Ecology, 91(12), 2010, pp. 3698–3706

Modeling spatial aggregation of finite populations

Tommaso Zillio and Fangliang He[1]

Department of Renewable Resources, 751 General Services Building, University of Alberta, Edmonton, Alberta T6G 2H1 Canada

Abstract. Accurate description of spatial distribution of species is essential for correctly modeling macroecological patterns and thus to infer mechanisms of species coexistence. The Poisson and negative binomial distribution (NBD) are most widely used to respectively model random and aggregated distributions of species in infinitely large areas. As a finite version of the Poisson distribution, the binomial distribution is used to model random distribution of species populations in finite areas. Despite that spatial aggregation is the most widespread pattern and no species in nature are distributed in infinitely large areas, no model is currently available to describe spatial aggregation for species distributed in finite areas. Here we develop a finite counterpart of the NBD to model aggregated species in finite landscapes. Similar to the NBD, this new model also has a parameter k measuring spatial aggregation. When $k \to \infty$, this model becomes the binomial distribution; when study area approaches infinite, it becomes the NBD. This model was extensively evaluated against the distributions of over 300 tree species in a 50-ha stem-mapping plot from Barro Colorado Island, Panama. The results show that when sampling area is small (relative to the study area), the new model and the NBD are of little difference. But the former correctly models spatial distribution at the finite limit at which the NBD fails. We reveal serious theoretical pathologies by using infinite models to approximate finite distribution and show the theoretical and practical advantages for using the new finite model for modeling species–area relationships, species occupancy and spatial distribution of rare species.

Key words: Barro Colorado Island, Panama; binomial distribution; finite area; negative binomial distribution; Poisson distribution; presence probability; spatial aggregation; spatial distribution; species–area relationship.

Introduction

Spatial distribution of species is a fundamental ecological structure that is formed by the interplay of many mechanisms (e.g., competition, dispersal, reproductive behaviors, habitat heterogeneity, and disturbances; Janzen 1970, Connell 1971, Chapin et al. 1989, Howe 1989, He et al. 1997, Harms et al. 2001, Hubbell 2001, Fragoso et al. 2003, Valencia et al. 2004, Seidler and Plotkin 2006, Law et al. 2009, Li et al. 2009, Wiegand et al. 2009) and that underpins many macroecological patterns (e.g., species–area curves, range distributions, beta diversity, and phylogeography; Arrhenius 1921, Condit et al. 2002, He and Legendre 2002, Green and Ostling 2003, Harte et al. 2005, Morlon et al. 2008, Shen et al. 2009). The two most widely used spatial models are Poisson probability distribution for modeling spatial randomness and negative binomial distribution (NBD) for spatial aggregation (Bliss and Fisher 1953, Boswell and Patil 1970, He and Gaston 2000, Plotkin and Muller-Landau 2002, Green and Plotkin 2007).

The Poisson distribution postulates that the probability that an individual of a species is found in a given area is independent of the presence of other individuals in the same area. The Poisson distribution can be easily simulated from the homogeneous Poisson point process (Diggle 2003). In contrast, the negative binomial distribution arises from contagious processes by which a cell that already has an individual would be more likely to contain more individuals, whereas those empty cells are apt to remain empty (Boswell and Patil 1970). Although there is no stationary point process that directly generates the NBD (Diggle 2003), the distribution can describe well point patterns generated from the Neyman-Scott, Thomas, or Cox point process.

Although the Poisson and NBD can arise from different spatial point processes and respectively describe random and aggregated patterns, the two models share a common feature: both describe the spatial distribution of species in infinitely large areas. This is evident from the following formulation of the two models where the domains (i.e., the range of n) vary from 0 to infinite:

$$\mathrm{Pois}(n) = \frac{\mu^n e^{-\mu}}{n!} \qquad n = 0, 1, 2, \ldots \tag{1}$$

$$\mathrm{NBD}(n) = \binom{n+k-1}{n} \left(\frac{\mu}{\mu+k}\right)^n \left(\frac{k}{\mu+k}\right)^k \qquad n = 0, 1, 2, \ldots \tag{2}$$

Manuscript received 30 November 2009; revised 26 April 2010; accepted 4 May 2010. Corresponding Editor: M. Fortin.
[1] E-mail: fhe@ualberta.ca

where μ is the expected number of individuals per cell, expressed as $\mu = NA_1/A$ where A_1 is cell size and N is the total number of individuals of a species in the entire study area A and k is the aggregation parameter of the NBD. Small k indicates high aggregation, large k less aggregation. The NBD becomes the Poisson model if $k \to \infty$.

Although the Poisson and negative binomial distributions are useful for many theoretical analyses, the assumption of infinite study area has never been met in reality. Because of this limitation, both models have been used as an approximation to what is true. As is well known, for a finite study area the exact model for spatial randomness is the binomial distribution:

$$\text{Bin}(n) = \binom{N}{n}(a)^n(1-a)^{N-n} \qquad n = 0, 1, 2, \ldots N \tag{3}$$

where $a = A_1/A$ is the relative cell size and A is the area of a study (it is finite). It is easy to show that when $N \to \infty$ and $a \to 0$ so that aN (=μ) is a constant, Eq. 3 becomes Eq. 1.

A widespread application of the above models is to model the occupancy of species, or the probability of species occurrence in an area (He and Gaston 2000, MacKenzie et al. 2002, Harte et al. 2005, Royle et al. 2005). For the binomial distribution the occurrence probability is simply $p_{\text{Bin}} = 1 - (1 - a)^N$, while the probabilities for the Poisson and the NBD are $p_{\text{Pois}} = 1 - e^{-aN}$ and $p_{\text{NBD}} = 1 - (1 + [aN/k])^{-k}$, respectively. The occurrence probability for the binomial distribution is 1 when $a = 1$, but it is smaller than 1 for other two models. This property renders the binomial distribution, not the Poisson, to be an important null model for the species–area relationship (Arrhenius 1921, Coleman 1981). For the binomial random placement species–area curve, the number of species predicted at $a = 1$, as expected, is equal to the total number of species of the study area. This is, however, not true for the species–area curve constructed from the Poisson or the NBD due to the problem of infinite area assumption. To solve this problem, He and Legendre (2002) modify the NBD occurrence probability to be $p_{\text{NBD}} = 1 - (1 - a)(1 + [aN/k])^{-k}$ to respect the constraint of $p_{\text{NBD}} = 1$ when $a = 1$. A similar solution is also given by Pielou (1975). But these are ad hoc modifications without any theoretical justification and they work only if we consider the occurrence probability, but not if we examine the full probability distribution.

