Estimating Abundance from Occurrence: An Underdetermined Problem

Fangliang He[1,*] and Kevin J. Gaston[2]

1. Department of Renewable Resources, University of Alberta, Edmonton, Alberta T6G 2H1, Canada;
2. Biodiversity and Macroecology Group, Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom

Submitted August 17, 2006; Accepted June 22, 2007; Electronically published September 4, 2007

Keywords: negative binomial distribution, presence-absence map, spatial scales, species abundance, species occupancy.

He and Gaston (2000) address a practically important yet challenging question: how can information about abundance be derived from presence-absence data? that is, how to get the most (abundance) from the least (occurrence). The challenge lies not so much in the method as in that it is an underdetermined problem, because we have only one distribution map in hand but two parameters (abundance N and aggregation parameter k) of the negative binomial distribution (NBD) moment estimator. Following Kunin (1998), we took an empirical approach to generating a second, coarser-scale map from the one available, so as to give a simultaneous equation system:

$$\begin{cases} N = M_1 k\left[\left(1 - \dfrac{m_1}{M_1}\right)^{-1/k} - 1\right] \\ N = M_2 k\left[\left(1 - \dfrac{m_2}{M_2}\right)^{-1/k} - 1\right] \end{cases}, \tag{1}$$

where M_1 and M_2 are the total numbers of cells for map 1 and map 2, respectively, and m_1 and m_2 are the numbers of occupied cells of the respective maps.

Critically, this empirical approach has been shown in practice to work well, or at least substantially better than existing alternatives, for estimating the abundances of more than 800 species of tree (He and Gaston 2000). However, the assumption of a constant NBD k across scales is a simple violation of a theoretical premise of the NBD. It is well established in statistics that k increases proportionally with scale; that is, if x_1 and x_2 are from an NBD with aggregation parameter k, then $y = x_1 + x_2$ follows an NBD with $2k$. We have long been aware of this property of the NBD and of the associated literature (e.g., the early contribution of Bliss and Owen [1958] and the synthetic volumes of Johnson and Kotz [1969] in statistics and Krebs [1999] in ecology). Indeed, the issue has been made explicit in many of our previous publications (e.g., He 1999; He and Gaston 2000; He and Reed 2006). For example, He (1999, p. 40) states, "At coarse-scale (with cells resulting from grouping C adjacent fine-scale cells), one can look at the number of empty cells resulting from the known number of occupied fine-scale cells that are placed in the coarse-scale cells. It will be assumed that this placement is also contagious, following the NBD with k replaced by Ck." The scale dependence is also both acknowledged and assessed by He and Gaston (2000, p. 556), but its effect on abundance estimation at the extent of 50-ha plots is observed to be small: "Although k in equation (5) changes with scale, the change is also limited (table 2)." This scale dependence of k is again recognized by He and Reed (2006, p. 99): "We realize that this assumption [constant k] does not necessarily hold in reality (Pielou 1957, Taylor et al. 1978). However, if the difference in scale for the two maps is not large, this assumption may be plausible."

Although the NBD is considered to be the model that best describes the distribution of the majority of species in nature (Boswell and Patil 1970; Krebs 1999), much empirical evidence has shown that in practice, the theoretical proportional dependence of k across scales is unlikely to hold. For example, Plotkin and Muller-Landau (2002) find that for tree species on Barro Colorado Island (BCI), Panama, k and spatial scale have the relationship $k(a) = 0.8604 + 0.002923a^{0.5450}$, where a is grid cell size in square meters. A similar relationship, $k(a) = k_0(a/a_0)^{0.55}$, is found to extrapolate k very precisely across spatial scales for trees on both the BCI plot and a similar plot in the Pasoh Reserve, Malaysia (He and Hubbell 2003;

* Corresponding author; e-mail: fhe@ualberta.ca.

