Available online at www.sciencedirect.com

SCIENCE DIRECT®

Ecological Modelling 185 (2005) 1–12

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Selecting a distributional assumption for modelling relative densities of benthic macroinvertebrates

Brian R. Gray [*]

Upper Midwest Environmental Sciences Center, U.S. Geological Survey, 2630 Fanta Reed Road, La Crosse, WI 54603, USA

Received 5 May 2004; received in revised form 3 November 2004; accepted 7 November 2004

Abstract

The selection of a distributional assumption suitable for modelling macroinvertebrate density data is typically challenging. Macroinvertebrate data often exhibit substantially larger variances than expected under a standard count assumption, that of the Poisson distribution. Such overdispersion may derive from multiple sources, including heterogeneity of habitat (historically and spatially), differing life histories for organisms collected within a single collection in space and time, and autocorrelation. Taken to extreme, heterogeneity of habitat may be argued to explain the frequent large proportions of zero observations in macroinvertebrate data. Sampling locations may consist of habitats defined qualitatively as either suitable or unsuitable. The former category may yield random or stochastic zeroes and the latter structural zeroes. Heterogeneity among counts may be accommodated by treating the count mean itself as a random variable, while extra zeroes may be accommodated using zero-modified count assumptions, including zero-inflated and two-stage (or hurdle) approaches. These and linear assumptions (following log- and square root-transformations) were evaluated using 9 years of mayfly density data from a 52 km, ninth-order reach of the Upper Mississippi River ($n = 959$). The data exhibited substantial overdispersion relative to that expected under a Poisson assumption (i.e. variance:mean ratio $= 23 \gg 1$), and 43% of the sampling locations yielded zero mayflies. Based on the Akaike Information Criterion (AIC), count models were improved most by treating the count mean as a random variable (via a Poisson-gamma distributional assumption) and secondarily by zero modification (i.e. improvements in AIC values = 9184 units and 47–48 units, respectively). Zeroes were underestimated by the Poisson, log-transform and square root-transform models, slightly by the standard negative binomial model but not by the zero-modified models (61%, 24%, 32%, 7%, and 0%, respectively). However, the zero-modified Poisson models underestimated small counts ($1 \leq y \leq 4$) and overestimated intermediate counts ($7 \leq y \leq 23$). Counts greater than zero were estimated well by zero-modified negative binomial models, while counts greater than one were also estimated well by the standard negative binomial model. Based on AIC and percent zero estimation criteria, the two-stage and zero-inflated models performed similarly. The above inferences were largely confirmed when the models were used to predict values from a separate, evaluation data set ($n = 110$). An exception was that, using the evaluation data set, the standard negative binomial model appeared superior to its zero-modified counterparts using the AIC (but not percent zero criteria). This and other evidence suggest that a negative binomial distributional assumption should be routinely considered when modelling benthic macroinvertebrate data from low flow environments. Whether negative binomial models

---

[*] Tel.: +1 608 781 6234; fax: +1 608 783 6066.
E-mail address: brgray@usgs.gov.

doi:10.1016/j.ecolmodel.2004.11.006

should themselves be routinely examined for extra zeroes requires, from a statistical perspective, more investigation. However, this question may best be answered by ecological arguments that may be specific to the sampled species and locations.

Keywords: *Hexagenia*; Hurdle models; LTRMP; Mayflies; Negative binomial distribution; Two-stage models; Zero-inflated count models

1. Introduction

The selection of a distributional assumption is a frequent challenge when modelling macroinvertebrate density data. Such data typically exhibit overdispersion with respect to that expected under a standard count distribution, that of the Poisson. Further, overdispersion may derive from multiple sources, including unobserved heterogeneity, zeroes in excess of those expected under a Poisson or generalized Poisson distributional assumption, and contagion (Cameron and Trivedi, 1998).

Unobserved heterogeneity arises when observations derive from processes that are incompletely observed or that are incompletely represented by covariates. Such variation may be modelled by assuming the count mean varies randomly. If the mean is assumed to vary as a gamma-distributed random variable, for example, then outcomes may be modelled using negative binomial regression (Johnson et al., 1993).

Zeroes in excess of those expected under a Poisson assumption will arise naturally as a result of heterogeneity of the mean (Cameron and Trivedi, 1998). However, in the absence of heterogeneity or when accommodating heterogeneity fails to account for observed zeroes, then zeroes may be modelled using zero-modified count models.

Common zero-modified count models include the two-stage (or hurdle) and zero-inflated count models (Cameron and Trivedi, 1998). The former treats zeroes and positive counts separately, but assumes that positive counts arise as a result of conditions that result in passing a threshold or hurdle. Zero-inflated models assume an unobserved or latent “zero process” that augments the zeroes arising from the count distribution. Such zeroes may arise from undercounting (i.e. “false zeroes”) and from sampling habitat that is unsuitable for the organism in question. Regression may be performed under both approaches.

The value of accommodating heterogeneity using both negative binomial distributional assumptions and zero-modified approaches was addressed using 9 years of mayfly relative density data from Navigation Pool 13, a ninth-order reach of the Upper Mississippi River. Negative binomial and zero-modified negative binomial models were compared with their Poisson counterparts, and all were compared with models based on log- and square root-transformations. Relative density values were coarsely adjusted for undercounting in standard and two-stage negative binomial models.

2. Methods

2.1. Mayfly sampling protocol

Mayflies were sampled annually from 1993 to 2001 in Navigation Pool 13 (hereafter, Pool 13) of the Upper Mississippi River. Pool 13 drains approximately 222,000 km^2 of the Upper Mississippi River basin, and attains mean annual discharges that have varied from 736 to 2600 m^3/s since impoundment in 1940 (U.S. Geological Survey, 2001).

Sampling was conducted using a stratified randomized sampling design. Strata, defined primarily by geomorphic features (Wilcox, 1993), included contiguous backwaters (backwaters connected by surface flow with the river’s main channel during normal discharge; 2779 ha), main channel borders (areas in the main channel located between the navigation channel and the main channel bank; 1145 ha), impounded area (large, mostly open-water areas located in the downstream portion of navigation pools; 3672 ha) and side channels (channels other than the main channel; 985 ha). Backwater components of the reach that were isolated from the river were not sampled because such areas were not accessible by boat. Also, the main navigation channel was not sampled because of safety concerns and because the channel is composed primarily of large diameter substrates that are not suitable for benthic taxa groups of primary interest