Unlike the binomial distribution, which is a finite counterpart of the Poisson model for spatial randomness, a finite version of the negative binomial distribution for spatial aggregation is not yet available. The objective of this study is to develop a finite version of the negative binomial distribution. This new distribution describes spatial aggregation of species in finite areas. As a desirable property, this model is expected to approach the binomial distribution when species becomes randomly distributed. To be more specific, the model is expected to be related to other models as shown in Fig. 1.

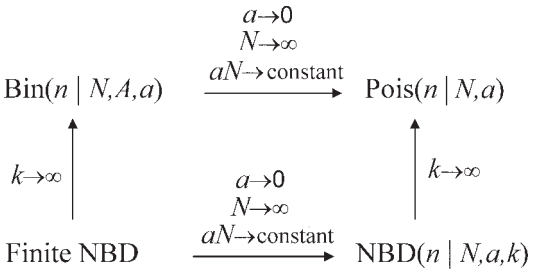

FIG. 1. The relationship between Poisson, binomial, negative binomial, and finite negative binomial probability models.

With few exceptions, ecological samples are always taken from finite areas or from finite populations. Like the negative binomial distribution, the finite negative binomial distribution should find wide applications in ecology.

Derivation of the Finite Negative Binomial Distribution

Stochastic migration or death/birth processes

Just as the negative binomial distribution can arise from a variety of random processes (Boswell and Patil 1970), the finite negative binomial distribution (FNBD) may also be generated from different processes. Here we propose two stochastic processes (immigration–emigration and birth–death) that yield the FNBD.

Suppose we have N individuals in an area A, of which $n \le N$ are in $A_1 \in A$. Let A_2 be the complementary area of A_1, so that $A = A_1 + A_2$. We can set up a dynamic process in which at every time step every individual, independent of all others, will relocate randomly in one of the two partitioning areas (A_1 and A_2). To simplify the notation and not to cause confusion, we denote $a = A_1/A$ (proportion of a sample area, not absolute area). Within a time step, a particular individual will have probability $a = A_1/A$ to be in A_1 and probability $1 - a = A_2/A$ to be in A_2, independent of its previous position. If we focus on A_1 we have the following rates of change:

$$\lambda(n) \sim \frac{A_1}{A}(N - n)$$

$$\mu(n) \sim \frac{A_2}{A}n \tag{4}$$

where $\lambda(n)$ is the rate for the transition $n \to n + 1$ in A_1 and $\mu(n)$ the rate for the transition $n \to n - 1$ in A_1. Rate of change $\lambda(n)$ can also be interpreted as the per capita immigration rate and $\mu(n)$ as the per capita emigration rate. The equilibrium solution of $P(n)$, the probability distribution of n in A_1, can be found by writing the master equation of the process:

$$\frac{\partial P(n)}{\partial t} = \lambda(n-1)P(n-1) + \mu(n+1)P(n+1) - (\lambda(n) + \mu(n))P(n) \quad (5)$$

and requiring that at equilibrium $\partial P(n)/\partial t = 0$. It is apparent that the solution has to respect the following condition, called *detailed balance*:

$$\mu(n+1)P(n+1) = \lambda(n)P(n).$$

By iterating this last equation, we find

$$P(n) = P(0)\prod_{i=0}^{n-1}\frac{\lambda(i)}{\mu(i+1)} \quad (6)$$

where $P(0)$ is determined by normalizing $\sum_n P(n) = 1$. Substituting the rates in Eq. 4 into the above equation, we obtain

$$P(n) = P(0)\prod_{i=0}^{n-1}\frac{A_1(n-i)}{A_2(i+1)} = P(0)\left(\frac{A_1}{A_2}\right)^n\frac{N!}{n!(N-n)!}.$$

This is a binomial distribution with normalizing factor $P(0) = (1 + A_1/A_2)^{-N}$. This distribution is the expected random distribution model since individuals move independently between A_1 and A_2.

Now let's add a clustering term to the above process, i.e., a term that will make the allocation of individual follow a contagious process by which individuals tend to allocate to the area that are already occupied. One way to create such contagious process is to modify the rates in Eq. 4 as

$$\lambda(n) \sim \left(\frac{A_1}{A} + c\frac{n}{N}\right)(N-n)$$

$$\mu(n) \sim \left(\frac{A_2}{A} + c\frac{N-n}{N}\right)n \quad (7)$$

where c is a clustering parameter: the higher the value of c, the more the individuals are attracted by the presence of others. These rates in Eq. 7 can be interpreted as: at one time step an individual will relocate in one of the two areas, with a probability that is proportional to a weighted average of the relative area and the relative abundance of the conspecifics in that area, with weights 1 and c, respectively.

Substituting the rates of Eq. 7 into Eq. 6, we obtain

$$\begin{aligned} P(n) &= P(0)\\ &= \prod_{i=0}^{n-1}\frac{(aN/c+i)(N-i)}{\Big((1-a)N/c+N-i-1\Big)(i+1)}\\ &= P(0)\frac{\Gamma(aN/c+n)\Gamma\Big((1-a)N/c+N-n\Big)N!}{\Gamma(aN/c)\Gamma\Big((1-a)N/c+N\Big)(N-n)!n!}\\ &= P(0)N!\binom{n+aN/c-1}{n}\\ &\quad\times\binom{N-n+(1-a)N/c-1}{N-n}. \end{aligned}$$

Recognizing

$$\binom{A}{B} = \binom{A}{A-B}$$

and using a Vandermonde's-like convolution (Graham et al. 1994:169),

$$\sum_{n=0}^{C-D}\binom{n+A}{B}\binom{C-n}{D} = \binom{A+C+1}{B+D+1}$$

we can normalize the above $P(n)$ and derive the desired finite negative binomial distribution model:

$$\mathrm{FNBD}(n\,|\,N,a,k) = \frac{\binom{n+k-1}{n}\binom{N-n+k/a-k-1}{N-n}}{\binom{N+k/a-1}{N}} \quad (8)$$

where $k = aN/c$ ($=\mu/c$) is the aggregation parameter that has the same definition as the k of the negative binomial distribution (see *Properties of the FNBD*). As expected, the distribution in Eq. 8 is symmetric for the simultaneous change $a \to 1-a$ and $n \to N-n$. Also note that this function is very different from a hypergeometric distribution, despite the apparent similarity.

