A General Approach to Modeling and Analysis of Species Abundance Data With Extra Zeros

H. M. PODLICH, M. J. FADDY, and G. K. SMYTH

A general method for the analysis of ecological count data with extra zeros is presented using a Markov birth process representation of discrete distributions. The method uses a nonparametric formulation of the birth process to model the residual variation and therefore allows the data to play a greater role in determining an appropriate distribution. This enables a more critical assessment of covariate effects and more accurate predictions to be made. The approach is also presented as a useful diagnostic tool for suggesting appropriate parametric models or verifying standard models. As an illustrative example, data describing abundance of a species of possum from the montane ash forests of the central highlands of Victoria, southeast Australia, is considered.

Key Words: Covariate effects; Extended Poisson process model; Penalized likelihood; Prediction.

1. INTRODUCTION

Ecological count data describing the abundance of species frequently contain many zeros, particularly when the species is rare and endangered. These data usually consist of the number of sightings of the species in a defined area and measurements of corresponding habitat variables on that site. Accurate assessment of the significance of each of these habitat covariates is important to enable effective monitoring and management of the species. If the significance of covariates is over- or understated, then poorer predictions of the mean abundance of animals at a site may follow. Therefore, statistical models must take account of the extra zeros and accurately account for any residual variation.

Standard approaches proceed by adding a probability at zero to truncated standard distributions such as a Poisson or negative binomial distribution. The resulting models are

H. M. Podlich is from the School of Land and Food Sciences, The University of Queensland, Australia (E-mail: h.podlich@mailbox.uq.edu.au). M. J. Faddy is a Professsor at the School of Mathematics and Statistics, The University of Birmingham, U.K. G. K. Smyth is Senior Research Scientist at the Walter and Eliza Hall Institute for Medical Research, Melbourne, Australia.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 7, Number 3, Pages 1–11
DOI: 10.1198/108571102221

referred to as added zero or extra zero Poisson and negative binomial models, respectively. More recently, Welsh, Cunningham, Donnelly, and Lindenmayer (1996) used the term conditional to describe such models. There are also zero-inflated models (Lambert 1992), which model the response as a mixture of a Bernoulli distribution and a Poisson or negative binomial distribution.

Recently, Faddy (1998) proposed a new method for the analysis of count data with extra zeros using a Markov birth process representation of discrete distributions with non-negative support. These models allow for control of the residual dispersion as covariates are incorporated into the model and therefore allow for a more critical assessment of their effects. The appropriate distribution describing the residual variation was determined using a parametric formulation. In this article, we extend this approach by using penalized likelihood to estimate an appropriate formulation nonparametrically, thereby allowing the data to play a greater role in describing the residual distribution. The resulting models can be used in their own right for assessing covariate effects and making predictions or as a useful exploratory tool for suggesting parametric formulations such as that proposed in Faddy (1998).

In Section 2, the model framework will be described, and in Section 3, model fitting is discussed. The approach will then be applied to some data describing abundance of the Leadbeater's possum species in Victoria, Australia, in Section 4. Comparisons with results obtained from the completely parametric approach of Faddy (1998) and the more usual extra zeros models will be of interest.

2. MODELING

The model framework arises from noting that any distribution $\{p_n, n = 0, 1, \ldots, i\}$ can be represented as the distribution of a Markov birth process $\{X(t), t \geq 0\}$ with $X(0) = 0$ and transition rates $\{\lambda_n,\ n = 0, 1, \ldots, i\}$ at some arbitrary time, $t = 1$ say. Such a representation is unique (Faddy 1997), i.e., there is a one-to-one relationship between a distribution and transition rate sequence satisfying the system of differential equations (Cox and Miller 1965, chap. 4)

$$\begin{aligned} {p_0}'(t) &= -\lambda_0 p_0(t) \quad \text{with} \quad p_0(0) = 1 \\ {p_n}'(t) &= -\lambda_n p_n(t) + \lambda_{n-1} p_{n-1}(t) \quad \text{with} \quad p_n(0) = 0,\ n = 1, \ldots, i. \end{aligned} \tag{2.1}$$

The underlying birth process is an extension of the simple Poisson process for which events occur at a constant rate $\lambda_n = \lambda$. Therefore, distributions or models constructed from transition rates that depend on n, the number of accumulating events, can be referred to as extended Poisson process models (EPPMs).