Am. Nat. 2007. Vol. 170, pp. 655–659.
0003-0147/2007/17004-42031$15.00.
DOI: 10.1086/521340

however, the results of the Pasoh plot were not shown in that article). Based on this scaling function, we can estimate k at any other scale if k_0 at a base scale a_0 (e.g., 10 m × 10 m) is known. That $k = k_0C^{0.55}$, not $k = k_0C$, suggests that although the distributions of the 1,100 tree species in the BCI and Pasoh plots can be adequately modeled by the NBD at a single spatial scale, the model is not spatially invariant—the aggregation parameter does not follow the theoretical premise. Species are more aggregated at large scales than is predicted from the NBD of fine scales.

So, why do He and Gaston (2000) assume a constant k across scales? During the course of our study, we explored many alternative approaches to estimating abundance from occurrence, including using k and Ck in the two simultaneous equations (eq. [1]). But none of these worked. If we substitute k and Ck into equations (eq. [1]), they will not have any solution because the two lines do not cross, as has also been observed by Conlisk et al. (2007). Our assumption of a constant k across scales is the best solution we have found thus far in order to solve for N. We disagree with Conlisk et al. that this was a mistake or a misunderstanding of the nature of k. It was a purposeful and pragmatic assumption.

Conlisk et al. (2007) have shown that presence-absence data alone are not sufficient for accurately estimating k. The estimator of He and Gaston (2000) actually provides a lower bound of abundance. If these abundance estimates are to be improved on, it will be necessary either to estimate k from something other than the presence-absence data in hand (which fails to address the question of how best to estimate abundance simply from presence/absence) or to seek new methods that can better use this presence-absence information. One possible solution lies in the above empirical k-a scaling relationships for estimating k, but the use of such relationships will result in the suppression of any specifics of the distributions of individual species, greatly weakening the usefulness of such an approach. In what follows, we briefly elaborate a number of theoretical and practical issues that bear on the development of alternative approaches and on some of the other points made by Conlisk et al. (2007).

Two Limits of the Classical Occupancy Model

He and Reed (2006) consider the estimation of abundance from distribution as a classical occupancy problem that stems from the birthdays problem first explicitly described by P.-S. Laplace. Given that the birthdays of N people occur randomly and independently within a year of 365 days, Laplace wanted to know the number of days when nobody has a birthday. What we are interested in for the estimation of abundance from distribution is the analogous reverse problem: given u empty days out of 365, how many people were born? For random and independent occupancy, it is well established that u follows the classical distribution (Barton and David 1962; Johnson et al. 1993):

$$p(u) = \binom{M}{u}\sum_{i=0}^{M-u}(-1)^i\binom{M-u}{i}\left(1-\frac{u+i}{M}\right)^N, \qquad (2)$$

for $u = 0, 1, \ldots, M$, where M is the total number of cells (e.g., 365 days) and N is the total number of individuals (people).

The expectation and variance of u are, respectively,

$$\mathrm{E}(u) = M\left(1-\frac{1}{M}\right)^N, \qquad (3a)$$

$$V(u) = M(M-1)\left(1-\frac{2}{M}\right)^N + M\left(1-\frac{1}{M}\right)^N - M^2\left(1-\frac{1}{M}\right)^{2N}. \qquad (3b)$$

If N and M are large in such a way that mean abundance $\mu = N/M \le$ constant $(c) < \infty$, it is easy to show, using Taylor expansions, that the expectation and variance in equation (3) are approximately

$$\mathrm{E}(u) = Me^{-\mu}, \qquad (4a)$$

$$V(u) = Me^{-2\mu}(e^{\mu} - 1 - \mu). \qquad (4b)$$

This arrives at the same lower abundance limit that A. Chao (personal communication) developed from a more general model. Note that Kolchin et al. (1978) show that if $0 < \mu \le c < \infty$, the distribution of u is well approximated by a normal distribution with mean and variance given by equations (4a) and (4b), respectively.