The finite negative binomial model in Eq. 8 may also arise from a birth–death process instead of immigration–emigration dynamics as given in the above. In this case, the $\lambda(n)$ and $\mu(n)$ in Eq. (7) are the per capita birth and death rates, respectively. Furthermore, the birth (and death) events are not an independent but a contagious process, meaning that a birth has the tendency to induce more births and a death to induce more deaths. The birth/death argument may be more suitable to plant communities since plants are sessile.

Properties of the FNBD

The general shape of the FNBD is shown in Fig. 2, compared against the NBD and the binomial distributions. Here are some properties of the FNBD distributions:

1) Regardless of the value of k, the FNBD has expectation $\mathrm{E}(n) = aN = \mu$ which is the average number of trees per cell. This expected value must hold regardless of distribution, whether it is the binomial, negative binomial or Poisson.

2) The FNBD has variance $\sigma^2 = (1-a)\mu(k+\mu)/(k+a)$. This variance tends to the binomial variance $(1-a)\mu$ when $k \to \infty$, and to the negative binomial variance $\mu(1+\mu/k)$ when $a \to 0$, $N \to \infty$, and $aN \sim$ constant. Taking into account these relationships, we obtain all the limits shown in Fig. 1. Therefore, the NBD clustering parameter k is equivalent to the FNBD k.

3) The FNBD is invariant under the simultaneous substitution $a \leftrightarrow (1-a)$ and $n \leftrightarrow (N-n)$. This means

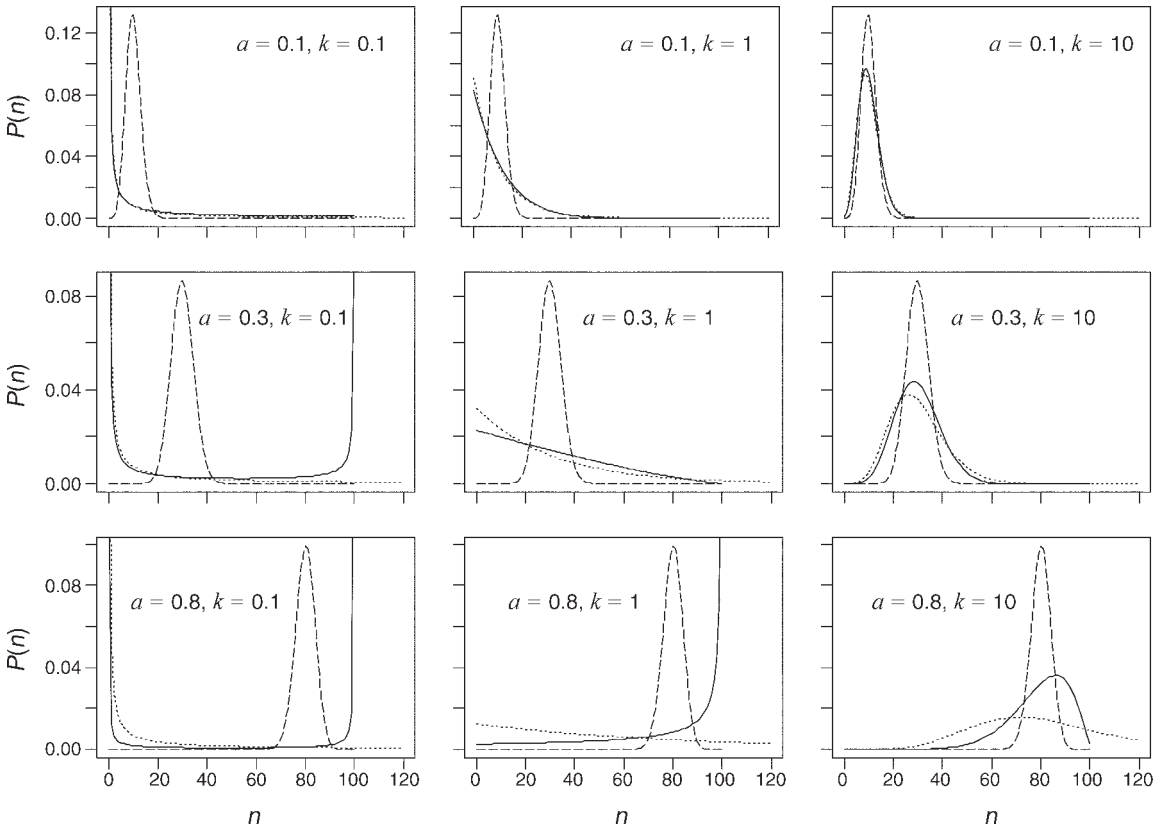

FIG. 2. Probability densities of the finite negative binomial (FNBD, solid lines) compared to the negative binomial (NBD, dotted lines) and the binomial (dashed lines) distributions. Here the total number of individuals of a species in the entire study area $N = 100$; sampling area $a = 0.1, 0.3, 0.8$; and clustering parameter $k = 0.1, 1, 10$. Contrary to the NBD, the FNBD vanishes for $n > N$. As expected, the FNBD and NBD are almost identical when sampling area a is small but progressively diverge when a becomes larger. At larger values of a the FNBD correctly tends to a spike at $n = N$, where n is the number of iterations. Note the expected value of all three distributions is $\mu = aN$, where μ is the expected number of individuals per cell.

that there is complete symmetry between the situations in A_1 and A_2.

4) When $a = 1$ the FNBD becomes a Kronecker delta $\delta_{n,N}$, i.e., it has value 1 when $n = N$ and 0 for all other values of n. In this respect it behaves like the binomial distribution, but unlike the negative binomial and the Poisson distributions. As highlighted before, this property is necessary for a correct treatment of finite samples.