This construction can be easily modified to allow for an excess of zero counts. From the solution to (2.1), note that $p_0(t) = e^{-\lambda_0 t}$, where we take $t = 1$ without any loss of generality. The probability at zero is therefore determined from the value of λ_0, and extra zeros could be accommodated by modeling the transition rate at $n = 0$ separately from the other transition rates. The resulting models thus have more similarities with the extra zeros

Poisson and negative binomial models than Lambert's zero-inflated models.

Suppose the ith observation, y_i, is observed with corresponding vector of covariates $\mathbf{x}_i^{\mathrm{T}}$. Let $\lambda_i(n), n = 0, 1, \ldots, y_i$, denote the transition rate sequence that determines p_{y_i}. Then the proposed model for the transition rates is of the form

$$\lambda_i(n) = \begin{cases} \lambda_{i0} & n = 0 \\ \lambda_{i1} h(n) & n \geq 1, \end{cases} \tag{2.2}$$

where $\lambda_{i0}, \lambda_{i1} > 0$ and $h(n)$ is to be estimated nonparametrically from the data. To allow for the assessment of covariate effects, the parameters λ_{i0} and λ_{i1} in (2.2) can be modeled as log-linear functions of the covariates,

$$\begin{aligned} \lambda_{i0} &= \exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0) \\ \lambda_{i1} &= \exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_1). \end{aligned}$$

For ease of notation in discussions that follow, we drop the ith case dependence and refer to these parameters as λ_0 and λ_1. When these parameters are modeled as functions of covariates, we use the term semiparametric to describe such a model since the covariate effects are parametric; however, the distributional form is to be estimated nonparametrically. Also note, from $p_0 = e^{-\lambda_0}$ and with a log-linear form for λ_0, these extra zeros models correspond to the use of a complementary log-log link for $1 - p_0$.

It is the form of $h(n)$ in (2.2), referred to as the n-dependence form, that determines the dispersion properties of the model. For example, $\lambda_0 = \lambda_1 h(1)$ and $h(n)$ constant corresponds to the ordinary Poisson model, while $\lambda_0 = \lambda_1[2h(1) - h(2)]$ and $h(n)$ linear increasing corresponds to the negative binomial model. Therefore, a model with $\lambda_0 < \lambda_1 h(1)$ and $h(n)$ constant has similarities with an extra zeros Poisson model, while $h(n)$ linear increasing and $\lambda_0 < \lambda_1[2h(1) - h(2)]$ has similarities with an extra zeros negative binomial model. Numerical experiments suggest that a model with $h(n)$ decreasing allows for less variation than one with $h(n)$ constant. Also, a model with $h(n)$ concave increasing allows for variation between $h(n)$ constant and linear increasing, while a convex increasing $h(n)$ allows for more variation than the linear increasing case. For example, Figure 1 shows probability distributions corresponding to $h(n)$ decreasing (–), constant (-··-), concave increasing ($\cdots$), linear increasing, (- - -) and convex increasing (-·· ··-). Although the mean is the same for each probability distribution (as is the probability of zero), increasing variation is apparent from $h(n)$ decreasing to $h(n)$ convex increasing. Hence, different forms for $h(n)$ will result in different distributions for the residual dispersion, allowing models applicable to a broad range of extra zeros data. Further, because we do not assume any particular form for $h(n)$, the data are allowed to determine an appropriate functional form.

Faddy (1998) assumed $h(n)$ took the particular form

$$h(n) = n^c \qquad c \leq 1. \tag{2.3}$$

The models (2.3) provide a flexible class of monotonic relations for the n-dependence: $c = 0$ constant, $c = 1$ linear increasing, $0 < c < 1$ concave increasing, and $c < 0$ decreasing. [Unfortunately, this model does not admit $c > 1$ because a dishonest distribution results (see

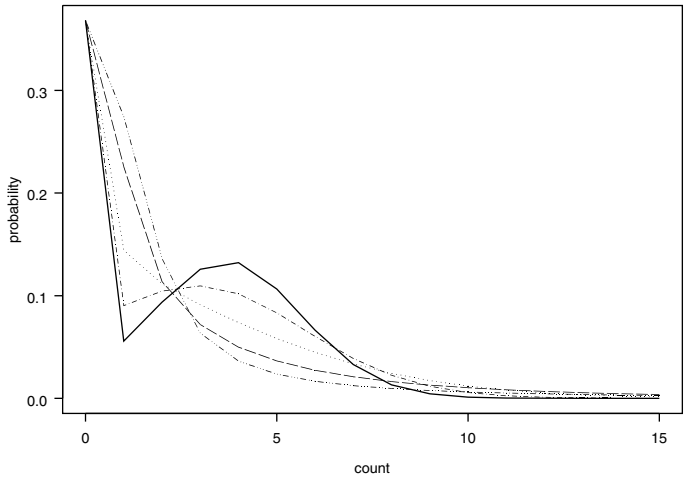

Figure 1. Probability Distributions With Equal Mean and Probability of Zero for Various $h(n)$.

