ELSEVIER

Ecological Modelling 157 (2002) 179–188

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Generalized additive modelling and zero inflated count data

Simon C. Barry [a,*], A.H. Welsh [b]

[a] *Bureau of Rural Sciences, Agriculture, Fisheries and Forestry-Australia, PO Box E11, Kingston, ACT 2604, Australia*
[b] *Faculty of Mathematical Studies, University of Southampton, Southampton SO17 1BJ, UK*

Abstract

This paper describes a flexible method for modelling zero inflated count data which are typically found when trying to model and predict species distributions. Zero inflated data are defined as data that has a larger proportion of zeros than expected from pure count (Poisson) data. The standard methodology is to model the data in two steps, first modelling the association between the presence and absence of a species and the available covariates and second, modelling the relationship between abundance and the covariates, conditional on the organism being present. The approach in this paper extends previous work to incorporate the use of Generalized Additive Models (GAM) in the modelling steps. The paper develops the link and variance functions needed for the use of GAM with zero inflated data. It then demonstrates the performance of the models using data on stem counts of *Eucalyptus mannifera* in a region of South East Australia.

Keywords: Abundance models; Statistical models; Count data; Prediction; Distribution modelling; Zero inflated data; Generalized additive models

1. Introduction

Models of species abundance data are potentially important tools for the management of wildlife populations. If the relationship between the abundance of organisms and known geographical and environmental variables can be described, it can be used to ensure that conservation criteria are met. A recent review of available techniques is given by Guisan and Zimmermann (2000).

Describing the empirical relationship between a response variable (abundance) and covariates is the aim of statistical modelling. In a statistical model we hypothesize plausible relationships between the variables and then use the data to verify and quantify these relationships. It is thus important that the observed features of the data are captured effectively by our model.

A distinctive feature of data collected on the abundance (counts) of organisms is that it is often zero inflated. By this we mean that the data contains more zeros than might be predicted from standard error models used with generalized linear models (GLMs) (McCullagh and Nelder,

* Corresponding author
E-mail address: simon.barry@brs.gov.au (S.C. Barry).

0304-3800/02/$ - see front matter © 2002 Published by Elsevier Science B.V.
PII: S0304-3800(02)00194-1

1989). If we ignore this feature of the data and apply standard Poisson error models then problems with inference can occur. These problems occur because the Poisson assumption is not an adequate approximation to the conditional distribution of the data.

This problem of extra zeros is very common in field collected data, and is a particular form of overdispersion (McCullagh and Nelder, 1989). While simple assumptions might lead to the expectation of data with a Poisson distribution, the complexity of real world processes and the limitations of the data we typically collect mean that simple theoretical models rarely hold.

The problem of extra zeros has been previously noted in the ecological literature. An early example of an article considering this problem is Austin and Cunningham (1981) who modelled stem counts in Eucalypt species. More recently Leathwick and Austin (2001), while modelling competitive interactions between tree species, have commented on the lack of tractable techniques to flexibly model zero inflated count data.

Mullahy (1986), Heilbron (1994), Welsh et al. (1996) have considered empirically modelling such data in general. They recommend modelling the data in two steps. First, they model the presence–absence component of the data via a GLM, usually with the logistic link. They then model the observed abundance, conditional on the response being greater than 0. For this second stage they consider both the truncated Poisson and if over-dispersion is present, the truncated negative binomial distribution. Optionally, in the final stage of the analysis the two models can be combined to produce marginal predictions.

The two stage approach has a number of advantages. First, by allowing separate models (terms) for the two components, a richer class of potential abundance distributions can be modelled compared with simply assuming the data is approximately Poisson. The second advantage is that the two models are directly interpretable in terms of the empirical features of the data. The third advantage is that the approach will better model the variability seen in this data. This will ensure that the information in the data is used appropriately.

The approach of Welsh et al. (1996) assumes a fully parametric specification of the model. While there is much flexibility available through this formulation, with ecological data our responses are often complicated functions of the explanatory variables. We need a method that can routinely attempt to model the observed features of the data. This can then be used as either a final model or as an exploratory step to suggest appropriate parametric models. Of course it is imperative that all models be assessed rigorously at the end of any model building exercise.

