On Predicting Abundance from Occupancy

Andrew R. Solow[1,*] **and Woollcott K. Smith**[2]

1. Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543; 2. Statistics Department, Temple University, Philadelphia, Pennsylvania 19122

Submitted November 25, 2009; Accepted March 12, 2010; Electronically published May 13, 2010

abstract: There is growing interest in predicting the abundance of a species in a region from the occupancy of cells in a uniform grid overlaid on the region. When the number of individuals in each cell follows a negative binomial distribution, prediction is in general not possible from the number of unoccupied cells alone. A prediction method based on the number of unoccupied cells and the number containing a single individual is described and shown to work well on simulated and real data.

Keywords: negative binomial distribution, optimal prediction, species abundance, species occupancy.

Introduction

A problem of interest in statistical ecology is predicting the number of a sessile organism in a region from the number of uniform cells overlaid on the region containing no individuals—or, equivalently, at least one individual (Hui et al. 2009). Here, we apply the term "prediction" to random variables and the term "estimation" to fixed but unknown parameters. The practical issue here is that it is often much less costly to determine that a cell is empty than to count the number of individuals in it. In a widely cited contribution, He and Gaston (2000) proposed a predictor under the assumption that the cell counts are independent, identically distributed negative binomial random variables. The negative binomial distribution is commonly used as a model for clustered counts. Conlisk et al. (2007) pointed out that the method of He and Gaston (2000) is based on an incorrect assumption about the aggregation of negative binomial counts and that prediction is not strictly possible without additional information. In response, He and Gaston (2007) argued that, for a number of data sets, this assumption is nearly correct and the method works well. The purpose of this note is to show that the problem can be avoided altogether if prediction is based on the number of cells containing no individuals and the number containing exactly one individual.

* Corresponding author; e-mail: asolow@whoi.edu.

Am. Nat. 2010. Vol. 176, pp. 96–98. 0003-0147/2010/17601-51759$15.00.
DOI: 10.1086/653077

Model and Method

Let the random variable N_j be the number of individuals in cell j ($j = 1, 2, \ldots, J$). Following He and Gaston (2000), we assume that the counts in different cells are independent and that N_j has a negative binomial distribution with probability mass function

$$\Pr(N_j = n_j) = \frac{\Gamma\{[\mu/(\gamma - 1)] + n_j\}}{n!\Gamma[\mu/(\gamma - 1)]}\left(\frac{1}{\gamma}\right)^{[\mu/(\gamma-1)]}\left(1 - \frac{1}{\gamma}\right)^{n_j} = p(n_j), \quad (1)$$

where $\mu > 0$ is the mean of N_j and $\gamma > 1$ is the ratio of the variance of N_j to μ. He and Gaston (2000) used a different parameterization, but we believe this one is more useful.

Let the random variable $N = \sum_{j=1}^{J} N_j$ be the total number of individuals in the region. Interest centers on predicting N from the observed number m_0 of unoccupied cells. The optimal predictor is the conditional expected value of N, given m_0:

$$E(N|m_0) = \frac{(J - m_0)\mu}{1 - p(0)}, \quad (2)$$

where from equation (1) $p(0) = \gamma^{[\mu/(1-\gamma)]}$. The expression in equation (2) is the product of the number of occupied cells and the expected number of individuals in a cell, given that it is occupied. To estimate $E(N|m_0)$, it is necessary to estimate μ and γ. This is not possible from m_0 alone.

The idea proposed here is to base the prediction of N on both m_0 and the number m_1 of cells containing exactly one individual. The optimal predictor in this case is

$$E(N|m_0, m_1) = m_1 + m_{>1}\frac{\mu - p(1)}{1 - [p(0) + p(1)]}, \quad (3)$$

where $m_{>1} = J - (m_o + m_1)$ and, from equation (1), $p(1) = \mu\gamma^{[\mu/(1-\gamma)]-1}$. As before, to estimate $E(N|m_o, m_1)$, it is necessary to estimate μ and γ. Unlike before, this is now possible.