Faddy 1997).] The more general approach here would provide a useful diagnostic tool for investigating the appropriateness of such monotonic forms. Also, by smoothing the entire sequence $h(n)$ for $n \geq 0$, the need for modeling the zero probability separately as in (2.2) can be investigated.

The method proposed for nonparametrically estimating the n-dependence form, $h(n)$, is based on the roughness penalty approach (see Green and Silverman 1994). Here a set of knots are defined over the range of nonzero data, and the values of the function $h(\cdot)$ at these knots are computed by maximizing an adjusted log-likelihood function, known as a penalized log-likelihood, that penalizes rough forms for $h(n)$.

Let $\mathbf{h} = (h_1\ h_2\ \cdots\ h_q)$ denote the values of the function $h(\cdot)$ at the q knots. Also let $\mathbf{y}$ denote the data vector with elements y_i and $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, the vector of covariate coefficients for the zero and nonzero counts, respectively. Then the penalized log-likelihood function,

$$\ell_p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{h}) = \ell(\boldsymbol{\beta}, \mathbf{h} \mid \mathbf{y}) - \alpha P(\mathbf{h}), \qquad \alpha > 0, \tag{2.4}$$

is maximized over $\boldsymbol{\beta}$ and $\mathbf{h}$, where $\ell(\boldsymbol{\beta}, \mathbf{h} \mid \mathbf{y}) = \Sigma_i \log(p_{y_i})$ denotes the ordinary log-likelihood, $P(\mathbf{h})$ is a penalty function measuring the roughness of the n-dependent form $\mathbf{h}$, and α is a smoothing parameter. The smoothing parameter controls the trade-off between goodness of fit, as measured by the log-likelihood, and smoothness of the n-dependent form, as measured by the roughness penalty. The larger the value for α, the smoother the form for $h(n)$.

We propose the use of a penalty function that takes account of the discrete nature of our smoothing problem,

$$P(\mathbf{h}) = \sum_{j=2}^{q-1} (h_{j-1} - 2h_j + h_{j+1})^2 + \gamma \sum_{j=2}^{q} (h_j - h_{j-1})^2, \qquad \gamma > 0,$$

i.e., the penalty function comprises the summed second differences squared and some multiple of the summed first differences squared. We choose γ small (e.g., 10^{-4}), resulting in a zero penalty when the function $h(\cdot)$ is constant, so that the limiting $\alpha \to \infty$ n-dependence form will be this constant form using a single degree of freedom.

Selection of an appropriate smoothing parameter value, α, can be made by displaying $h(n)$ graphically for a range of smoothing parameters and visually comparing the forms obtained in relation to the log-likelihoods. Such a subjective approach to selection of α is often applied and avoids the need to estimate the equivalent degrees of freedom for the nonparametric form.

3. MODEL FITTING

To form the log-likelihood, $\ell(\boldsymbol{\beta}, \mathbf{h} \mid \mathbf{y})$ in (2.4), probabilities p_{y_i} are computed by solving the Chapman–Kolmogorov system of differential equations (2.1) given in Section 2. The solution for each observation y_i can be expressed in terms of the matrix exponential of the Q-matrix of transition rates,

$$\mathbf{Q} = \begin{pmatrix} -\lambda(0) & \lambda(0) & 0 & \cdots & 0 & 0 \\ 0 & -\lambda(1) & \lambda(1) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda(y_i - 1) & \lambda(y_i - 1) \\ 0 & 0 & 0 & \cdots & 0 & -\lambda(y_i) \end{pmatrix},$$

with p_{y_i}, the probability of obtaining the count y_i, being the last element of the probability vector satisfying

$$(p_0\ p_1\ \cdots\ p_{y_i-1}\ p_{y_i}) = (1\ 0\ \cdots\ 0\ 0)\exp(\mathbf{Q}). \tag{3.1}$$

