003). This species is of considerable ecological and commercial importance, and is the most heavily targeted recreationally fished species in this region. Full details of the locations, methodology, and design of the monitoring program, including an analysis of its first three years of data, are provided in Willis et al. (2003). The monitoring program consisted of BUV surveys (Willis and Babcock 2000), a method which records video footage from a camera mounted on a steel frame, to which bait is attached to attract carnivorous fish. The data here consisted of counts of the maximum number of legal-sized (>27 cm fork length) individuals of snapper seen in any single frame from each 30-min baited underwater video recording. This standardized sampling unit ensured no individual fish was counted more than once (Willis et al. 2003), and was consistently applied across the entire study design. Here, the aim of the analysis was to estimate the ratio of the density of legal-sized snapper in areas within the reserve relative to areas outside of the reserve. The variation attributable to reserve status was then compared to estimates of other spatial and temporal components of variation inherent in the study design.

The monitoring program was designed as follows. A c. 10 km length of coast that includes the reserve was divided into six areas: three within the reserve and three outside the reserve (see Willis et al. 2003: Fig. 1). In a single survey, five ($\pm$ 2) BUV replicate drops were done at random locations within each area. Surveys were repeated in each of nine years, specifically 1997–2001, 2003, 2004, 2006, and 2010. In each year, surveys were done in spring (1997, 2004), autumn (1999, 2003, 2006, 2010), or both (1998, 2000, 2001), resulting in an unbalanced design. This yielded four factors in a hierarchical mixed-effects ANOVA sampling design: reserve status (fixed), area (random, nested in status), year (random), and season (fixed). There were 348 data values in total, 191 of which were zeros.

Statistical methods

Data were analyzed using a set of models with either Poisson or negative binomial error distributions. Each of these base distributions was applied with each the following four methods for zero inflation described in the previous sections, namely, no zero inflation (model 1) constant zero inflation (model 2), zero inflation with its own separate linear predictor (model 3), and the linked model (model 4). In addition, various combinations of the four factors and their interactions (i.e., the predictor variables) were considered within each class.

A Bayesian approach was used because of the relative ease with which complex models can be fitted, including those with nonlinear and hierarchical model structures and nonstandard (e.g., zero-inflated) error distributions, as required here. Parameter estimates were obtained using Markov chain Monte Carlo (MCMC) methodology, implemented in the software OpenBUGS (Lunn et al. 2009) and called from within R (R Development Core Team 2012) using the R2OpenBUGS library (Sturtz et al. 2005). Standard noninformative prior distributions were used for all estimated parameters (see Appendix A: Table A1). Three MCMC chains were run

TABLE 1. Comparison of a selection of candidate models for estimating counts of legal-sized snapper (*Pagrus auratus*) from a marine reserve monitoring program in New Zealand.

Model no.	Model for log(λ)	Model for logit(π)	Deviance and model complexity				Posterior predictive checks			
			DIC	$\bar{D}$	p_D	p	Total n_0	Total t	Bin $\overline{\in}_{n_0}$	Bin $\overline{\in}_t$
1	$R + S + A + Y$	0	1032.8	1012.7	20.1	21	180	709	0.89	5.5
2	$R + S + A + Y$	α	1030.2	1008.7	21.5	22	184	680	0.93	5.2
3.1	$R + S + A + Y$	R	1031.4	1005.1	26.3	23	186	679	0.91	5.1
3.2	$R + S + A + Y$	$R + S$	1031.8	1002.8	29.0	24	187	675	0.89	5.0
3.3	$R + S + A + Y$	$R + S + A$	1036.3	1004.3	32.0	31	186	679	0.89	5.0
3.4	$R + S + A + Y$	$R + S + A + Y$	1041.7	1003.9	37.8	41	186	679	0.87	5.0
4	$R + S + A + Y$	$\gamma_0 + \gamma_1 \log(\lambda)$	1018.0	994.9	23.1	23	190	670	0.86	4.8