5) When $a = 1/2$ and $k = 1$, the FNBD is constant with value $1/(N + 1)$. This result is reminiscent of the hypothesis of equal allocation probabilities (HEAP) model of Harte et al. (2005) and more recently of the MaxEnt model proposed by Harte et al. (2008).

Parameter estimation

Like the negative binomial distribution, the FNBD also has two parameters (N and k). They can be either estimated using the moment method or the likelihood method. The moment estimates can be directly derived from properties (1) and (2) of the previous section:

$$\hat{N} = \frac{\bar{n}}{a} \quad \text{and} \quad \hat{k} = \frac{(1-a)\bar{n}^2 - as^2}{s^2 - (1-a)\bar{n}}$$

where

$$\bar{n} = \frac{1}{m}\sum n_i \quad \text{and} \quad s^2 = \frac{1}{m}\sum (n_i - \bar{n})^2$$

where m is the number of quadrats.

The maximum-likelihood estimates of N and k can also be numerically solved by maximizing the likelihood function of the FNBD with respect to N and k:

$$L(N,k) = \prod_{i=1}^{m} \frac{\binom{n_i + k - 1}{n_i}\binom{N - n_i + k/a - k - 1}{N - n_i}}{\binom{N + k/a - 1}{N}}$$

where $\{n_i\}$ is the abundances in a set of sampled subareas. In numerical computation, it is easier to maximize the log-likelihood function $\log(L(N, k))$ than maximizing the likelihood function.

In the rest of this study N was assumed to be known since census data were used for testing the FNBD, k was the only parameter to be estimated and we used the maximum-likelihood method for estimating k.

TABLE 1. Comparison of the maximum-likelihood fitting of the finite negative binomial distribution (FNBD) and negative binomial distribution (NBD) for 190 Barro Colorado Island, Panama (BCI) species with abundances ≥ 50.

Scale (m × m)	a	χ^2 test		Likelihood ratio test		Correlation between FNBD k and other clumping indices		
		FNBD fails	NBD fails	FNBD wins	NBD wins	$r_{k\text{FNBD},k\text{NBD}}$	$r_{k\text{FNBD,dist}}$	$r_{k\text{FNBD,quadrat}}$
20 × 20	0.0008	1	1	0	0	0.997	0.629	−0.188
50 × 50	0.005	4	4	1	4	0.997	0.662	−0.0236
100 × 100	0.02	1	1	6	19	0.994	0.690	0.0622
200 × 200	0.08	0	0	32	75	0.939	0.691	0.0786
400 × 400	0.32	0	2	64	108	0.938	0.656	0.0985
500 × 500	0.50	1	5	75	110	0.916	0.675	0.170
600 × 600	0.72	16	8	120	69	0.802	0.474	0.256

Notes: Rarer species were not compared because the maximum-likelihood estimation of NBD was not reliable for rare species although it was not a problem for the FNBD; $a = A_1/A$ is the relative cell size, and A is the area of a study. The numbers under the columns for "χ^2 test" and "likelihood ratio test" report the number of species a model either favors or fails. If rare species were included, the FNBD would outperform the NBD in every test. The last three columns show the correlation of FNBD clustering parameter k with NBD k, clumping index of nearest distance, and clumping index of quadrat count. Note that $r_{k\text{FNBD,dist}}$ and $r_{k\text{FNBD,quadrat}}$ were calculated from data log–log transformation.

Empirical tests

We tested the FNBD, along with the NBD, using the tree distribution data from the 50-ha (1000 × 500 m rectangle) stem-mapping plot in Barro Colorado Island (BCI), Panama. In the plot, trees and saplings with diameter at breast height ≥1 cm were mapped and identified to species. There are 305 species in the BCI plot (1990 census).

To model the tree distribution, we divided the 50-ha plot into a grid of cells of varying sizes, or by randomly sampling subareas with sample size $m = 10$–1000, depending on the relative size of the subarea. The subarea sampling is necessary when working with a larger region where a small number of subareas can only be taken. In this study, when randomly sampling subareas, the shape of the subarea was a rectangle 2:1 (with the same shape of the overall plot). This shape of sampling area allowed us to study tree distribution in subareas that are larger than half of the 50-ha plot (i.e., $a > 0.5$). However, our analysis did not depend on the shape of sampling areas. For instance, cell size of 10 × 10 m can be equally used.

Using the maximum-likelihood method, we fitted the FNBD and NBD to all the BCI species. The goodness-of-fit was evaluated using the χ^2 test and likelihood ratio test. We further tested whether the FNBD and NBD k's were reliable measures of spatial aggregation by comparing the k values against two independent aggregation indices. The first one is nearest neighbor index calculated as r_a/r_e, where r_a is the nearest neighbor distance averaged over the trees of a species and r_e is the nearest neighbor distance expected from random distribution. A value of one indicates random distribution, larger than one regular distribution, and smaller than one aggregation. The second index is calculated from variance-mean quadrat count as $s^2/\bar{x}$, where $\bar{x}$ is the average number of trees per quadrat for a given quadrat size and s^2 is the variance of the quadrat count. It is one if the species is at random distribution. It is larger than one for aggregated species, and smaller than one for regular species.

We constructed species–area curves using those fitted FNBD and NBD for the BCI plot. The presence probability of the FNBD for a species is

$$P(n>0\,|\,N,a,k) = 1 - \frac{\Gamma(N+k/a-k)\Gamma(k/a)}{\Gamma(N+k/a)\Gamma(k/a-k)}. \quad (9)$$

By definition of Gamma function, it is easy to verify that the presence probability is 0 when $a \to 0$ and it is 1 when $a \to 1$. The species–area relationship is simply the sum of Eq. 9:

$$\langle S\rangle_a = \sum_{i=1}^{S} P(n>0\,|\,N_i,a,k_{i,a}) \quad (10)$$

where $\langle S\rangle_a$ is the average number of species at sampling area a, S is the total number of species in the community, N_i and $k_{i,a}$ are respectively the abundance of species i and the fitted value of k at each relative area a. The presence probability and the species-area relationship for the NBD can be similarly defined from Eq. 2.

