Acta Oecologica 37 (2011) 170–178

Contents lists available at ScienceDirect

Acta Oecologica

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

Original article

Comparing distribution models for small samples of overdispersed counts of freshwater fish

Lise Vaudor [a,*], Nicolas Lamouroux [a], Jean-Michel Olivier [b]

[a] CEMAGREF MALY, 3 bis quai Chauveau, CP 220, 69336 Lyon Cedex 09, France
[b] UMR 5023, CNRS, Ecologie des Hydrosystèmes Fluviaux, Université Lyon 1, 69622 Villeurbanne, France

ARTICLE INFO

Article history:
Received 24 June 2010
Accepted 12 January 2011
Available online 23 February 2011

Keywords:
Abundance
Negative binomial distribution
Poisson distribution
Zero-inflation
Profile likelihood
Confidence intervals

ABSTRACT

The study of species abundance often relies on repeated abundance counts whose number is limited by logistic or financial constraints. The distribution of abundance counts is generally right-skewed (i.e. with many zeros and few high values) and needs to be modelled for statistical inference. We used an extensive dataset involving about 100,000 fish individuals of 12 freshwater fish species collected in electrofishing points (7 m^2) during 350 field surveys made in 25 stream sites, in order to compare the performance and the generality of four distribution models of counts (Poisson, negative binomial and their zero-inflated counterparts). The negative binomial distribution was the best model (Bayesian Information Criterion) for 58% of the samples (species–survey combinations) and was suitable for a variety of life histories, habitat, and sample characteristics. The performance of the models was closely related to samples' statistics such as total abundance and variance. Finally, we illustrated the consequences of a distribution assumption by calculating confidence intervals around the mean abundance, either based on the most suitable distribution assumption or on an asymptotical, distribution-free (Student's) method. Student's method generally corresponded to narrower confidence intervals, especially when there were few (≤ 3) non-null counts in the samples.

1. Introduction

A common issue in ecology consists of modelling the abundance of organisms and the environmental determinants of this abundance, using repeated counts. In many cases, authors pay minor attention to the general shape of the distribution of counts and focus on summary properties such as the variance-to-mean relationship. Nevertheless, the need to characterise this distribution is increasingly acknowledged in the literature for ecological and statistical reasons.

Ecological interest lies in the determination of the biotic and abiotic factors that can influence the distribution of abundance counts. For example, abundance heterogeneity might be linked to the heterogeneity of the environment, and the high frequency of zero counts can be due to unsuitable habitats (Martin et al., 2005). More generally, variations in the distribution shape might reflect differences in species behaviour, due to e.g. population density, season, presence of predators or competitors (Taylor, 1984). Abundance samples repeated in time and space are particularly important for identifying the generality of a given distribution model and inferring the limits of its relevance.

Concerning statistical reasons, inference and estimation of abundance often require a distributional assumption for abundance counts. An inappropriate assumption has many implications, including misleading confidence intervals around the estimate of mean abundance (Rosenblum and Van der Laan, 2008), or unreliable error rates with traditional statistical analyses such as linear models (Power and Moser, 1999; McArdle and Anderson, 2004).

Poisson-mixture distributions are widely recognised as potentially appropriate distribution models for count data. In particular, the negative binomial has a long history as a suitable model for clumped count data (Anscombe, 1949; Bliss and Fisher, 1953; Evans, 1953) and is increasingly used in abundance count studies. Recent developments have discussed extending the Poisson or negative binomial distributions into models that account for extra zeros (Martin et al., 2005; Wenger and Freeman, 2008). Some studies have used extensive ecological datasets to compare the performance of these distributions under a variety of environmental and biological conditions (Welsh et al., 1996; Ridout et al., 1998; Gray, 2005; Martin et al., 2005; Warton, 2005; Potts and Elith, 2006;

---

[*] Corresponding author. Tel./fax: +33 472208914.
E-mail addresses: lise.vaudor@cemagref.fr (L. Vaudor), nicolas.lamouroux@cemagref.fr (N. Lamouroux), jean-michel.olivier@univ-lyon1.fr (J.-M. Olivier).

1146-609X/$ – see front matter © 2011 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.actao.2011.01.010

Wenger and Freeman, 2008). Nevertheless, few studies of this kind concern freshwater fish (Wenger and Freeman, 2008; Lewin et al., 2009).

Freshwater systems are often heavily impacted by human activities, and the study of fish abundance and of its environmental determinants is of particular importance to develop objective diagnostic and management tools for these systems (Lamouroux et al., 1999; Lewin et al., 2009). Estimating fish abundance is particularly challenging in medium and large rivers due to the overdispersion of counts, that can result from schooling behaviour and/or strong habitat preferences (Lamouroux et al., 1999). The rarity of many species and the limitations in sampling effort due to budget constraints adds further difficulties (Lewin et al., 2009).

Studies comparing distribution models for count data generally focus on modelling the distribution of one or a few large size samples (i.e. set of counts) rather than the distribution of many small size samples, repeated in space and time (Warton, 2005). Still, repeated small size samples are extremely common in ecological studies (for instance for monitoring species). When sample size is limited, due to logistic or financial constraints, it is of particular importance that inference is carried out based on a sound, parsimonious, and general distribution model. It is therefore important to identify which sample characteristics affect the performance and generality of candidate distribution models.