In this paper we extend the arguments of Welsh et al. (1996) to provide greater flexibility in potential mean models. Specifically, we will again use the two step approach, but will now model each component as a Generalized Additive Model (GAM). We then illustrate the use of the techniques by analyzing data on stem counts for a Eucalypt species in South East Australia previously modelled by Austin and Cunningham (1981).

2. Extending GAMs to zero truncated data

2.1. Introduction

The modelling of zero inflated data is reviewed in Ridout et al. (1998). In summary, as discussed by Welsh et al. (1996), there are several main approaches in practical use. First, we can ignore the zero inflation, which means that inferences will be incorrect and parameter estimates biased. Second, we can assume an underlying mixture of Poisson and Bernoulli variables and model this mixture process. While in certain applications this model is appropriate, the mixture assumption is often unrealistic as there are no clear means of logically separating the zeros into different classes. In the third approach we can model the Bernoulli process and then model abundance conditionally on the organism being present. This is the approach taken in Welsh et al. (1996). This approach seeks to describe the empirical relationship between abundance and the covariates, and it is argued that this provides a more flexible and

interpretable empirical description of zero inflated data.

In the following we will consider the case where we have a sample $(y_1, x_1), \ldots, (y_n, x_n)$, where y_i is the response for the ith observation and x_i are the associated covariates. We will assume that responses are conditionally independent of each other and, therefore, that we can construct the likelihood via the product of the components. The Zero Inflated Poisson (ZIP) approach to modelling such data is as follows. First we define a derived presence–absence variable y_i^* as follows.

$$y_i^* = \begin{cases} 1 & y_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

We model y_i^* in terms of the covariates x_i. This model is the presence–absence component of the analysis.

In the second stage we model the abundance conditional on the response being greater than zero. To do this we subset the data to include only those cases where $y_i > 0$. We then try to describe the variation of y_i in this subset in terms of the covariates. The innovation of this paper is to explore the use of a GAM for the model in the two cases.

2.2. GAM models

GAM models are a flexible class of models that can be used either as the main analysis tool (for example Leathwick and Austin, 2001) or as an exploratory tool before constructing a more formal analysis.

The modelling of data via a GAM is relatively standard and will not be fully discussed here. The interested reader is referred to Hastie and Tibshirani (1990), Chambers and Hastie (1992) We will instead briefly review the major components of a GAM and the statistical functions that are needed to specify the model. We describe the model for the ith observation and for clarity we will suppress the i subscript.

To model data via a GAM we need to specify several components. We first have a linear predictor. In the parametric case we use:

$$\eta = x\beta$$

where, η is termed the linear predictor, x is a row vector of covariates and β is a column vector of the associated regression parameters. The extension of the GAM is to consider linear predictors of the form:

$$\eta = z\beta + s(x_1) + \ldots + s(x_k)$$

where, the $s()$ are smooth functions (not necessarily the same at each appearance) and z is a subset of x. It is these smooth terms which give the flexibility to the model. The choice of linear predictor is the major difference between the GAM and GLM.

To implement a GAM model we need to define three functions in addition to the linear predictor. These are the link function, the variance function and the deviance function.

The link function $g()$ is defined as:

$$g(\mu) = \eta,$$

where, μ is the mean. This function relates the mean response to the linear predictor. The link function defines the scale upon which the model terms are additive.

The next component to specify is a variance function, $V(\mu)$, which is parameterized in terms of the underlying mean. This function describes how the variance of the response varies with the mean. The variance function determines how the variance changes with the mean and ensures that the information in the sample is correctly weighted.

The last component to specify is the deviance–quasi-deviance, $D(y; \mu)$. This function measures the discrepancy between the data and the model and is useful for performing the inferences used in model selection. The definition of variance function and deviance are directly related, as one defines the other.

3. Derivation of link and variance functions

3.1. Binary data

We choose to use logistic regression for the Bernoulli component of the model. In this case the choice of functions is standard (McCullagh and Nelder, 1989). The link function is:

$$g(x) = \log\left(\frac{x}{1-x}\right).$$

In the binary case we have for the variance function:

$$V(\mu) = \mu(1-\mu)$$

When $V(\mu) = \mu(1-\mu)$ we obtain the deviance:

$$D(y;\ \mu) = (1-y)\log(1-\mu) + y\log(\mu).$$

3.2. Truncated Poisson data

In the second stage we consider the subset of data for which $y_i > 0$.

