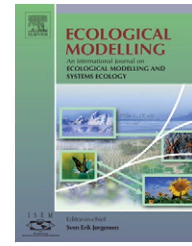
ECOLOGICAL MODELLING 199 (2006) 153–163

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Comparing species abundance models

Joanne M. Potts[a,*], *Jane Elith*[b]

[a] *Arthur Rylah Institute for Environmental Research, Department of Sustainability and Environment, 123 Brown St, Heidelberg, Victoria 3084, Australia*

[b] *School of Botany, The University of Melbourne, Parkville, Victoria 3010, Australia*

ARTICLE INFO

Article history:
Published on line 13 July 2006

Keywords:
Zero-inflation
Over-dispersion
Poisson regression
Abundance data
Model evaluation

ABSTRACT

Five regression models (Poisson, negative binomial, quasi-Poisson, the hurdle model and the zero-inflated Poisson) were used to assess the relationship between the abundance of a vulnerable plant species, *Leionema ralstonii*, and the environment. The methods differed in their capacity to deal with common properties of ecological data. They were assessed theoretically, and their predictive performance was evaluated with correlation, calibration and error statistics calculated within a bootstrap evaluation procedure that simulated performance for independent data.

The hurdle model performed best, with the highest correlations between the observed and predicted abundances. This model was also well calibrated, giving the closest agreement between observed and predicted abundances. The negative binomial was the worst performing model. It had weaker correlations than the other models and resulted in a strong, inconsistent bias in predictions. The standard Poisson model which accommodates neither zero-inflation nor over-dispersion gave accurate estimates of regional population abundance, but at the individual population level they were inconsistent and biased.

The strong performance of the hurdle model, coupled with theoretical properties that suit it for these data and for the ecology of this species, suggest that it is a useful alternative to other modelling methods. The gains in performance have practical advantages where predictions are used by conservation planners to understand population dynamics or to assess the relative risks of alternative management scenarios.

1. Introduction

Environmental managers require estimates of species abundance in a broad range of situations: classifying species according to the IUCN Red List for threatened species (IUCN, 2001); conducting population viability analyses (Possingham et al., 2001); managing fire regimes (e.g. the endangered shrub *Grevillea caleyi*, Regan et al., 2003); monitoring (e.g. population changes of pest species over time, Hone, 1999); and reintroducing or translocating animals (Lubow, 1996). Obtaining such estimates can be resource demanding because surveys are expensive and time-consuming, especially if the species is rare or occurs in remote locations.

Mathematical models that quantify the relationship between a species' abundance and environmental characteristics may be used to complement survey work. Predictions of abundance can then be made at unsurveyed locations and used to guide management decisions. The choice of a mathematical model should be governed by knowledge of the species and characteristics of the available data. Accommodating characteristics of the data in a model can increase its complexity and thus decrease the ease with which it is

* *Corresponding author*. Tel.: +61 3 9450 8701; fax: +61 3 9450 8799.
E-mail address: Joanne.Potts@dse.vic.gov.au (J.M. Potts).

0304-3800/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2006.05.025

developed, interpreted and understood. There is a trade-off between complicated models that account appropriately for characteristics of the data and simpler models that are easier to develop but may be sub-optimal.

This study explores these trade-offs using data collected on an Australian threatened plant species, *Leionema ralstonii* (F. Muell.) Paul G. Wilson (Rutaceae). The species is protected under both state and federal legislation due to its small geographical range (~40,000 ha in the southeast corner of New South Wales) and population size (~18,100 individual plants distributed across 71 discrete populations on rocky outcrops, NSW National Parks and Wildlife Service, 2003). Predictions of species abundance were required to guide management of the species and as inputs for a population viability analysis (Potts et al., submitted for publication).