The matrix-exponential vector operation (3.1) can be performed, for example, using the software developed by Sidje (1998). In Podlich, Faddy, and Smyth (1999), it is shown how derivatives of the log-likelihood can also be computed by extending the Q-matrix. These are not only useful for iterative maximization of the log-likelihood function but they also enable the computation of asymptotic standard errors of parameter estimates and therefore routine application of inferential techniques such as Wald tests. However, for large counts or large numbers of parameters, computation can become expensive since the size of the matrix to be exponentiated is large. In this situation, probabilities and derivatives can be computed from a saddlepoint approximation (Daniels 1982),

$$\hat{p}_{y_i} = \frac{\displaystyle\prod_{j=0}^{y_i-1} \lambda(j) e^{-\tilde{\theta}}}{\displaystyle\prod_{j=0}^{y_i} (\lambda(j) - \tilde{\theta}) \left\{ 2\pi \sum_{j=0}^{y_i} \frac{1}{(\lambda(j) - \tilde{\theta})^2} \right\}^{1/2}} \left\{ 1 + \frac{1}{8}\rho_4 - \frac{5}{24}\rho_3^2 \right\},$$

with $\tilde{\theta}$ satisfying

$$1 = \sum_{j=0}^{y_i} \frac{1}{\lambda(j) - \tilde{\theta}}$$

and with

$$\rho_r = \frac{(r-1)! \sum_{j=0}^{y_i} \frac{1}{(\lambda(j) - \tilde{\theta})^r}}{\left\{ \sum_{j=0}^{y_i} \frac{1}{(\lambda(j) - \tilde{\theta})^2} \right\}^{r/2}}.$$

All computations involving nonparametric estimation are only exact for zero counts, with the saddlepoint approximation being used to compute probabilities and derivatives for the nonzero counts. Software for these purposes has been developed in the popular statistical package S-Plus and is available from http://www.maths.uq.edu.au/~hmp/.

4. LEADBEATER'S POSSUM DATA

As an illustrative example of the semiparametric approach, we revisit the data considered in Faddy (1998).

The data consist of the numbers of Leadbeater's possums (*Gymnobelideus leadbeateri*) observed in a sample of 151 three-hectare sites across Central Victoria, Australia. Covariates describing various habitat characteristics were measured on each site. Those believed to be relevant to the abundance of this species (Welsh et al. 1996) are

lstags	$\log_e$(no. of trees with hollows + 1),
age	forest age,
baa	basal area of *Acacia* species on site,
slope	slope of the site,
aspect	aspect of the site (SW-NW, NW-NE, NE-SE, SE-SW)
bark	score for degree of decorticating or peeling bark,
shrubs	number of shrubs on the site.

All of these covariates are quantitative with the exception of the *aspect* covariate, which is a factor variable at four levels. Variables such as *lstags, baa,* and *bark* are believed to be important because they are directly related to the possums' survival, corresponding to the possums' shelter (hollow logs) and food source (grubs, etc., living in trees and bark).

Table 1 gives the actual frequencies of observed counts of Leadbeater's possums sampled over the 151 sites. Note the high proportion of zeros, with over 60% of the data in this category.

From the above frequency distribution, we are able to construct the empirical sequence of $\hat{\lambda}$'s that is the transition rate sequence giving rise to the observed distribution of counts.

Table 1. Frequencies of Leadbeater's Possum Counts

n	0	1	2	3	4	5	6	7	8	9	10
freq_n	95	9	10	12	8	9	0	4	1	1	2

This can be constructed by taking $\hat{\lambda}(0) = -\log(\hat{p}_0) = -\log(95/151)$ and successively solving numerically for $\hat{\lambda}(n+1)$, $n = 0, 1, \ldots, 9$, using the matrix-exponential vector operation (3.1). The resulting sequence is shown in Figure 2 with $\pm$standard errors indicated by the bars. The value for $\hat{\lambda}(6)$ is infinite since there were no observed counts of 6 and is therefore not shown, while the value for $\hat{\lambda}(10)$ is exactly zero as 10 represents the maximum count. Note the increase in the transition rates from $n = 0$ to $n = 1$, reflecting the large number of zeros, and it appears that the rates are fairly constant thereafter.