We consider deriving sensible choices for the link, variance and deviance function to be used when we are modelling truncated data. An obvious choice is to consider deriving these quantities from the models used for the data in the GLM case. As our basic model we consider the truncated Poisson distribution.

Before we proceed we must be clear about the motivation of our analysis. The non zero abundance follow some model $u(y|x)$ where y are the observed abundance and x are the associated covariates. The density function $u()$ is defined over the positive integers and in practice its functional form can be difficult to specify a priori. An approach in this case is to construct $u()$ by assuming a standard form, in this case Poisson with the mean modelled via the log link. We then truncate the distribution to arrive at $u()$. The truncated Poisson analysis is derived as follows. First assume that our data y is generated from a Poisson distribution with mean λ. In this case we have:

$$f(y;\ \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}.$$

Now consider data generated from this model where we only observe the results if $y > 0$. This distribution is:

$$u(y|y>0;\ \lambda) = \frac{f(y)}{Pr(y>0)} - \frac{\lambda^y \exp(-\lambda)}{y!\{1-\exp(-\lambda)\}}.$$

This provides us with a likelihood for positive data while giving a clear link between the linear predictor and the mean and variance of the truncated data, which we will now explore.

The truncated Poisson distribution leads to the mean of the observed data, μ being related to λ via:

$$\mu = h(\lambda) = \frac{\lambda}{1-\mathrm{e}^{-\lambda}}$$

so $\lambda = h^{-1}(\mu)$.

In the Poisson model we typically model the mean by:

$$\log(\lambda) = \eta.$$

Combining the previous two equations leads to:

$$\log(h^{-1}(\mu)) = \eta$$

so the link function is $g(x) = \log h^{-1}(x)$.

We also require the variance of Y, which is:

$$\mathrm{Var}(Y) = \mu(1+\lambda-\mu)$$

$$= \mu(1+h^{-1}(\mu)-\mu) = V(\mu).$$

Next, the deviance is:

$$\begin{aligned} D(\mu;\ y) = 2[&y\log h^{-1}(\mu) - h^{-1}(\mu) \\ &- \log[1-\exp\{h^{-1}(\mu)\}] - y\log h^{-1}(y) \\ &+ h^{-1}(y) + \log[1-\exp\{h^{-1}(y)\}]] \end{aligned}$$

Note that we cannot invert h explicitly. This is inconvenient as the fitting algorithm requires multiple evaluations of the link, variance and deviance. For moderate samples ($< 100\,000$) we can calculate the inverse numerically as needed.

3.3. Truncated negative binomial data

The above discussion assumes that the appropriate distribution for our non-zero data is the truncated Poisson. This must always be verified in practice. It is entirely plausible that this approximation will be inadequate with certain data. In particular, we may find that the data is overdispersed, and we must take account of this extra variation in any analysis we perform, otherwise errors can occur. The important point is that the link and variance function must always be subject to empirical validation.

A possible solution for overdispersion is to consider variance functions of the form:

$$V(\mu) = \sigma^2\{\mu(1 + h^{-1}(\mu) - \mu)\}.$$

This allows the estimation of the scale of the variation, while still imposing a particular relationship between the mean and variance of the observations. While this should be useful in many circumstances there will still be occasions where further modifications are required.

Welsh et al. (1996) considered using the truncated negative binomial distribution for the truncated counts. To use this model in the context of GAM modelling we need to derive an appropriate link and variance function. The negative binomial distribution is defined as:

$$f(y;\ \phi,\ \alpha) = \frac{\Gamma(y + \phi\alpha)\phi^{\phi\alpha}}{y!\Gamma(\phi\alpha)(1 + \phi)^{y+\phi\alpha}}.$$

where, $\phi = 1/\delta\alpha$, $\Gamma()$ is the gamma function and α and δ are the modelled parameters. The mean of this distribution is α and the variance is $\alpha + \delta\alpha^2$. Now consider data generated from this model where we only observe the results if $y > 0$. This distribution is:

$$u(y|y > 0;\ \delta,\ \alpha) = \frac{f(y)}{\Pr(y > 0)}$$

$$= \frac{\Gamma(y + \phi\alpha)\phi^{\phi\alpha}}{y!\Gamma(\phi\alpha)(1 + \phi)^{y+\phi\alpha}} \frac{(1 + \phi)^{\phi\alpha}}{(1 + \phi)^{\phi\alpha} - \phi^{\phi\alpha}}.$$