RESULTS

We fitted the FNBD to all the BCI species and compared the performance of the FNBD and NBD for species with abundance ≥50 in Table 1. The results of the χ^2 test and likelihood ratio test are somewhat mixed: the performance of FNBD and NBD in describing species distribution varies across scales. It is clear that when cell size a is small or N is large, the FNBD and NBD are virtually identical, as expected. The superiority of the FNBD increases with cell size as the FNBD is the exact model while NBD is an infinite approximation to a finite problem. This is also reflected by the strong but decreasing correlation between FNBD k and NBD k with the increase of cell size (Table 1). The FNBD k is closely correlated with the nearest-neighbor-based clumping index ($r_{k\text{FVDB,dist}}$ in Table 1) but not so with the quadrat based clumping index ($r_{k\text{FNDB,quadrat}}$). Fig. 3 shows the fitted distributions for three species from BCI and the dependence of the aggregation parameter k's of

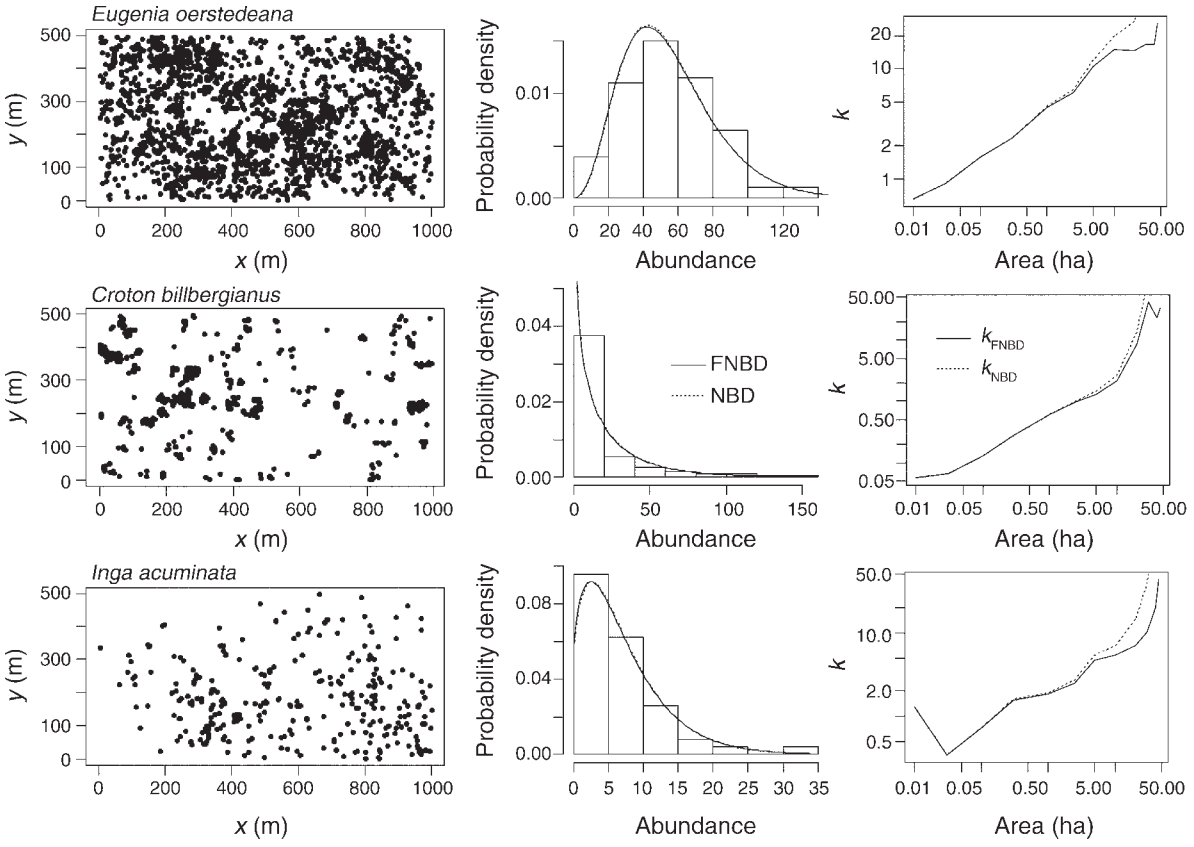

FIG. 3. Spatial distributions of three Barro Colorado Island, Panama (BCI) species (*Eugenia oerstedeana*, *Croton billbergianus*, *Inga acuminata*), their probability distribution of quadrat count (at scale = 100 × 100 m) and the fits of the FNBD and NBD, and the change of aggregation parameter k's of FNBD and NBD with sampling scale. Note that the fit of the FNBD and NBD to the three species are nearly identical (indistinguishable).

FNBD and NBD on the sampling scale. It is clear that k_{FNBD} and k_{NBD} are nearly identical at small scales, but differ at larger scales.

The relationship between k values and abundance for BCI species reveals that the species may be divided into two different groups: randomly dispersed (high k values) and clustered (intermediate k values) with a few species highly clustered (low k values; Fig. 4). Those highly aggregated species are evident only when using the FNBD, and comprises only the rare species with abundance < 30. In each of the two groups, k has a very weak but positive relationship with abundance. Note the values of k for the randomly distributed species are not accurate because a random distribution would have k_{FNBD} (and k_{NBD}) $= \infty$; indeed numerically the upper bound of the k's is of the order of 10^8 and thus the determination of the actual values of the k's in these cases are not reliable. Practically, because of the properties of NBD and FNBD, any k larger than 10 will make these two models indistinguishable from the Poisson and the binomial distribution, respectively.

It is worthwhile comparing the performance of the FNBD and NBD with rare species. The maximization of the log-likelihood function of the NBD often failed to converge when N is small (<50) but this was not a problem for the FNBD even if for singleton species ($N = 1$). For those rare species with which the maximum-likelihood function of the NBD appeared to converge, the estimation was not reliable. For example, for the four obviously aggregated species in Fig. 5 the likelihood maximization of the NBD leads to unrealistically large k values, incorrectly suggesting the species are randomly distributed. Only for *Marila laxiflora*, the NBD produces a reasonable k pointing to spatial aggregation of the species. In contrast, the FNBD always correctly models the distribution of all species.