Another limiting distribution of equation (2) is the Poisson. For a common species and a large area, M and N are both large. The expectation of the number of empty cells u is approximated by equation (4a). If also $N \gg M$, the chance that a cell will be empty is small. The classical occupancy model can be approximated by the Poisson distribution (Barton and David 1962):

$$p(u) = \frac{\lambda^u e^{-\lambda}}{u!}, \qquad (5)$$

for $u = 0, 1, 2, \ldots, M$, where $\lambda = \mathrm{E}(u)$.

Estimating Abundances of Aggregated Species

The limiting Poisson distribution suggests that equation (4a) should apply well to nearly saturated maps (or abundant species), where u is a rare event. However, the difference between equation (4a) and equation (3a) is very little (only fractions of an individual), and both can substantially underestimate the abundances of real species (He and Gaston 2000; He and Reed 2006). The majority of species have abundances much larger than that predicted from the classical occupancy model. The reason is simply that the individuals of most species are not randomly distributed but aggregated. The NBD estimate (eq. [1]) takes aggregation into account and is a useful model. However, the presence-absence data required for the He and Gaston (2000) model include no information on the specific arrangement of the occurrences; any permutation would give the same outcome, as observed by Conlisk et al. (2007).

The spatial information in an occurrence map could be used if the map is properly scaled. A potential approach is to use the fact that, for contagious placement of individuals, the distribution of empty cells u has the limiting normal distribution (Barton and David 1959; Kolchin et al. 1978),

$$f(u_1) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(u_1-\lambda_1)^2/2\sigma_1^2},$$

where $\lambda_1 = M_1(1 + N/M_1k)^{-k}$, $\sigma_1^2 = M_1e^{-2\mu_1}(e^{\mu_1} - 1 - \mu_1)$, $\mu_1 = N/M_1$, and u_1 and M_1 are, respectively, the number of empty cells and the total number of cells of the fine map.

At this fine scale, the number of empty cells resulting from the unknown number N of organisms is distributed according to $f(u_1)$. At the coarse scale (with cells resulting from grouping C adjacent fine-scale cells), one can look at the number of empty cells resulting from the known number of occupied fine-scale cells that are placed in the coarse-scale cells. This placement will also be contagious, following the NBD with k replaced by Ck. This assumption is somewhat ad hoc, but it does reflect the fact that contagion should not be confined to fine-scale cells only and the fact that one fine-scale cell being occupied will likely influence the probability that neighboring fine-scale cells are also occupied. Under this condition, we can formulate a second NBD, with k replaced by Ck, as

$$f(u_2) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(u_2-\lambda_2)^2/2\sigma_2^2},$$

where $\lambda_2 = M_2[1 + (M_1 - u_1)/M_2Ck]^{-Ck}$, $\sigma_2^2 = M_2e^{-2\mu_2}(e^{\mu_2} - 1 - \mu_2)$, $\mu_2 = (M_1 - u_1)/M_2$, and u_2 and M_2 are, respectively, the number of empty cells and the total number of cells of the coarse map.

A joint probability for observing u_1 empty cells at the fine scale and u_2 empty cells at the coarse scale is the product of the above two probability density functions (PDFs), $L = f(u_1)f(u_2)$, with the log-likelihood function being

$$l = -\left[\frac{1}{2}\ln(\sigma_1^2) + \frac{(u_1 - \lambda_1)^2}{2\sigma_1^2}\right] - \left[\frac{1}{2}\ln(\sigma_2^2) + \frac{(u_2 - \lambda_2)^2}{2\sigma_2^2}\right]. \quad (6)$$

However, it should be noted that this log likelihood is approximate because the PDFs at the two scales are unlikely to be independent.