Now for the truncated negative binomial distribution we have the mean:

$$\mu(\delta,\ \alpha) = \frac{(1 + \phi)^{\phi\alpha}}{(1 + \phi)^{\phi\alpha} - \phi^{\phi\alpha}}$$

and variance:

$$\mathrm{Var}(Y;\ \delta,\ \alpha) = \frac{(1 + \phi)}{\phi}\ \mu + \mu(\alpha - \mu).$$

To sensibly discuss the link function in this case we must assume that the dispersion parameter δ is fixed. In fitting the model we will use an iterative technique that alternates fitting the GAM and the dispersion parameter. This is described in Section 4. In this case we model:

$$\log(\alpha) = \eta$$

so:

$$\mu = h(\alpha) = \frac{(1 + \phi)^{\phi\alpha}}{(1 + \phi)^{\phi\alpha} - \phi^{\phi\alpha}}$$

and the link is:

$$g(\mu) = \log[h^{-1}(\mu)].$$

The inverse h cannot be found explicitly, but is well defined and can be approximated numerically. The deviance is:

$$D(\mu;\ y) =$$

$$-2[\log\{u(y|y > 0;\ \delta,\ \alpha)\} - \log\{u(\mu|y > 0;\ \delta,\ \alpha)\}]$$

4. Fitting

Algorithms for fitting GAM models are described in Hastie and Tibshirani (1990), Chambers and Hastie (1992) The models in this paper have been implemented within the S-PLUS (Insightful Corporation) language using the existing gam() technology. Specifically, new family functions have been developed for the truncated Poisson and the truncated negative binomial (conditional on fixed δ). These allow the fitting of GAMs using the standard S-PLUS gam() function. In practice, when fitting negative binomial distributions it is rare to know the value of δ a priori. To estimate the parameters we use the following recursive approach.

1) Choose a starting value for δ.
2) Fit the required model using the truncated negative binomial family function, assuming δ is fixed.
3) Maximize the truncated negative binomial log likelihood over δ, conditional on the fitted values found in 2.
4) Iterate between 2 and 3 until convergence.

Unsupported copies of the family objects and related functions are available from the first author.

5. Results

5.1. Background

We apply the techniques to data from a stratified sample of 261 40×40 m plots. For each plot the stem count of a woodland tree, *Eucalyptus mannifera* was measured. The data are described fully in Austin and Cunningham (1981).

In the original paper the purpose of the analysis was to assess the species response to measured environmental gradients such as temperature and rainfall. The paper recognized the zero inflation and attempted to model it with the available techniques. The major difficulties of the analysis were modelling the zero inflation and specifying an appropriate parametric mean function to model the species response. Both of these features are dealt with cleanly in the present formulation. The available variables for each site were as follows:

- Mean annual rainfall (mm)
- Mean summer rainfall (mm)
- Mean winter rainfall (mm)
- Mean daily temperature (°C)
- July minimum temperature (°C)
- Topography (categorical)
- Nutrient (categorical)
- Stem count for *E. mannifera*

For both the binary component and the count component of the model a full model was first fitted with linear predictor:

$$\begin{aligned}\eta = {} & s(\text{Mean annual rainfall}) \\ & + s(\text{Mean summer rainfall}) \\ & + s(\text{Mean winter rainfall}) \\ & + s(\text{Mean daily temperature}) \\ & + s(\text{July minimum temperature}) \\ & + \text{Topography} + \text{Nutrient}\end{aligned}$$

Models were fitted using the gam routines in S-PLUS 2000 (Insightful Corporation). Variable selection was performed using backwards elimination with the AIC statistic (see Chambers and Hastie, 1992) being used to differentiate between potential models. Continuous predictors were considered for the model in three formulations, either as smooth terms (i.e. using the default degrees of freedom in the fitting package), linear terms or null (i.e. excluded). Where appropriate, scaled deviances where used with the scale parameter estimated from the data.