Notes: For all models shown here, the base distribution for counts was the negative binomial. Four classes of zero-inflated models were used, as indicated by the model numbers: (1) no zero inflation, (2) constant zero inflation, (3) a separate linear predictor for zero inflation, and (4) zero inflation linked to the average of the count process. For model 3, submodels 3.1–3.4 contain increasing numbers of parameters in the separate linear predictor for zero inflation, as indicated. Predictor variables are denoted as follows: R, reserve status; S, season; A, area; Y, year. Models were compared using the Deviance Information Criterion (DIC) and its summands, the expected deviance ($\bar{D}$) and the effective number of parameters (p_D). The actual number of stochastic parameters (p) is also provided. The mean of the posterior predictive distributions for the total number of zeros (total n_0) and the total count (total t) is presented. These may be compared with the same values from the observed data, 191 and 660, respectively. Finally, estimates of the mean absolute error for each of n_0 and t, pooled at the level of replicate bins, provide the "mean bin misclassification rate" (Bin $\overline{\in}_{n_0}$) and the "mean bin absolute deviation" (Bin $\overline{\in}_t$). For these measures, smaller values indicate more accurate predictions. See Appendix A for further details of the model and posterior predictive checks.

for each model and convergence was evaluated using the Brooks-Gelman-Rubin $\hat{r}$-statistic (Gelman and Rubin 1992, Brooks and Gelman 1998). Each chain was run for 500 000 iterations, excluding a 250 000 burn-in, and thinned at the rate of 1/50 so that three sets of 10 000 values each were kept.

The predictive performance of the models was compared using the Deviance Information Criterion (DIC; Spiegelhalter et al. 2002) and a number of other measures of performance and complexity. Some authors have suggested that the DIC is inappropriate for mixture models when observation-level latent parameters are used in the likelihood function (e.g., Lawson and Clark 2002). Here, a partially marginalized form of the likelihood was used, thereby avoiding this problem (Millar 2009; see Supplement). The summands of the DIC were also used: the average deviance ($\bar{D}$) provided an estimate of the overall goodness of fit of the models to the data, while the effective number of parameters (p_D) provided a measure of model complexity. The latter was calculated as half the variance of the posterior deviance (Gelman et al. 2004), which is a less conventional alternative to the formulation proposed by Spiegelhalter et al. (2002) with some important potential advantages (Link and Barker 2009). The actual number of stochastic parameters (p) was also noted. The remaining measures of model fit were obtained from posterior predictive checks (Gelman et al. 2004), using 5000 replicate data sets that were generated at random for each model with samples from the joint posterior distributions of the parameters (y^{rep}). The y^{rep} were then compared with the observed data (y) on two summary statistics, namely the number of zero counts and the total number of fish. Comparisons were made at two different levels, first by pooling at the level of the whole data set (i.e., grand totals) and then at the level of the 72 bins of replicates as delineated by the four factors of the study design (see Appendix A: Eqs. A.9–A.12). A full description of the finally selected model, with further details of the calculations of derived parameters, components of variation and posterior predictive checks, are provided in Appendix A. Data and code are provided as a Supplement.

Results

The DIC criterion favored models based on the negative binomial distribution rather than their Poisson-based counterparts, and using only the main effects as predictors for λ. Among these models, the lowest average deviance and the lowest DIC were obtained for model 4, where π was functionally linked to λ (Table 1). This model was also the most accurate for predicting the number of zeros and also the total counts at either the level of the entire data set or the level of the replicate bins. Despite requiring far fewer parameters, it provided better predictions than even model 3.4, where all available factors were used to predict π.

Estimates for some key parameters obtained from model 4 are given below (see Appendix B: Table B1 for a more complete list). The shape of the relationship between π and λ was negative (Fig. 1). The posterior mean for the overall mean count of legal-sized snapper per BUV deployment inside the reserve (averaged across areas, seasons and years) was 3.02, with a 95% credible interval (CI) of 1.12–6.18. Outside the reserve, it was 0.20 (CI = 0.03–0.59). The reserve effect, calculated as the ratio of the mean count per BUV deployment inside : outside the reserve, was estimated to have a posterior median of 16.36 (CI = 4.10–90.00). A comparison of the components of variation showed that reserve status had a far greater influence on counts of snapper than did season, year or area (Appendix B: Table B1). With zero inflation incorporated into the

model, the degree of overdispersion was moderate, with δ estimated to be 2.9 (CI = 1.5–5.0).