We applied this two-scale method to estimating the abundances of the Pasoh species. The results showed that all three models performed well for rare species, and there is little difference among them. For relatively abundant species, they can differ substantially. For example, *Aidia wallichiana* is a small tree with 2,793 observed stems. Based on the occurrences in 12.5 × 12.5-m and 25 × 25-m areas, the abundances estimated by models (1), (3), and (6) are 2,200.8, 1,887.4, and 1,977.2, respectively. It is clear that model (6) to some degree outperforms the classical occupancy model (eq. [3]). Overall, the He and Gaston (2000) model outperforms all of the others: in the vast majority of cases, it still generates the best estimate of the three models. Nevertheless, model (6) suggests that occurrence data do contain spatial information useful for describing species distribution, and spatial scaling across scales should be a key approach to deducing this information because different occurrences at a fine scale will result in different occurrences at coarse scales (He and Condit 2007). The remaining challenge is how to make best use of this spatial arrangement in order more accurately to estimate abundances.

Does the Negative Binomial Distribution Estimator at Two Scales Work?

Although the constant k across scales assumed by He and Gaston (2000) is inconsistent with the theoretical scaling property of the negative binomial distribution (NBD), empirical results have suggested that the method is a practically sensible approach if abundance has to be estimated strictly from presence-absence data. Indeed, additional results for 300 tree species from the 50-ha BCI plot in Panama reinforce this conclusion (fig. 1). The NBD method clearly outcompetes the performance of the random-placement model (eq. [3]) that is considered as a "universal

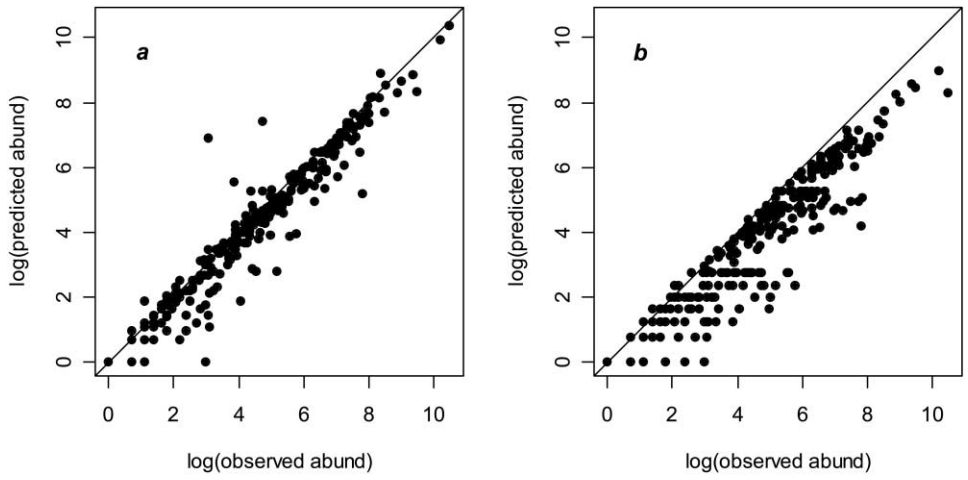

Figure 1: Negative binomial distribution–predicted versus observed abundances (*a*) and random-placement model–predicted versus observed abundances (*b*) for 300 Barro Colorado Island tree species. For each species, the two coarsest (unsaturated) distribution maps were used to estimate abundance. This means that for very abundant species, 10 × 10- and 20 × 20-m occurrence maps were used, while for rare species, 100 × 100- and 250 × 250-m maps were used. Note the estimation is much improved if 10 × 10- and 20 × 20-m maps are applied to all species.

lower bound" of abundance (A. Chao, personal communication). Thus, our method still represents the best empirical approach currently available. In situations where more accurate abundance estimates are needed, we would recommend the use of the two-scale NBD method (eq. [1]). At worst, the second scale map is redundant; there is no other cost in using this method.

Among other minor points, Conlisk et al. (2007) also offer the criticism that N is not soluble if $m = M$. This follows obviously, and rather trivially, from He and Gaston (2000). A map is saturated when $m = M$, but a saturated map contains no effective information about abundance (He and Condit 2007).