In the initial stage of modelling we verified that the zeros in the data could not be successfully modelled via a Poisson GAM. To do this we fitted the full model specified in the previous section and calculated the statistic:

$$\sum_i \{I(y_i = 0) - \exp(-\hat{\mu}_i)\}$$

where, y_i is the response for the ith plot and μ is the expected value for the ith plot calculated from the fitted model. This statistic compares the actual number of zero observations with the number predicted by the model conditional on the covariates. To examine the significance of this a parametric bootstrap, based on the fitted model, was used to estimate the sampling distribution of the statistic. This test gave unequivocal evidence of zero inflation (percentile P-value $= 0$).

Given this result we examine the two components of the model separately.

5.2. Presence–abscence component

The model selected for the presence–abscence data was:

$$\begin{aligned}\eta = {} & s(\text{Mean annual rainfall}) \\ & + s(\text{Mean summer rainfall}) \\ & + s(\text{Mean winter rainfall}) \\ & + s(\text{Mean daily temperature}) \\ & + \text{July minimum temperature} + \text{Topography}\end{aligned}$$

The standard logistic link and Bernoulli variance function was used. This is only a marginal simplification from the full specification. The fitted curves are shown in Fig. 1. Note that nonlinear terms are maintained in the final model.

5.3. Count data

When the data was restricted to exclude all absence records a sample size of 79 plots was obtained. For the count data we have two

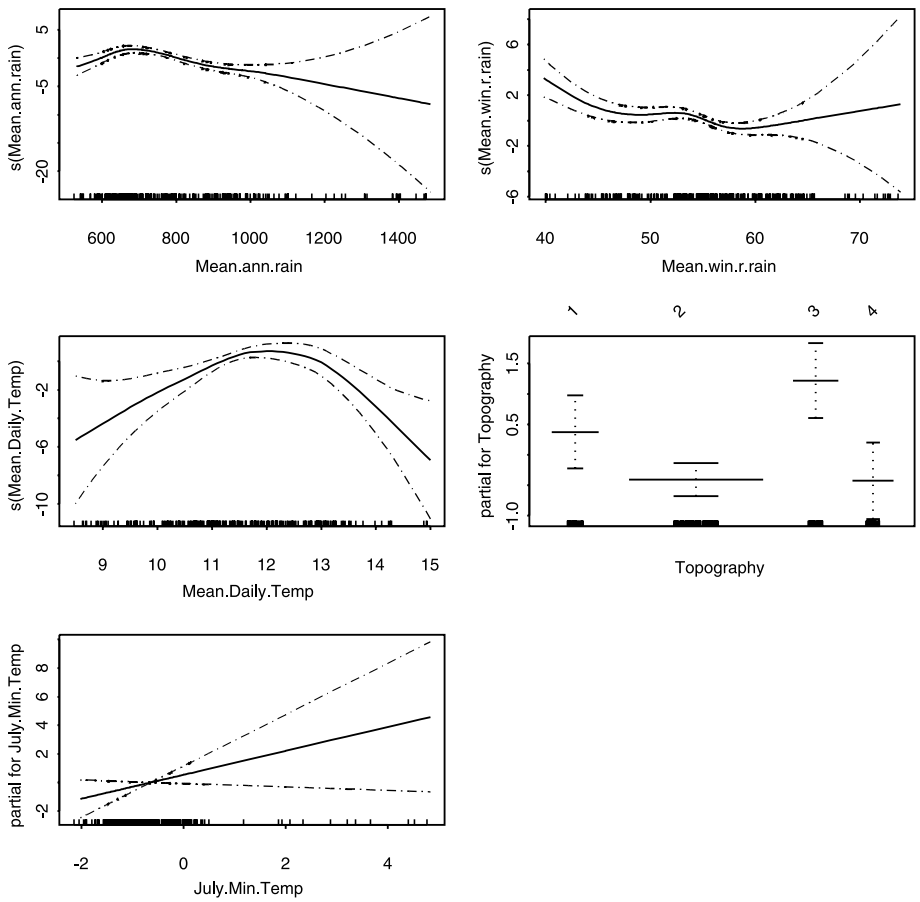

Fig. 1. Additive terms of final model for the presence–absence component. Dashed lines give approximate, pointwise, 95% confidence intervals. Refer to text for details of terms.

competing models, the truncated Poisson and the truncated negative binomial. In addition we can consider an overdispersed version of the truncated Poisson, leading to three candidate specifications.