Let the random variables M_o, M_1, and $M_{>1}$ be the number of cells containing no individuals, exactly one individual, and greater than one individual, respectively. The joint distribution of M_o, M_1, and $M_{>1}$ is multinomial with J trials and probabilities $p(0)$, $p(1)$, and $1 - [p(0) + p(1)]$. The log likelihood for μ and γ on the basis of the observed values m_o, m_1, and $m_{>1}$ is

$$\log L(\mu, \gamma) = m_o \log p(0) + m_1 \log p(1) + m_{>1} \log\{1 - [p(0) + p(1)]\}. \quad (4)$$

The maximum likelihood (ML) estimates $\hat{\mu}$ and $\hat{\gamma}$ of μ and γ are found by maximizing equation (4) numerically. This is an example of ML estimation based on a censored sample. These estimates can be substituted into equation (3) to estimate the optimal predictor of N. For later use, we will write this predictor as

$$\hat{N} = m_1 + m_{>1} f(\hat{\mu}, \hat{\gamma}), \quad (5)$$

where $f(\hat{\mu}, \hat{\gamma})$ is the ML estimate of $[\mu - p(1)]/\{1 - [p(0) + p(1)]\}$.

Some Results

We conducted a small simulation experiment to assess the performance of $\hat{N}$. For selected combinations of J, μ, and γ, we simulated independent negative binomial counts N_1, N_2, …, N_J and found the corresponding value $\hat{N}$. The process was repeated 200 times for each combination. In table 1, the bias and root mean squared error (RMSE) of $\hat{N}$, expressed as a proportion of the expected value $J\mu$ of N, is reported. It is clear from table 1 that $\hat{N}$ underpredicts N. However, the relative underprediction bias is small provided that J is not too small and γ is not too large. Similarly, the relative RMSE is generally small provided that J is not too small and γ is not too large. In overall terms, given its simplicity, $\hat{N}$ appears to perform reasonably well in most of the cases considered here.

Turning to an application to real data, figure 1 of He and Gaston (2000) shows the locations of $N = 591$ individuals of the midcanopy tree species *Dacryodes rubiginosa* in a 50-ha plot in the Pasoh Forest in Malaysia. We applied the prediction method described above to these data, which were kindly provided by F. He, using the grid of 25 × 25-m cells also shown in figure 1 of He and Gaston (2000). Of the $J = 800$ cells in this grid, $m_o = 525$ contain no individuals and $m_1 = 141$ contain exactly one individual. For these counts, $\hat{N} = 573$, which underpredicts N by only around 3%.

Table 1: Relative bias and root mean squared error (RMSE) of prediction for selected combinations of J, μ, and γ based on 200 simulated data sets

J, μ, and γ	Relative bias	Relative RMSE
200:		
1:		
2	−.016	.103
5	−.080	.314
10	−.172	.610
2:		
2	−.024	.137
5	−.068	.275
10	−.127	.500
500:		
1:		
2	−.002	.062
5	−.035	.154
10	−.055	.382
2:		
2	−.012	.082
5	−.023	.151
10	−.044	.222
1,000:		
1:		
2	.032	.041
5	−.018	.113
10	−.048	.261
2:		
2	−.013	.036
5	.000	.096
10	−.042	.164

It is possible to go beyond point prediction to construct a prediction interval for N by the following bootstrap procedure. Form a bootstrap sample of J cells by sampling with replacement from the original set of cells. Let m_o^* and m_1^* be the number of cells in this bootstrap sample that contain zero individuals and one individual, respectively. Let $\hat{\mu}^*$ and $\hat{\gamma}^*$ be the ML estimates of μ and γ based on m_o^* and m_1^*, and let $\hat{N}^* = m_1 + m_{>1} f(\hat{\mu}^*, \hat{\gamma}^*)$ be the predictor of N based on $\hat{\mu}^*$ and $\hat{\gamma}^*$. The lower and upper bounds of a $1 - \alpha$ bootstrap prediction interval for N are given by $2\hat{N} - \hat{N}^*_{1-\alpha/2}$ and $2\hat{N} - \hat{N}^*_{\alpha/2}$, where $\hat{N}^*_{1-\alpha/2}$ and $\hat{N}^*_{\alpha/2}$ are the upper and lower $(\alpha/2)$ quantiles, respectively, of the distribution of $\hat{N}^*$ found by repeated bootstrap sampling. For the Pasoh Forest data, the 0.95 prediction interval based on 200 bootstrap samples was (520, 612), which comfortably covers the true value.