2. Technical review

This paper focuses on regression methods within a generalised linear model framework (McCullagh and Nelder, 1989). These types of models are used frequently to quantify the relationship between species abundances and environmental characteristics (e.g. Wintle et al., 2005; Leathwick et al., this issue). Regression models are typically described in terms of their systematic component in which the response is linked to the environmental data, and their stochastic structure that describes the error distribution (Venables and Ripley, 2002). In order to focus on the link function separately to the rest of the systematic component, in this paper we use the terms model structure and model specification. Model structure includes both the choice of environmental characteristics (the explanatory variables) assumed to affect species abundance (the response variables) and the shape of the modelled responses (linear, quadratic and so on). Model specification defines how these variables are related using a 'link' function. The choice of 'link' function allows the response variable to be non-linearly related to the explanatory variables.

When the response variable is count data (as is the case when working with abundance observations), the response variable can be linked to the explanatory variables using a log transformation (McCullagh and Nelder, 1989):

$$\log(p) = \ln(p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_N X_N \quad (1)$$

where p is the probability of an event occurring, X_N the Nth independent variable and β_N is the regression coefficient. In our example, an event is the mean rate at which individuals occur on each outcrop (termed μ). This model is referred to as a standard Poisson regression and is the simplest and most commonly specified model for count data.

This model specification assumes equi-dispersion, meaning if Y is Poisson distributed, then the expectation of Y is equal to the variance of Y. Since the variance is not constant, the regression is intrinsically heteroskedastic (i.e. the variance increases with increasing mean). Violating the assumption of equi-dispersion has similar consequences to violating the assumption of homoskedasticity in linear regression (Cameron and Trivedi, 1998). The standard errors of the predictions are biased because the different populations have different variances.

If the variance exceeds (or is less than) the mean, then the data are said to be over- (or under-) dispersed (Cox, 1983). An indication of the magnitude of over- or under-dispersion can be obtained by comparing the sample mean and variance of the dependent count variable. Over-dispersion can be reduced using explanatory variables. When working with ecological data the equi-dispersion assumption is commonly violated, especially if the data are zero-inflated (Cameron and Trivedi, 1998).

Zero-inflated data contain substantially more zeros that the specified distribution suggests (Tu, 2002). They occur because the data generating process adds an additional mass at zero, inflating the probability of observing a zero above that which is consistent with the specified distribution. It may therefore be a mis-specification to assume that the zero and non-zero observations come from the same source. Visual inspection of a histogram of the observed data might suggest a spike of zero observations if zero-inflation is present.

Count data for rare species commonly are zero-inflated. The species may be observed absent at many sites because of true negative or false negative observations (Martin et al., 2005b). We can think about these in terms of the source of the error (i.e. the uncertainty). True negative observations are attributable to structural zeros (i.e. unsuitable habitat) or environmental process (i.e. suitable but unoccupied habitat because the species does not saturate its environment). The latter are also known as stochastic zeros. False negatives are attributable to experimental design (i.e. survey site is utilised by the species, but not during the survey period) or observer error (i.e. species is present but not detected). If not modelled properly, the presence of excess zeros can violate the distributional assumptions of the analysis, lead to invalid scientific inferences and create computational difficulties (Tu, 2002). Zero-inflation may cause over-dispersion, but it is possible for either of these two features to occur independently in any data set. Formal statistical tests are available for both equi-dispersion and zero-inflation (see Cox, 1983; Böhning, 1994; van den Broek, 1995; Ridout et al., 2001; Hall and Berenhaut, 2002).

If a data set is zero-inflated and/or violates the equi-dispersion assumption, the standard Poisson regression is still commonly used (Cameron and Trivedi, 1998). We believe this is because the Poisson model is easy to implement and available in a number of statistical packages. Incorrectly specifying a Poisson distribution in the presence of zero-inflation and/or over-dispersion has two important consequences. Firstly, it will result in incorrect predictions at each site, although the average prediction across all sites will be consistent with that observed (Cameron and Trivedi, 1998; Barry and Welsh, 2002). Secondly, it will cause overly optimistic conclusions about the statistical significance of the explanatory variables (i.e. reduced standard errors of the coefficients). This means that under common model-building procedures such as stepwise selection, incorrect variables are more likely to be retained (Fitzmaurice, 1997). Both of these consequences are important for environmental managers, as either the predictions and/or the model structure may influence decision making.