Overdispersion was clearly present so only the overdispersed truncated Poisson model was considered during variable selection. Both candidate models produced identical specifications for the mean, which was

$$\eta = s(\text{Mean winter rainfall}) + s(\text{July minimum temperature})$$

This is both simpler than the 0/1 component and functionally different, but any comparison must be made recognizing the different sample sizes involved and the difficulty in comparing the effect of terms in two regression models when different terms are present in each model. Note that there is no theoretical reason to expect the two models to have identical covariates as the responses are fundamentally different.

Choosing between the models on the basis of fit is complicated as they are non-nested. Fig. 2 displays the absolute value of the raw residuals of the fit against the fitted values and compares this with the variance function estimated by the two models. Examining the fitted values it is clear that the simple Poisson model is a gross underestimate of the variation of the data. The choice between the overdispersed Poisson and the negative binomial formulation is less clear. For small values of the mean the Poisson model appears to

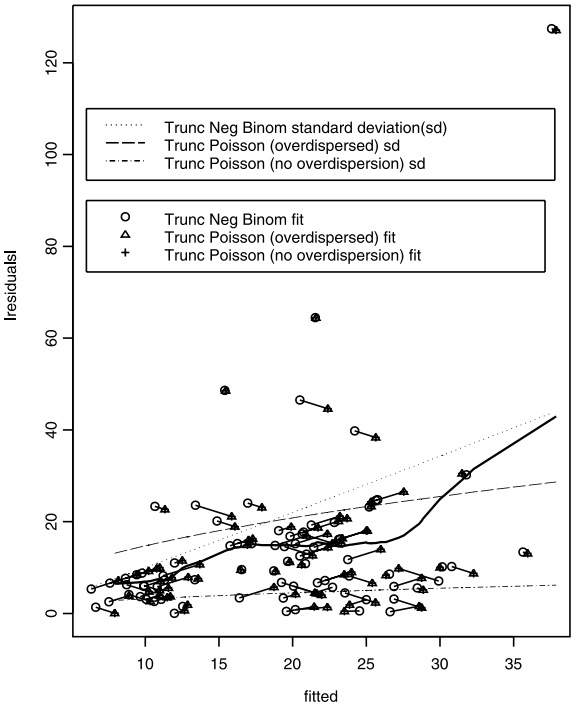

Fig. 2. Plot of fitted values against residuals for the three reduced candidate models. Points that relate to the same observation are joined by a solid line, and allow comparison of the fits. The standard deviation (S.D.) estimated by the fitted models are overlaid, as well as a nonparametric smooth of the residuals (solid line).

overestimate the variance for this case. The negative binomial model more accurately reflects the variance in this region of the fitted values. For larger fitted values there is little information to contrast the two models.

We choose the negative binomial based on its superior reflection of the variance structure. The fitted curves are shown in Fig. 3. Note that a model with only linear terms would be clearly inadequate.

The chosen models would represent a basis for predicting stem counts in the environments that the data represent. The ecological interpretation of the abundance model is more problematic. While theory might suggest the plausibility of bimodal responses to particular gradients in the presence of competition, it is more likely that the data are exhibiting the effect of spatial disequilibrium, such as a missing spatially varying covariate. In the original analysis Austin and Cunningham (1981) broke the data into four regions and fitted a separate analysis at each location. Space restrictions do not allow us to repeat this analysis here.

6. Discussion

The results of the analysis demonstrate an example where the pattern of presence–abscence of a species, conditional on the covariates, is markedly different from the pattern of abundance. Incorporating this structure into the analysis allows us to improve our predictions of the mean. If we attempt, following a classical analysis, to model this structure as extra-Poisson variation, with variance of $Y \propto \sigma^2\mu$, we will face two consequences. First, the model for the variance is not correctly specified for almost all conceivable zero contamination patterns. Second, this analysis addresses the novel feature of the data by trying to adjust the uncertainty of the predictions (by inflation) instead of improving their precision. It is thus an inefficient approach to the analysis.