It is possible to apply our methodology to smooth these empirical rates and, through the Q-matrix exponential, estimate the observed frequency distribution. However, because we are mainly interested in the model with covariate effects, we do not present these results in detail but simply note that they do suggest that a constant rate sequence over $n \geq 1$ would be appropriate. On introducing covariates into the model, we might expect to see a decreasing transition rate sequence, as there would be less residual dispersion relative to the model with no covariates.

Figure 3 shows the fitted smoothed n-dependent forms, $h(n)$, obtained for various values of the smoothing parameter α when all of the covariate effects are included in the model. Knots have been defined at each integer between $n = 1$ and $n = 10$, the maximum count, and the penalized log-likelihood maximized over the values of $h(n)$ at these knots as well as the covariate parameters. Placing a knot at every integer up to the maximum count may be computationally prohibitive if much larger counts are present. In this situation, a smaller number of knots can be defined over the range of the data and the 'in between' values of $h(n)$ determined using a cubic spline interpolant.

The plots in Figure 3 do indeed show decreasing sequences of $h(n)$ for $n \geq 1$ as the smoothing parameter gets larger, with the bump in the transition rates from $n = 5$ to $n = 6$

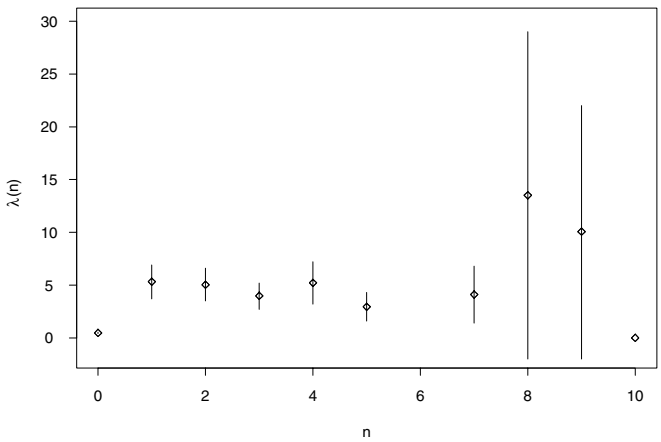

Figure 2. Empirical Transition Rate Sequence (Leadbeater's Possum Counts).

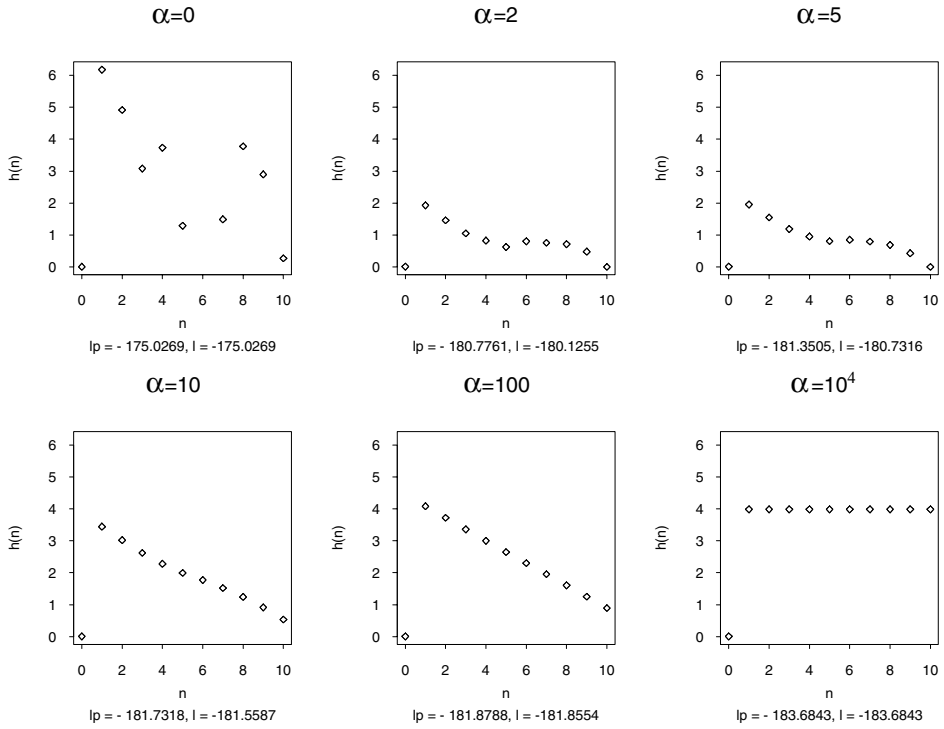

Figure 3. Nonparametric n-Dependent Form With Covariate Effects (Leadbeater's Possum Counts).