As is known from Introduction, although the species–area curves constructed from the NBD and the FNBD are indistinguishable for small sampling areas, e.g., $a < 0.3$, the curve constructed from the NBD (and the Poisson distribution) is not appropriate for finite area as shown by Fig. 6. When a is larger than 0.3, the NBD species–area curve substantially underestimates the number of species, while the FNBD curve works very well. When $a = 1$ the FNBD species–area curve is exact by definition (equaling 305), while the NBD is always smaller than the true number of species of the community (Fig. 6b).

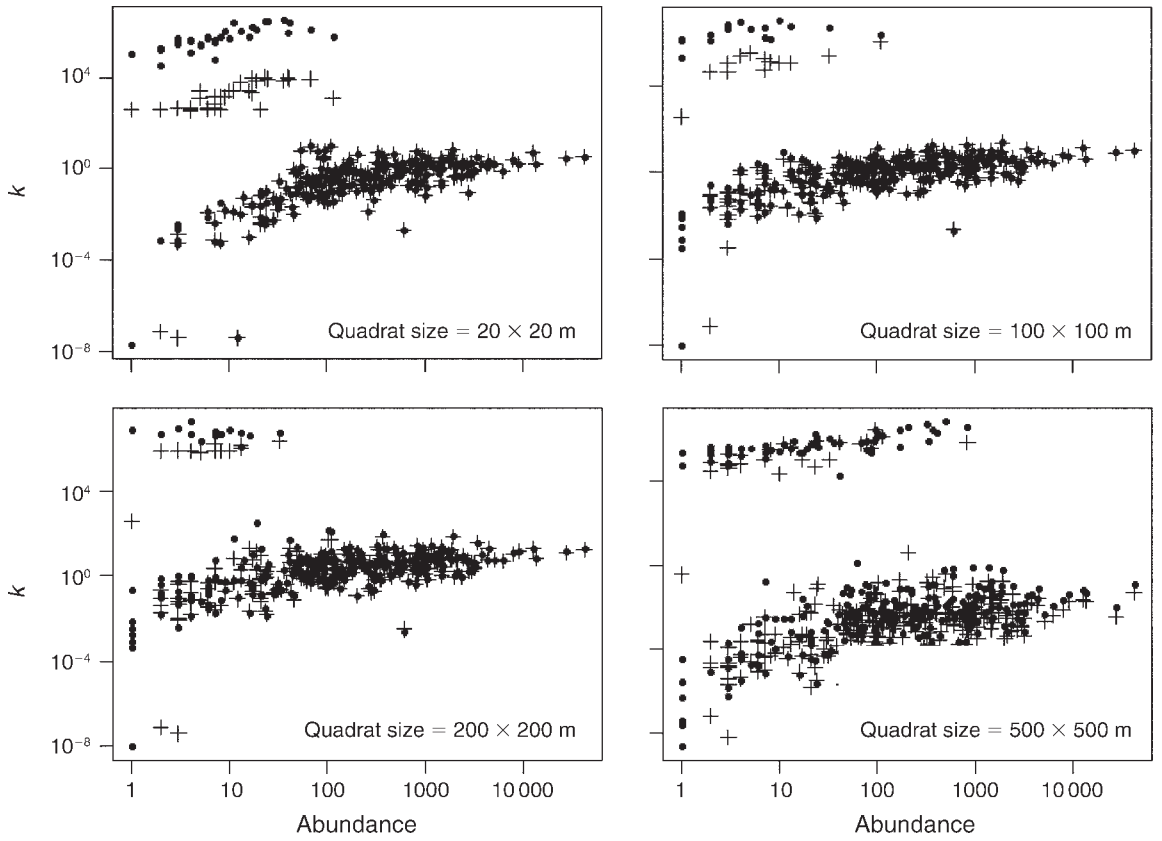

FIG. 4. Relationship between the clustering parameter k and abundance for the NBD (dots) and the FNBD (crosses) for the 305 BCI species, for quadrat size varies from 20 × 20 m up to 500 × 500 m. Note that because the maximization of the log-likelihood function for the NBD did not converge for many rare species ($N < 50$), only those species with $N \geq 50$ (190 species) are included in the figure.

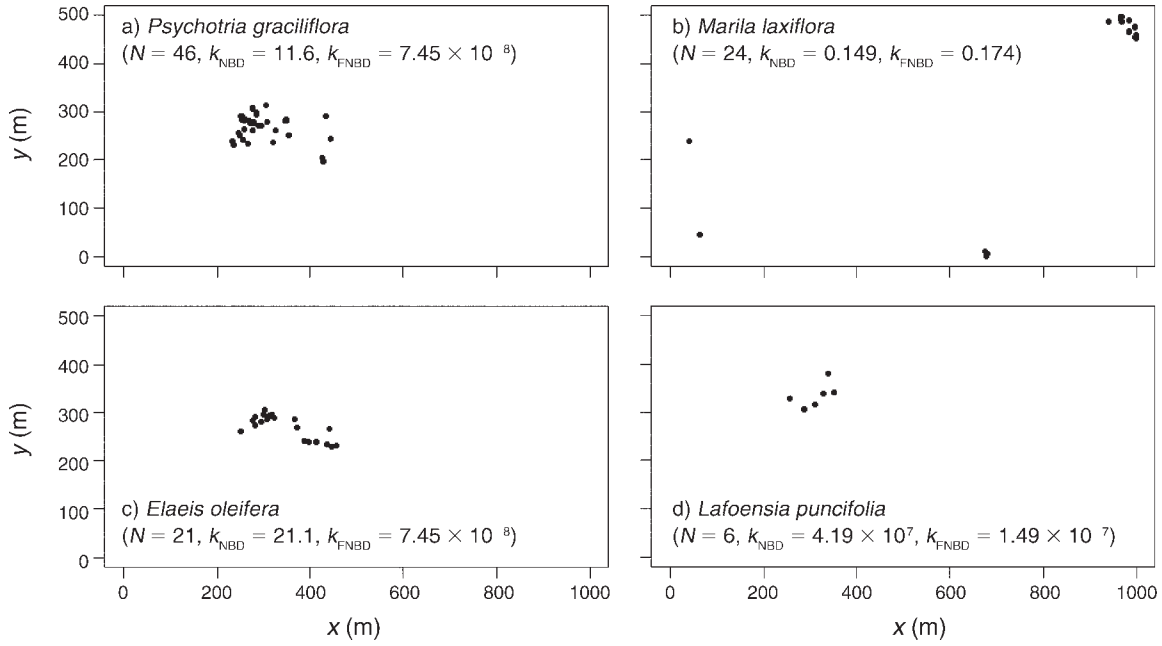

FIG. 5. Spatial distributions of four rare BCI species with $N < 50$. The k values (at cell size = 500 × 500 m) for the four species vary considerably between the NBD and FNBD, except for *Marila laxiflora*. NBD tends to inflate k_{NBD} (indicating random distribution) while FNBD has small k_{FNBD} that correctly detect the species as clustered.