In advocating the two step approach to modelling zero inflated count data it is reasonable to consider the circumstances where the method may fail or be inefficient. Some aspects of this are discussed in Welsh et al. (1996). In principle we argue that though some small biases may potentially be introduced by the two step approach if the model does not hold exactly, the potential advantages dominate the negative aspects. As a purely illustrative example, consider data that does follow a Poisson model with log link and variance proportional to the mean. If we unwisely tried to model this as ZIP data, simple calculations show that the logistic model is mis-specified, because the effects of the covariates on the probability of presence or absence are not linear on the logistic scale. In a practical sense the logistic model may well approximate the actual probabilities of $y = 0$, with the GAM model providing even greater flexibility in modelling these non linear patterns.

When modelling species abundance the aim of the analysis is sometimes to make predictions for a range of unsurveyed sites to produce a map of distribution. Standard GAM theory can be used to

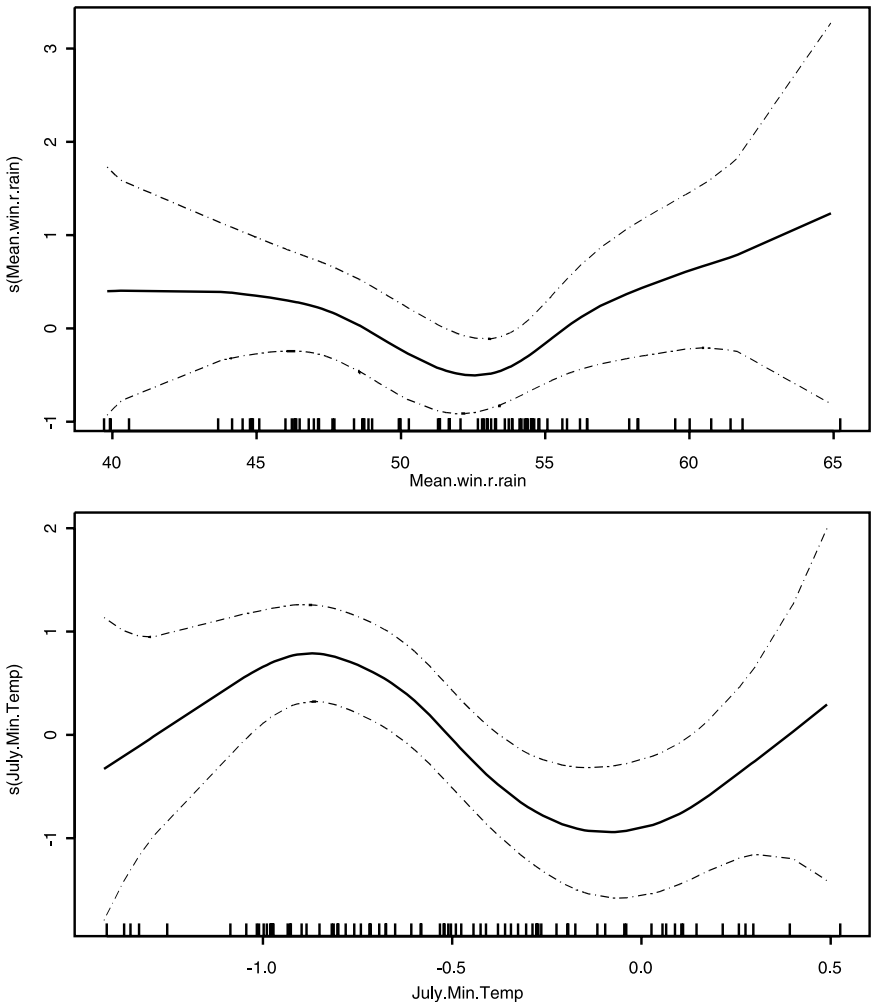

Fig. 3. Final model for the count data based on the truncated negative binomial. Dashed lines give approximate, pointwise, 95% confidence intervals. Note the model is different from the presence–absence component.

produce these predictions. While it may philosophically be preferable to produce separate maps for each component of the model, in practice, it may not be feasible because of end user demands. In this case the models can be combined to produce a single prediction for each cell. This technique is sketched in Appendix A.