Discussion

The purpose of this note has been to describe a simple occupancy-based predictor of abundance that is in the spirit of, but avoids the problem with, the predictor of He and Gaston (2000). The earlier predictor requires that each cell is searched until it is determined to be empty or until the first individual is found. The predictor proposed here requires that each cell is searched until it is determined to be empty or to contain exactly one individual or until the second individual is found. The feasibility of both of these sampling schemes will depend on circumstances. Censored sampling (and a related method called binomial sampling) has been used for different purposes in the management of insect and plant pests (e.g., Binns and Nyrop 1992; Gold et al. 1996).

The basic idea of this note can be extended by basing prediction on the numbers of cells containing exactly k individuals for $k = 0, 1, \ldots, K$ for a specified choice of K. Binns and Bostanian (1988) showed that, in estimating the parameters of the negative binomial distribution, little is gained by increasing the censoring point K beyond μ. This suggests that the method based on m_0 and m_1 alone is appropriate in situations where the combination of cell size and abundance results in an average cell count of around 1.

The analysis here has focused on the case where total abundance N is a random variable. Conlisk et al. (2007) also discussed the case in which N is treated as fixed but unknown. In that case, the cell counts follow the Polya-Eggenberger distribution with N and γ (but not μ) as parameters. The joint probability mass function of M_0 and M_1 can be found using the results of Charalambides (2005). This allows ML estimation of N based on m_0 and m_1 (but not on m_0 alone), although the required computation is somewhat delicate. In most practical cases, we would expect the two methods to give similar results.

Like other occupancy-based abundance predictors, the one proposed here assumes that the underlying distribution of cell counts is negative binomial. This or another distributional assumption cannot be checked from the censored counts alone. The negative binomial distribution is a flexible model commonly used for cell counts, so the assumption seems reasonable. Nevertheless, it would be interesting to assess the robustness of this predictor to departures from the negative binomial assumption.

Acknowledgments

Helpful comments from E. Conlisk, F. He, and two anonymous reviewers are acknowledged with gratitude.

Literature Cited

Binns, M. R., and N. J. Bostanian. 1988. Binomial and censored sampling in estimation and decision making for the negative binomial distribution. Biometrics 44:473–483.

Binns, M. R., and J. P. Nyrop. 1992. Sampling insect populations for the purpose of IPM decision making. Annual Review of Entomology 37:427–453.

Charalambides, C. A. 2005. Derivation of a joint occupancy distribution via a bivariate inclusion and exclusion formula. Metrika 62:149–160.

Conlisk, E., J. Conlisk, and J. Harte. 2007. The impossibility of estimating a negative binomial clustering parameter from presence-absence data: a comment on He and Gaston. American Naturalist 170:651–654.

Gold, H. J., J. Bay, and G. G. Wilkerson. 1996. Scouting for weeds, based on the negative binomial distribution. Weed Science 44:504–510.

He, F., and K. J. Gaston. 2000. Estimating species abundance from occurrence. American Naturalist 156:553–559.

———. 2007. Estimating abundance from occurrence: an underdetermined problem. American Naturalist 170:655–659.

Hui, C., M. A. McGeoch, B. Reyers, P. C. Le Roux, M. Greve, and S. L. Chown. 2009. Extrapolating population size from the occupancy-abundance relationship and the scaling pattern of occupancy. Ecological Applications 19:2038–2048.

Associate Editor: Priyanga Amarasekare
Editor: Mark A. McPeek