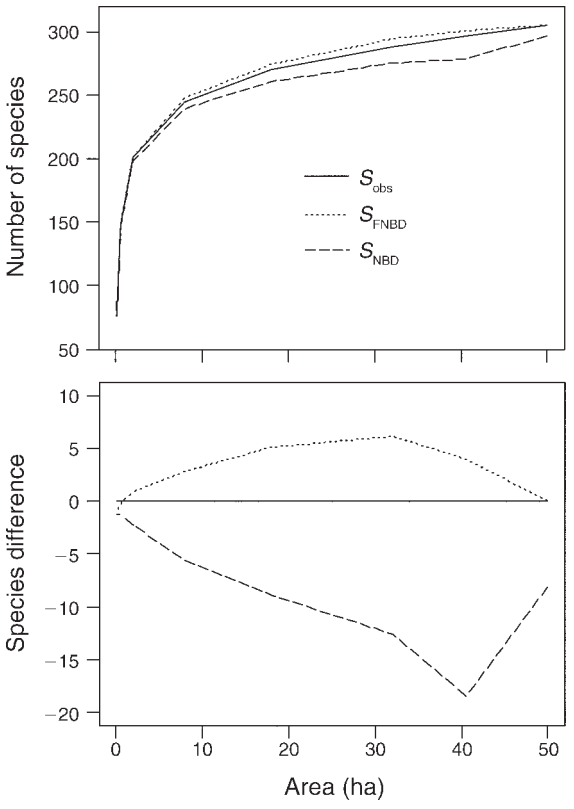

FIG. 6. (a) The NBD (dashed) and FNBD (dotted) species–area curves compared with the empirical curve (solid). (b) The difference between the NBD and FNBD species–area curves and the empirical species–area curve, showing the FNBD curve is a superior model.

DISCUSSION

Spatial models have played an increasingly important role in inferring mechanisms of species coexistence (Hubbell 2001, Seidler and Plotkin 2006, Law et al. 2009, Li et al. 2009, Wiegand et al. 2009) and in modeling diversity patterns (Condit et al. 2002, He and Legendre 2002, Harte et al. 2005, Morlon et al. 2008, Shen et al. 2009). Although the NBD is the most widely used model in the literature to describe spatial aggregation of species, it is only an approximation to problems of most applications where sampling area is small relative to the total area of a study. When sampling area is large the NBD is clearly inappropriate. The theoretical consequence by using a model for an infinite area (whether it is the Poisson distribution or NBD) to a finite area is serious as revealed by the pathology shown in Fig. 7. This pathology shows that whatever the value of sampling area a is, because n in the NBD is not upper bounded there is always a nonzero (albeit usually small) probability that we will sample $n > N$ individuals from an area that contains only N individuals! Using the infinite NBD to finite problem is amount to using a sampling with replacement to represent a problem that results from a sampling without replacement as noted by Plotkin and Muller-Landau (2002).

Another widely noticed consequence of using infinite models for finite area is reflected by the fact that the species-area curves constructed from the infinite models do not recover the total number of species at $a = 1$ (Fig. 6). The finite negative binomial distribution (Eq. 8) derived in this study is an exact model for spatial distribution in a finite area. Although the NBD can still be used as the limiting distribution for small a and large N, it is clear that it is not the "true" distribution. The NBD is just the result of a limit operation and the errors arising from its use, though small, are present at all scale.

An interesting distinction between the NBD and FNBD is how sampling area a and total abundance N are incorporated in the model. a and N are inseparable in NBD since they are coupled as a single parameter μ (tree density per cell; Eq. 2). In contrast, a and N are treated separately in the FNBD (Eq. 8). This very feature is linked to the failure of the NBD for describing a finite landscape: since it depends only on aN, the variable n (the number of trees in a cell) lacks a natural maximum (that we know to be N). The FNBD, on the contrary, goes naturally to zero when $n > N$.

The detection of spatial pattern for rare species has been elusive owing to insufficient sample size. Taylor et al. (1978) and He et al. (1997) find that spatial randomness is a dominant pattern for rare species in nature, while the opposite conclusion is made by Condit et al. (2000) due to the use of different measurements of spatial aggregation. As shown by the examples in Fig. 5, NBD and FNBD can also lead to opposite conclusions when they are used to model rare species. The NBD tends to erroneously detect rare species as being random, while the FNBD correctly identifies the rare species as clustered. As an extreme case, *Lafoensia punicifolia* (six individuals in the 50-ha BCI plot; Fig. 5d) was estimated

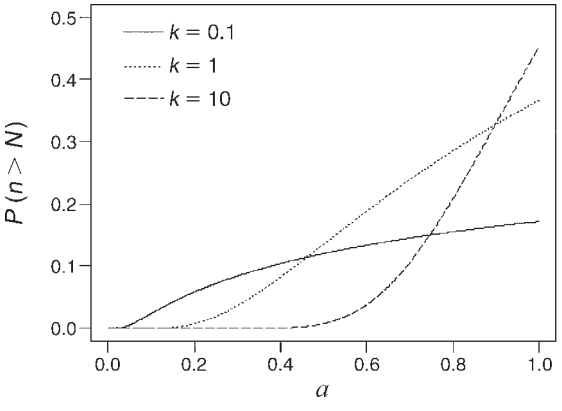

FIG. 7. Probability of sampling $n > N$ individuals from an area containing only N individuals vs. the relative area a when using the negative binomial distribution. Here $N = 100$ and $k = 0.1$, 1, and 10. This plot shows the inadequacy of the NBD to model finite populations.

to have $k_{NBD} = 41\,900\,000$ (suggesting a complete random distribution), but its k value of the FNBD model is nearly zero: $k_{FNBD} = 1.49 \times 10^{-7}$, indicating a highly clustered distribution. As shown in Table 1, FNBD k is a useful parameter for measuring spatial aggregation. It has a strong correlation with nearest neighbor distance clumping index but a very weak or no correlation with quadrat count variance-mean clumping index. This latter clumping index is known of little use for measuring spatial patterns (Hurlbert 1997). Our study has reinforced this result.