In this paper, we compared the performance of four distributions (Poisson, negative binomial and their zero-inflated counterparts) for modelling observed abundance counts of freshwater fish. Models were tested on a large existing dataset involving about 100,000 fish individuals of 12 species collected during 350 field surveys made in 25 stream sites of the Rhône basin (France) between 1985 and 2007. On each of these repeated surveys, a limited number (20–180) of independent abundance counts (electrofishing of a fixed area) was made. After selecting the best model for each survey–species combination according to the Bayesian Information Criterion (BIC), we used multinomial logistic regressions to identify which factors (e.g. total abundance, belonging to a species or a site) influenced this selection. Finally, we illustrated the consequences of a distributional assumption for estimating confidence intervals around the mean abundance estimate.

2. Methods

2.1. Sites and fish sampling

The initial fish dataset considered included 104,446 individuals belonging to 31 species collected during 350 field surveys made in 25 stream sites of the Rhône basin, between 1985 and 2007 (see Persat, 1988; Lamouroux et al., 1999, for previous use of part of the data). The sampling sites were 18 stream reaches of the Rhône river (the largest river in France) and seven reaches of its tributary the Ain river (Olivier et al., 2009). All sites were long enough to include several "pool-riffle" sequences, i.e. they reflected the habitat heterogeneity in which the fish communities mostly live. Fifteen of the Rhône sites were situated in a section of the river by-passed for hydropower generation, downstream from a dam (dams of Chautagne, Belley, Brègnier-Cordon, Pèage-de-Roussillon, Montèlimar). The three other Rhône sites and the seven Ain sites were not in by-passed sections but were also subject to artificial flow regimes. All sites were situated on large regulated rivers (several dozen metres wide, around 1 m deep at low flow) and sites had varied habitat characteristics (Lamouroux et al., 1999).

Each of the 350 field surveys was an electrofishing survey made during 1–4 consecutive days, generally at low to medium flow rates. During each field survey, 20–180 points were sampled using a "Point Abundance Sampling" strategy (Persat and Copp, 1990; Copp, 1992; Lamouroux et al., 1999). Points were electrofished by dropping or putting the electrode in the river and net-catching all fishes attracted. The attraction area of the electrode was approximately 7 m^2 (Regis et al., 1981). Electrofishing was made from a boat or wading in the stream where possible, and points were randomly or regularly spaced within the reach in areas where electrofishing was efficient (water depth < 2 m, water velocity $<1 \text{ m s}^{-1}$). The distance between points was >30 m, and we considered that the abundance counts were independent.

2.2. Statistical modelling

We modelled the abundance data separately for the different species–survey combinations (called samples hereafter). In other words, a sample x_i with i in 1, …, n, corresponded to n abundance counts (n is between 20 and 180, according to the survey) of a single species in a single stream site. Having few samples of abundance for the rarest species could make the analyses of the effect of species on model selection particularly unbalanced. To limit this problem we excluded from our study the species that were present in less than 100 surveys. Having really few individuals in a sample could make the use of a distribution model irrelevant. To limit this problem we excluded from our study the samples that had less than three individuals. This limit of three individuals is quite low, and the relevance of modelling a distribution where the total number of individuals is this low will be discussed.

We considered several candidate Poisson-mixture statistical distributions of abundance counts, that are routinely used to model count data. The Poisson (P) distribution accounts for discrete and positive counts but cannot account for overdispersion (when variance is higher than mean), which is a frequent feature of abundance counts. The negative binomial (NB) distribution is a generalisation of the P model, that allows for overdispersion (Johnson et al., 1992). This overdispersion may for instance be interpreted as resulting from the heterogeneity of habitat, or from a gregarious behaviour of the species under study. The P and NB distribution can also be extended to account for an additional proportion of extra zeros: they are then called zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB), respectively (Lambert, 1992). This kind of model has lately been much studied in the ecological literature (Welsh et al., 1996; Ridout et al., 1998; Gray, 2005; Martin et al., 2005; Potts and Elith, 2006; Wenger and Freeman, 2008). Zero-inflation is mainly interpreted as reflecting the existence of two kinds of zeros: structural zeros, which occur because the sampling is carried out in conditions such that the presence of an individual is impossible, and stochastic zeros, which occur simply by chance.

Many other candidate distribution models, and in particular hurdle models, were not included in our modelling exercise. Hurdle models (also called two-stage models) are another type of zero-modified counts models (Mullahy, 1986; Ridout et al., 1998; O'Neill and Faddy, 2003; Cunningham and Lindenmayer, 2005; Gray, 2005). Contrary to zero-inflated models that reflect a latent zero process increasing the proportion of zeros (Gray, 2005), they treat positive counts separately from zeros. With hurdle models, the occurrence of a positive count is treated as a binomial process, and the positive counts are generally modelled by a truncated-at-zero count data model. The underlying idea is that distinct ecological processes govern the presence or absence of individuals and the number of individuals given presence (Mullahy, 1986). We did not include these models in the present study because we considered that zero-inflated models were more realistic from an ecological point of view. Indeed, it has been demonstrated that freshwater fish species often have null counts in their suitable habitat: some of the zeros we observe are probably stochastic zeros (Lamouroux et al., 1999; Martin et al., 2005).