Discussion

The occupancy–abundance relationship in snapper

In the example just described, we considered several models for estimating counts of legal-sized snapper inside vs. outside of a marine reserve. These models differed in the way in which the probability of an excess zero, π, was structured. According to the DIC statistic, the best model was obtained when π was functionally linked to the conditional mean of the count distribution, λ. As expected, the relationship was negative (i.e., $\gamma_1 < 0$), indicating that the zero inflation reflected a positive relationship between occupancy and abundance for this species. Here, the specific relationship predicted π to be quite high (>0.5) when $\lambda < 1.2$, reducing to just 0.2 when $\lambda = 2.9$ (Fig. 1), indicating that excess zeros occurred primarily where the mean count was low. The ecological interpretation of this relationship is not straightforward, as it is unknown whether it resulted from small-scale behavioral processes, large-scale population processes, or some mixture of the two. For example, high probability of an excess zero could have arisen in areas of low density because fish were less likely to discover or approach the bait. At a larger scale, fish might actively seek to school more closely with other fish (creating less uniform and more clumped spatial distributions, and hence more zero counts) when the overall numbers of fish in an area are low. Causal mechanisms underlying observed occupancy–abundance relationships clearly require further study.

Advantages of the linked model

The distribution of individuals of a given species in nature can arise from a variety of interacting ecological processes which operate at numerous scales, including mortality, recruitment, dispersal, habitat specificity, environmental constraints, and interactions with other organisms. When a survey is done, a sampling process is superimposed onto this distribution to produce data. It is generally intended that a statistical model should reflect, as closely as possible, the underlying distribution and the sampling process that gave rise to the data to which it is applied. When considering the structure of a zero-inflated model, if the distribution of a species shows a strong relationship between occupancy and abundance at the relevant sampling scale, then explicitly incorporating this relationship may provide a better model which more accurately reflects the underlying patterns and processes. Such a model should then be favored over others by model selection criteria. If, on the other hand, the patterns in occupancy are distinct from those of abundance, then an alternative model may be more appropriate, either with constant π in the case where the excess zeros are not related to the predictor variables, or with a separate linear predictor for π in the case where they are.

When appropriate, the linked model (model 4) has several important advantages over the more commonly used alternatives (i.e., models 2 and 3). First is the principle of parsimony: we should seek to implement a model that adequately explains the variation in the data while using a minimal number of parameters. Although model 2 adds only one parameter to the base model, a constant rate of zero inflation across the whole data set may be overly simplistic in many situations. In contrast, using model 3 allows for a range of values of π but will generally require the introduction of many more parameters. Exploiting the link between occupancy and abundance can achieve the best of both worlds: π can take a wide range of values through the inclusion of just two extra parameters (γ_0 and γ_1).

From a practical point of view, if primary interest lies in estimating the overall mean (μ), modeling π separately from λ complicates the interpretation of the model with respect to μ. Because μ is a function of both π and λ, the overall effects of the predictor variables on μ are split between the two linear predictors. This can be considered an advantage in cases where different processes affect π and λ (Cunningham and Lindenmayer 2005). However, if not, model 4 has the advantage that the linear predictor η_λ is used to predict both π and λ, thereby simplifying the process of model selection and interpretation of the coefficients. The coefficients may be compared directly to identify those with the greatest influence on count values, as they are determined by both the mean of the count process and the excess zeros simultaneously. Model 4 also provides estimates of γ_0 and γ_1, which can yield useful insights regarding the shape of the relationship between zero inflation and abundance (see Appendix C).

Although implementation of the linked model is straightforward under the Bayesian paradigm (see Supplement), this approach introduces a complex nonlinear structure that is difficult to implement using frequentist methods. Nonetheless, there are important potential computational advantages to the linked model. Where occupancy is strongly related to abundance, having separate predictors in the model is likely to result in correlations among the two sets of coefficients. Intuitively, it is best to avoid this sort of redundancy in the parameters of a model. Practically, correlated parameters can result in problems with convergence and poor estimation when fitting the model. This was apparent in our analysis when the MCMC chains for class-3 models generally took longer to converge and had greater autocorrelation than the other models. Another consideration is that, as π approaches either 0 or 1, the data in its binary (i.e., observed or not observed) form has low variance and contains very little information. This can cause low statistical power and computational difficulties associated with extremely low (or high) values on the logit scale, because as π tends to 0 (or 1), logit(π) tends to $-\infty$ (or ∞) (Cunningham and Lindenmayer 2005). The linked model may help relieve

this problem because the parameter estimates are determined by the full range of integer values in the data.