As noted from this study, the FNBD should have broad applications, not only to modeling spatial distribution of individual species but also to modeling macroecological patterns such as species-area relationships and species occupancy. Its applications can be further extended to modeling other important patterns of ecological communities such as endemics-area relationships, range-area relationships and the relationship between population density and spatial variance (Gaston 1996, Green and Ostling. 2003, Harte et al. 2005).

Acknowledgments

The authors thank Guillaume Blanchet for stimulating discussion and the Center for Tropical Forest Science for generously providing the BCI data. The work was supported by the Sustainable Forest Management Network, the GEOIDE, and the NSERC of Canada.

Literature Cited

Arrhenius, O. 1921. Species and area. Journal of Ecology 9:95–99.

Bliss, C. I., and R. A. Fisher. 1953. Fitting the negative binomial distribution to biological data. Biometrics 9:176–200.

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

Chapin, F. S., III, J. B. McGraw, and G. R. Shaver. 1989. Competition causes regular spacing of alder in Alaska shrub tundra. Oecologia 79:412–416.

Coleman, B. D. 1981. Random placement and species-area relations. Mathematical Bioscsciences 54:191–215.

Condit, R., et al. 2000. Spatial patterns in the distribution of tropical tree species. Science 288:1414–1418.

Condit, R., et al. 2002. Beta-diversity in tropical forest trees. Science 295:666–669.

Connell, J. H. 1971. On the role of natural enemies in preventing competitive exclusion in some marine animals and in rain forest trees. Pages 298–312 *in* P. J. D. Boer and G. R. Gradwell, editors. Dynamics of populations. PUDOC, Wageningen, The Netherlands.

Diggle, P. J. 2003. Statistical analysis of spatial point patterns. Academic Press, London, UK.

Fragoso, J. M. V., K. M. Silvius, and J. A. Correa. 2003. Long-distance seed dispersal by tapirs increases seed survival and aggregates tropical trees. Ecology 84:1998–2006.

Gaston, K. J. 1996. Species-range-size distributions: patterns, mechanisms and implications. Trends in Ecology and Evolution 11:197–201.

Graham, R. L., D. E. Knuth, and O. Patashnik. 1994. Concrete mathematics. Second edition. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, USA.

Green, J. L., and A. Ostling. 2003. Endemics–area relationships: the influence of species dominance and spatial aggregation. Ecology 84:3090–3097.

Green, J. L., and J. B. Plotkin. 2007. A statistical theory for sampling species abundance. Ecology Letters 10:1037–1045.

Harms, K. E., R. Condit, S. P. Hubbell, and R. Foster. 2001. Habitat associations of trees and shrubs in a 50-ha neotropical forest plot. Journal of Ecology 89:947–959.

Harte, J., E. Conlisk, A. Ostling, J. L. Green, and A. B. Smith. 2005. A theory of spatial structure in ecological communities at multiple spatial scales. Ecological Monographs 75:179–197.

Harte, J., T. Zillio, E. Conlisk, and A. B. Smith. 2008. Maximum entropy and the state-variable approach to macroecology. Ecology 89:2700–2711.

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

He, F., and P. Legendre. 2002. Species diversity patterns derived from species–area models. Ecology 83:1185–1198.

He, F., P. Legendre, and J. V. LaFrankie. 1997. Distribution patterns of tree species in a Malaysian tropical rain forest. Journal of Vegetation Science 8:105–114.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, New Jersey, USA.

Hurlbert, S. H. 1997. Spatial distribution of the montane unicorn. Oikos 58:257–271.

Howe, H. F. 1989. Scatter- and clump-dispersal and seedling demography: hypothesis and implications. Oecologia 79:417–426.

Janzen, D. H. 1970. Herbivores and the number of tree species in tropical forests. American Naturalist 104:501–528.

Law, R., J. Illian, D. Burslem, G. Gratzer, C. V. S. Gunatilleke, and I. Gunatilleke. 2009. Ecological information from spatial patterns of plants: insights from point process theory. Journal of Ecology 97:616–628.

Li, L., Z.-L. Huang, W.-H. Ye, H.-L. Cao, S.-G. Wei, Z.-G. Wang, J.-Y. Lian, I.-F. Sun, K.-P. Ma, and F. He. 2009. Spatial patterns of tree species in a subtropical forest of China. Oikos 118:495–502.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248–2255.

Morlon, H., G. Chuyong, R. Condit, S. Hubbell, D. Kenfack, D. Thomas, R. Valencia, and J. L. Green. 2008. A general framework for the distance-decay of similarity in ecological communities. Ecology Letters 11:904–917.

Pielou, E. C. 1975. Ecological diversity. John Wiley and Sons, New York, New York, USA.

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

Royle, R. A., J. D. Nichols, and M. Kéry. 2005. Modelling occurrence and abundance of species when detection is imperfect. Oikos 110:353–359.

Seidler, T. G., and J. B. Plotkin. 2006. Seed dispersal and spatial pattern in tropical trees. PLoS Biology 4:e344.

Shen, G.-C., M.-J. Yu, X.-S. Hu, X.-C. Mi, H.-B. Ren, I-F. Sun, and K.-P. Ma. 2009. Species–area relationships explained by the joint effects of dispersal limitation and habitat heterogeneity. Ecology 90:3033–3041.

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.

Valencia, R., R. B. Foster, G. Villa, R. Condit, J. C. Svenning, C. Hernandez, K. Romoleroux, E. Losos, E. Magard, and H. Balslev. 2004. Tree species distributions and local habitat variation in the Amazon: a large forest plot in eastern Ecuador. Journal of Ecology 92:214–229.

Wiegand, T., I. Martínez, and A. Huth. 2009. Recruitment in tropical tree species: revealing complex spatial patterns. American Naturalist 174:E106–140.