at smaller values of α a result of the frequency of the count of 6 being zero. On comparing the n-dependent forms and log-likelihoods obtained from the range of smoothing parameter values, we are led to choose a linear decreasing form for $h(n)$ since the approximate linear decreasing form ($\alpha = 100$) represents a marginally significant improvement in log-likelihood over the constant ($\alpha = 10^4$) form for the extra 1 d.f. ($2 \times \Delta\ell = 2.90$), but smaller values of α do not represent any worthwhile improvements in log-likelihood for the increase in complexity of the n-dependent forms. An exact linear decreasing form can be obtained by taking $\alpha = 10^5$ and $\gamma = 0$, giving $\ell_p = \ell = -185.64$, since linear $h(n)$ results in the summed second differences squared in the penalty being zero.

To assess the significance of the habitat variables, those covariate coefficients with standard errors exceeding or close to the estimate were removed from the model. Standard errors can be computed using the Fisher information matrix from the ordinary log-likelihood without the penalty component. The $bark$ covariate effects in λ_0 and λ_1 were then marginally significant on their own; however, the change in log-likelihoods on removing both suggests that they should be retained ($2 \times \Delta\ell = 6.97$).

The estimates obtained for $h(n)$ with this reduced set of covariates do not change greatly from those in Figure 3 since the residual variation has not greatly changed when only the significant covariates are retained. The significant covariate effect estimates and their asymptotic standard errors from this model fit are given in Table 2.

Table 2. Significant Covariate Effect Estimates and Their Asymptotic Standard Errors

Coef	*Est. (SE)*
$lstags_0$	0.6614 (.2012)
baa_0	0.0743 (.0172)
$bark_0$	0.0502 (.0314)
$lstags_1$	0.3776 (.1139)
$slope_1$	−0.0441 (.0152)
$bark_1$	0.0329 (.0184)
$shrubs_1$	−0.1303 (.0377)

The resulting model can be used to make predictions of abundances given a set of habitat characteristics. Calculation of predicted values requires extrapolating the form for $h(n)$ beyond the data range, where there is no information to construct any estimates. Here we will simply take the transition rates to be constant beyond the range of the data. Means are then computed directly from the probability distribution (3.1) for suitably large y_i corresponding to the fitted transition rate sequence since no explicit expression for means is available. Standard errors for these predictions can be computed using the delta method.

Figure 4 shows estimated mean abundances and ±standard errors computed as a function of the most significant habitat variable $stags$, the number of trees with hollows. We have chosen the other values for the habitat characteristics from the data, one at a low value of the bark covariate and the other at a high value of the bark covariate. Comparisons are made here between estimates obtained from the chosen semiparametric model, the parametric model proposed in Faddy (1998), and the extra zeros Poisson model proposed in Welsh et al. (1996).

Our estimates are in good agreement with those from Faddy (1998), with the predictions from the parametric and semiparametric models almost indistinguishable in Figure 4. The analysis here therefore supports the parametric form used in Faddy (1998) and in particular verifies the assumption of monotonic n-dependence for $n \geq 1$. For other values of the

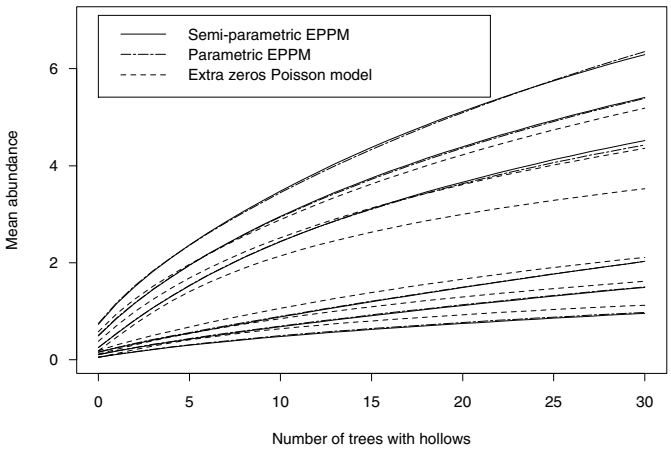

Figure 4. Predicted Abundances and ±Standard Errors.