Final Remarks

How best to estimate abundance from distribution largely remains an unsolved problem. Model (3a) provides a lower bound of abundance. Empirical results have shown that the two-scale NBD model (eq. [1]) is practically useful and provides more accurate estimation than model (6) and the random-placement model (eq. [3a]). It is, however, important to understand that the assumption of constant k across scales is ad hoc and inconsistent with the theoretical scaling property of the NBD. Model (1) is obtained from an unconditional NBD in which N is considered as a random variable. While it is conceptually useful to distinguish the unconditional and conditional NBDs, as Conlisk et al. (2007) point out, in practice the difference between the estimation methods derived from the two models is very small, with no practical significance.

Two important questions remain to be investigated. The first is how to deduce spatial information from occurrence data and incorporate this into abundance estimation. We consider scaling to be a key to answering this question. The second question is how to develop methods for estimating abundances across large landscapes. Such methods are urgently needed and extremely relevant to management and conservation at regional or landscape scales. The methods so far derived are applicable only to small areas and fail to provide any reasonable estimation at landscape scales, for example, 10 or 100 km^2.

Acknowledgments

We thank A. Chao, W. Reed, and an anonymous reviewer for helpful comments on an earlier draft that improved this note. This work was supported by the Alberta Ingenuity Fund and the Natural Sciences and Engineering Research Council of Canada.

Literature Cited

Barton, D. E., and F. N. David. 1959. Contagious occupancy. Journal of the Royal Statistical Society B 21:120–133.

———. 1962. Combinatorial chance. C. Griffin, London.

Bliss, C. I., and A. R. G. Owen. 1958. Negative binomial distributions with a common *k*. Biometrika 45:37–58.

Boswell, M. T., and G. P. Patil. 1970. Chance mechanisms generating the negative binomial distributions. Pages 3–22 *in* G. P. Patil, ed. Random counts in models and structures. Pennsylvania State University Press, University Park.

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:XXX–XXX.

He, F. 1999. Estimating abundance from binary maps. MS thesis. University of Victoria, Victoria, BC.

He, F., and R. Condit. 2007. The distribution of species: occupancy, scale and rarity. Pages 32–50 *in* D. Storch, P. Marquet, and J. Brown, eds. Scaling biodiversity. Cambridge University Press, Cambridge.

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

He, F., and S. P. Hubbell. 2003. Percolation theory for the distribution and abundance of species. Physical Review Letters 91:198103.

He, F., and W. Reed. 2006. Downscaling abundance from the distribution of species: occupancy theory and applications. Pages 89–108 *in* J. Wu, K. B. Jones, H. Li, and O. L. Loucks, eds. Scaling and uncertainty analysis in ecology: methods and applications. Springer, Dordrecht.

Johnson, N. L., and S. Kotz. 1969. Discrete distributions. Houghton Mifflin, Boston.

Johnson, N. L., S. Kotz, and A. W. Kemp. 1993. Univariate discrete distributions. Wiley, New York.

Kolchin, V. F., B. A. Sevast'yanov, and V. P. Chistyakov. 1978. Random allocations. Winston, Washington, DC.

Krebs, C. J. 1999. Ecological methodology. 2nd ed. Addison Wesley, Menlo Park, CA.

Kunin, W. E. 1998. Extrapolating species abundance across spatial scales. Science 281:1513–1515.

Pielou, E. C. 1957. The effect of quadrat size on the estimation of the parameters of Neyman's and Thomas' distributions. Journal of Ecology 45:31–47.

Plotkin, J. B., and H. C. Muller-Landau. 2002. Sampling the species composition of a landscape. Ecology 83:3344–3356.

Taylor, L. R., I. P. Woiwod, and J. N. Perry. 1978. The density dependence of spatial behaviour and the rarity of randomness. Journal of Animal Ecology 47:383–406.

Associate Editor: Andrew R. Solow
Editor: Donald DeAngelis