Here, we considered only the mixture-model approach to zero inflation, rather than the hurdle approach where the count distribution is truncated so that it cannot produce zeros. Although the mixture and hurdle models might give similar results, particularly if the mean abundance is high (Welsh et al. 1996, Wenger and Freeman 2008), it can be argued that truncating the count-generating distribution in this way is somewhat "artificial." There may be some practical advantages to eliminating zeros from the count model, such as orthogonality of the two distributions (Welsh et al. 1996, Cunningham and Lindenmayer 2005), but it is difficult to justify why zeros should be precluded from occurring in any stochastic ecological sampling process that produces counts of organisms. An ecological sampling process, even in the absence of excess zeros, could easily generate a zero if the mean is sufficiently low and/or the variance is sufficiently high. Furthermore, the mixture-model approach has the advantage of operating with conventional, parametric distributions.

Another advantage of the zero-inflated mixture model is that it can provide direct insights into the nature of the occupancy–abundance relationship for excess zeros. Under standard count distributions, such as the Poisson or negative binomial, the rate of occupancy is a fixed positive function of the mean abundance (λ) and, potentially, a dispersion parameter. Under a zero-inflated mixture model, a zero can arise from either the base count distribution (base zeros) or from the component of the model that allows for zero inflation (excess zeros). The base zeros will follow the occupancy–abundance relationship implicit in the base distribution. On the other hand, excess zeros may or may not be related to abundance, depending in part on the structure of the model for zero inflation. For the models described herein, model 2 asserts no relationship between excess zeros and abundance, and model 3 assumes no structural relationship, although one might be produced indirectly. Only model 4 directly relates excess zeros to the conditional mean abundance (λ). This parsimonious method allows for the excess zeros to follow a similar relationship with abundance to that of the base zeros. It also has the flexibility to allow for alternative forms of the relationship, if so required (Appendix C).

Conclusion

Many authors (e.g., Martin et al. 2005, Potts and Elith 2006) have stressed that failure to account for overdispersion and zero inflation in a statistical model can result in inaccurate point or interval estimation of parameters, which may then lead to spurious conclusions. Therefore, it is wise to put some effort into considering alternative model structures when standard models fail to provide an adequate representation of distribution of the data. For models that require zero inflation, the general approach of modeling the probability of an excess zero as a function of the conditional mean of the count values may be used to great advantage in cases where they are indeed related. Clearly, if no such relationship is present, other models for π should be considered instead. The general linear structure for the link between π and λ proposed here can produce a wide variety of useful relationships (Appendix C). Nevertheless, we encourage further research to evaluate alternative forms of this relationship in predictive models for real data, including alternative link functions for binary data (such as the complementary log–log or probit functions) and other forms of linkage suggested by empirical analyses of the occupancy–abundance relationship (He and Gaston 2003, Sileshi et al. 2009). We conclude that the incorporation of occupancy–abundance relationships into models of zero inflation provides a key tool to develop more accurate and parsimonious models of ecological count data.

Acknowledgments

The authors thank the Department of Conservation (DOC Inv 4238) and Massey University (specifically the Institute for Information and Mathematical Sciences and the New Zealand Institute for Advanced Study) for financial and logistic support. We also thank Trevor Willis and DOC for the provision of data. This article was improved by comments from an anonymous reviewer, Seth Wenger, and Robert Dorazio.

Literature Cited

Bliss, C. I., and R. A. Fisher. 1953. Fitting the negative binomial distribution to biological data. Biometrics 9:176–200.

Borregaard, M. K., and C. Rahbek. 2010. Causality of the relationship between geographic distribution and species abundance. Quarterly Review of Biology 85:3–25.

Brooks, S. P., and A. Gelman. 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7:434–455.