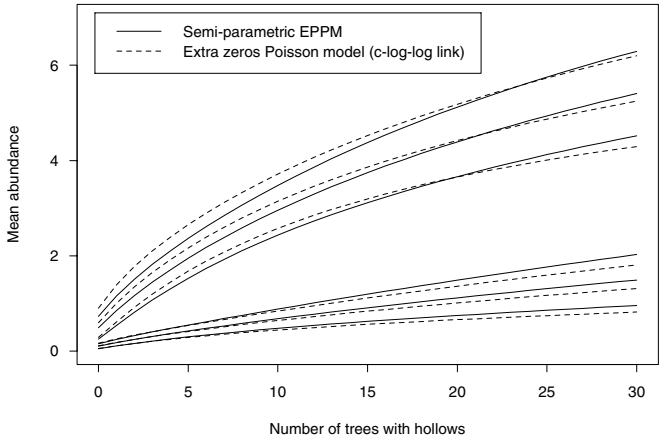

Figure 5. Predicted Abundances and ±Standard Errors.

smoothing parameter, α, the size of the standard errors in relation to the predictions becomes smaller as α decreases, reflecting the reduction in residual dispersion. Here, since the smaller positive values for α give a similar n-dependent form and log-likelihood to the linear decreasing form chosen above, the resulting predictions and standard errors are also similar. However, the constant n-dependent form results in slightly smaller predictions and larger standard errors relative to those shown.

Compared with the predictions from the extended Poisson process models in Figure 4, the extra zeros Poisson model tends to overestimate the abundances at the low value of the $bark$ covariate and underestimate the abundance at the high value of the $bark$ covariate. This latter model, using a logistic link for the probability at zero, resulted in a maximum log-likelihood $\ell = -187.80$, with the $bark$ covariate only included in the model for the equivalent of λ_1. If a complementary log-log link for the probability at zero is used instead, a lower log-likelihood of $\ell = -185.65$ results, with the $bark$ covariate effect for the equivalent of λ_0 now significant. Mean abundances and standard errors, shown in Figure 5, suggest predictions are now in closer agreement with the results from the extra zeros extended Poisson process models with similar covariate assessment.

5. CONCLUSION

The models proposed in this article offer a flexible framework for the analysis of count data with extra zeros, which frequently arise in studies of species abundance. It is the nonparametric formulation that allows the data to guide the appropriate model choice and therefore make for more reliable covariate assessment and model predictions. Although more standard extra zeros models are not explicitly special cases of those developed here, models with similar dispersion properties can be constructed. It is the greater generality of the modeling framework developed here that justifies this approach.

The use of the approach as an exploratory tool is particularly appealing. Through

graphical model selection, it is possible to obtain a good feel for whether there is a need to account for extra zeros in the first instance and then what model might be appropriate for a particular data set. This graphical analysis may serve to justify the use of a completely parametric model specification such as the flexible monotonic parametric forms of Faddy (1998) or more standard extra zeros models.

ACKNOWLEDGMENTS

Parts of this work were completed while the first author was a Ph.D. student in the Department of Mathematics at The University of Queensland.

[Received September 2000. Accepted August 2001.]

REFERENCES

Cox, D. R., and Miller, H. D. (1965), *The Theory of Stochastic Processes*, London: Methuen.

Daniels, H. E. (1982), "The Saddlepoint Approximation for a General Birth Process," *Journal of Applied Probability*, 19, 20–38.

Faddy, M. J. (1997), "Extended Poisson Process Modelling and Analysis of Count Data," *Biometrical Journal*, 39, 431–440.

Faddy, M. J. (1998), "Stochastic Models for Analysis of Species Abundance Data," in *Statistics in Ecology and Environmental Monitoring* (Vol. 2), eds. D. J. Fletcher, L. Kavalieris, and B. F. J. Manly, Dunedin: University of Otago Press, pp. 33–40.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, London: Chapman and Hall.

Lambert, D. (1992), "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing," *Technometrics*, 34, 1–14.

Podlich, H. M., Faddy, M. J., and Smyth, G. K. (1999), "Likelihood Computations for Extended Poisson Process Models," *InterStat*, September No. 1, 15 pp.

Sidje, R. B. (1998), "EXPOKIT: Software Package for Computing Matrix Exponentials," *ACM Transactions on Mathematical Software*, 24, 130–156.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F., and Lindenmayer, D. B. (1996), "Modelling the Abundance of Rare Species: Statistical Models for Counts With Extra Zeros," *Ecological Modelling*, 88, 297–308.