It is useful to note that the family functions here will provide maximum likelihood estimates when used with fully parametric linear predictors and the S-PLUS glm() function.

Further research is needed to explore models for integer valued data. A variety of approaches might be considered such as quasi-likelihood models or translated distributions. As an example, to simplify computation we could consider the translated Poisson distribution, $f(y;\ \lambda) = \text{Poisson}\ (y-1;\ \lambda)$, $y = 1, 2, \ldots$, with $\log(\mu - 1) = x^{\mathrm{T}}\beta$. The strengths and weaknesses of the different approaches will need to be considered.

Acknowledgements

Part of this research was supported by Australian Research Council Large Grant A00000506.

We thank the associate editor and two anonymous referees for their helpful comments, which improved the paper. Mike Austin provided the data and useful discussion.

Appendix A: Combining the models

The models we have derived provide two pieces of information. For a given cell, the presence–absence model predicts the probability that the species is present in that cell. The abundance model predicts how abundant a species should be if it is present. It is common in applications that we wish to provide a single map of our predictions at new sites. In this case we wish to predict the expected abundance of the species averaged over all sites, both where it is present and absent. In other applications we might be interested in examining each model separately.

To combine the two models we use the standard results that:

$$E[Y] = E[E[Y|I]]$$

$$\mathrm{Var}(Y) = \mathrm{Var}(E[Y|I]) + E(\mathrm{Var}[Y|I])$$

where, E and Var are the expectation and variance operators and we let Y be the observed abundance and I be an indicator of presence.

We take the following steps to predict the abundance at a site with covariates x. From the presence–absence model, estimate the probability that $y > 0$, namely $\hat{p}$. From the truncated model estimate the expected count $\hat{\mu}$ and dispersion $\hat{\sigma}^2$. We then estimate the mean abundance to be:

$$\mathrm{Predicted}(y) = \hat{p}\hat{\mu}$$

and the prediction variance:

$$\begin{aligned}\mathrm{Prediction\ variance}(y) &= \mathrm{Var}(y - \hat{p}\hat{\mu}) \\ &= \mathrm{Var}(y) + \mathrm{Var}(\hat{p}\hat{\mu})\end{aligned}$$

The variance of y, conditional on it being observed, is found from the variance function $V(\mu)$. The variance of y is thus:

$$\mathrm{Var}(y) = pV(\mu) + \mu^2 p(1-p).$$

The exact variance of $\hat{p}\hat{\mu}$ is difficult to calculate so we use a delta method approximation:

$$\mathrm{Var}(\hat{p}\hat{\mu}) = DVD^T$$

where

$$V - \begin{bmatrix} \mathrm{Var}(\hat{p}) & 0 \\ 0 & \mathrm{Var}(\hat{\mu}) \end{bmatrix}$$

$$D - [\mu \quad p]$$

and T is the transpose operator.

In practice μ and p are unknown and are replaced by their estimates.

References

Austin, M., Cunningham, R., 1981. Observational analysis of environmental gradients. Proceedings of the Ecological Society of Australia 11, 109–119.

Chambers, J., Hastie, T., 1992. Statistical Models in S. Wadsworth and Brooks, California.

Guisan, A., Zimmermann, N., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147–186.

Hastie, T., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London.

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

Leathwick, J., Austin, M., 2001. Competitive interactions between tree species in New Zealand's old growth indigenous forests. Ecology, in press.

McCullagh, P., Nelder, J., 1989. Generalized Linear Models. Chapman & Hall, New York.

Mullahy, J., 1986. Specification and testing of some modified count data models. Journal of Econometrics 33, 341–365.

Ridout, M., Demetrio, C., Hinde, J., 1998. Models for count data with many zeros. In: Proceedings of the 19th International Biometrics Conference, Cape Town, pp. 179–190.

Welsh, A.H., Cunningham, R., Donnelly, C., Lindenmeyer, D., 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecological Modelling 88, 297–308.