Brown, J. H. 1984. On the relationship between abundance and distribution of species. American Naturalist 124:255–279.

Clapham, A. R. 1936. Over-dispersion in grassland communities and the use of statistical methods in plant ecology. Journal of Ecology 24:232–251.

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

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian data analysis. Second edition. CRC Press, Boca Raton, Florida, USA.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7:457–472.

Ghosh, S. K., P. Mukhopadhyay, and J.-C. Lu. 2006. Bayesian analysis of zero-inflated regression models. Journal of Statistical Planning and Inference 136:1360–1375.

Hartley, S. 1998. A positive relationship between local abundance and regional occupancy is almost inevitable (but not all positive relationships are the same). Journal of Animal Ecology 67:992–994.

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

Heilbron, D. C. 1994. Zero-altered and other regression models for count data with added zeros. Biometrical Journal 36:531–547.

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

Lawson, A., and A. Clark. 2002. Comment on article by Spiegelhalter et al. Journal of the Royal Statistical Society B 64:624–625.

Link, W. A., and R. J. Barker. 2009. Bayesian inference: with ecological applications. Academic Press, Boston, Massachusetts, USA.

Liu, H., and K.-S. Chan. 2010. Introducing COZIGAM: An R package for unconstrained and constrained zero-inflated generalized additive model analysis. Journal of Statistical Software 35(11).

Lunn, D., D. Spiegelhalter, A. Thomas, and N. Best. 2009. The BUGS project: evolution, critique and future directions. Statistics in Medicine 28:3049–3067.

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

McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. Chapman and Hall, London, UK.

Millar, R. B. 2009. Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' factors. Biometrics 65:962–969.

Neyman, J. 1939. On a new class of "contagious" distributions, applicable in entomology and bacteriology. Annals of Mathematical Statistics 10:35–57.

Nielsen, S. E., C. J. Johnson, D. C. Heard, and M. S. Boyce. 2005. Can models of presence–absence be used to scale abundance? Two case studies considering extremes in life history. Ecography 28:197–208.

Potts, J. M., and J. Elith. 2006. Comparing species abundance models. Ecological Modelling 199:153–163.

R Development Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org

Sileshi, G., G. Hailu, and G. I. Nyadzi. 2009. Traditional occupancy–abundance models are inadequate for zero-inflated ecological count data. Ecological Modelling 220:1764–1775.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society B 64:583–639.

Sturtz, S., U. Ligges, and A. Gelman. 2005. R2WinBUGS: a package for running WinBUGS from R. Journal of Statistical Software 12:1–16.

Ver Hoef, J. M., and P. L. Boveng. 2007. Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? Ecology 88:2766–2772.

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

Wenger, S. J., and M. C. Freeman. 2008. Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. Ecology 89:2953–2959.

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

Willis, T. J., and R. C. Babcock. 2000. A baited underwater video system for the determination of relative density of carnivorous reef fish. Marine and Freshwater Research 51:755–763.

Willis, T. J., R. B. Millar, and R. C. Babcock. 2003. Protection of exploited fish in temperate regions: High density and biomass of snapper *Pagrus auratus* (Sparidae) in northern New Zealand marine reserves. Journal of Applied Ecology 40:214–227.

Wright, D. H. 1991. Correlations between incidence and abundance are expected by chance. Journal of Biogeography 18:463–466.

Supplemental Material

Appendix A

Formal description of a Bayesian mixed-effects model in which the probability of an excess zero is functionally linked to mean counts of snapper from baited underwater video deployments in a marine reserve monitoring program (*Ecological Archives* E093-237-A1).

Appendix B

A table presenting summary statistics for the posterior distributions of estimated parameters (*Ecological Archives* E093-237-A2).

Appendix C

A demonstration of various potential relationships between the mean of the count distribution (λ) and the probability of an excess zero (π) under the general form of the linked model, $\text{logit}(\pi) = \gamma_0 + \gamma_1 \log(\lambda)$ (*Ecological Archives* E093-237-A3).

Supplement

R and OpenBUGS code and data for fitting a linked zero-inflated negative binomial model applied to counts of legal-sized snapper from a marine reserve monitoring program (*Ecological Archives* E093-237-S1).