A NEW CLASS OF MODELS OF SPATIAL DISTRIBUTION

ERIN CONLISK,1,5 MICHAEL BLOXHAM,2 JOHN CONLISK,3 BRIAN ENQUIST,4 AND JOHN HARTE1

1Energy and Resources Group, University of California, Berkeley, California 94720 USA
2Mathematics and Computer Science, University of Redlands, Redlands, California 92373 USA
3Department of Economics, University of California, San Diego, California 92093 USA
4Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721 USA

Abstract. We analyze a new class of models of spatial distribution, developing mathematical properties and performing empirical tests. The models are based on a simple colonization rule operating on a rectangular grid. Two special cases within the class are traditional random placement and negative binomial models. Over three large data sets, these two cases are strongly outperformed by more flexible models within the class, in particular, models that allow more general patterns of aggregation. The models are simple and broadly applicable, with only one adjustable parameter, representing aggregation. Shortcomings of the colonization rule are studied, and extensions and applications of the models are discussed.

Key words: aggregation; colonization; negative binomial; serpentine grassland; spatial distribution; tropical forest.

INTRODUCTION

The abundances and spatial patterns of species shape our knowledge of biodiversity and vulnerability to extinction (Diamond et al. 1976, Ceballos and Ehrlich 2002, Wilson and Thomas 2004, Ceballos et al. 2005). Understanding spatial pattern helps to elucidate underlying biological processes (Brown 1995:19). Knowledge of pattern, e.g., range, abundance, and aggregation, combined with knowledge of process, e.g., reproduction, colonization, and the acquisition of resources, is essential to identifying dangerously small populations and protecting them with limited resources (Brown 1984, Soulé 1987, Hanski and Gyllenberg 1997). The more we understand natural systems, the better we will be at designing effective nature preserves, conducting efficient censuses, mitigating the effects of habitat loss, and predicting the effects of natural and anthropogenic disturbances.

A new model of the spatial distribution of species was presented in Harte et al. (2005). It made successful predictions about a wide variety of empirical patterns, within species and across communities. Here we build on Harte et al. (2005) by expanding the analysis to a broader class of models, providing much fuller mathematical properties and derivations, and performing new empirical tests based on likelihood and goodness of fit methods.

Traditional random placement and negative binomial models are special cases within the class studied here. All models in the class are simple, tractable, and broadly applicable. Detailed biological mechanisms are not modeled, circumventing the need for multiple, species-specific parameter settings. Simple models may provide good enough accuracy; and, if not, they may serve as comparisons against which to test more complex models.


The foundation of each model is a colonization rule through which individuals of a species locate in the cells of a rectangular grid until a given total abundance is achieved. There are two inputs: the abundances of the species studied and one adjustable aggregation parameter. The first colonizing individual locates in a cell purely at random, as if by seed dispersal from outside the area. As colonization proceeds, the location of each colonizing individual depends stochastically on the locations of previous colonists. The strength of this dependence is dictated by the aggregation parameter. When the parameter is large, a colonizing individual is likely to be born into an area with a relatively large number of conspecific individuals, as if seeds fall close to parents. When the parameter is small, a colonizing individual is less sensitive to previously colonized individuals, as if seeds may disperse more distantly, from parents inside or outside the area. A zero value of the parameter defines the random placement special case.

The outcome of colonization is a stochastic landscape matrix $N$, whose elements array abundances across the cells of the grid. A main goal of the theoretical analysis is a closed form solution for the distribution of the matrix $N$. The distribution summarizes predictions of the model. To our knowledge, plausible colonization...
rules typically do not allow closed-form solutions for the distribution of $N$. To achieve solutions here, the grid is created by sectioning, or subdivision, with the colonization rule applying subdivision by subdivision. Three specific models will be considered in detail, from which the extension to a class of related models will be apparent. In the “bisection model” (following Harte et al. 2005), the area under study is divided into left and right halves, each half is divided into upper and lower halves, and so on. The colonization rule locates a colonizing individual between the left and right halves, then between the between the upper and lower halves of that half, and so on. A second model, the “quadrisection” model, is based on successive quartering of an area. A third model, the “single-division” model, is based on a single division of the area into its final cells. This model lays a foundation for the bisection and quadrisection models. The single division model turns out to be the traditional negative binomial model conditioned on a given total abundance. The bisection and quadrisection models generate landscapes in which there is both within-cell clustering (a tendency for a cell to have many or few individuals) and cross-cell clustering (a tendency for well populated cells to aggregate, and for sparsely populated cells to aggregate). The single-division model generates within-cell clustering but not between-cell clustering. For this reason, empirical comparisons below strongly favor the bisection and quadrisection models over the single-division model. When the scale of a grid is changed from coarser to finer, part of the within-cell clustering for the coarser grid will likely shift to cross-cell clustering for the finer grid. An advantage of the bisection and quadrisection models is that they automatically incorporate this shifting through their sectioning schemes, using a single aggregation parameter to make predictions across all scales.

Under the colonization rule, the location of a colonizing individual depends only on previously colonized individuals of the same species. Thus, independence across species is implicit. Given independence, needed mathematical derivations are mostly concerned with a single species. Harte et al. (2005:192) discuss the empirical merit of independence.

Although our exposition is phrased as colonization of a barren area, a more general interpretation applies. As discussed in the concluding section, the distribution of $N$ is unchanged if random deaths occur, followed by replacement births located according to the colonization rule. Thus, the models could describe either initial colonization or mature communities in which initial colonists have long since died. Steady state is not assumed; the given total abundance might be equilibrium or transitional. Various biological mechanisms, though not explicitly modeled, are implicit in the given abundances.

Model comparisons and other empirical tests are based on three large plant censuses: a 64-m$^2$ California serpentine grassland census covering 37 182 individuals over 24 species (Green et al. 2003); a 50-ha moist tropical census in Panama covering 235 308 individuals over 305 species (Condit et al. 2000); and a 9.7-ha rectangular subarea, covering 12 851 individuals over 138 species, from a 16-ha dry tropical census in Costa Rica (Enquist et al. 1999).

**THREE MODELS FROM A BROADER CLASS**

*Grid geometry as the setting for colonization*

The grid geometry of each model is described by two indices ($c$, $i$) and by the spatial arrangement of divisions. A rectangular study area is first evenly divided into $c$ rectangular cells, then each of those cells is subdivided into $c$ rectangular cells, and so on through $i$ rounds of divisions. The number of final grid cells, or scale, is $c^i$. Fig. 1 demonstrates the spatial arrangement assumptions for the three models. The figure shows $4 \times 4$ final grids; larger grids are to be constructed by the apparent extensions. For the single-division model, the initial rectangle is divided once ($i = 1$) into $c$ final cells ($c = 16$).
on the figure), and the $c$ cells are arrayed as a square. For the bisection model, the initial rectangle is halved ($c = 2$) vertically, then each half is halved horizontally, and so on. Four rounds ($i = 4$) of halvings yield a $4 \times 4$ grid. For the quadrisection model, the initial rectangle is quartised into a $2 \times 2$ grid ($c = 4$), then each of those grid cells is divided into another $2 \times 2$, and so on. Two rounds ($i = 2$) achieve a $4 \times 4$ grid.

The grids are the setting for individual colonization. A colonizing individual is thought of as first located in one of the $c$ cells of the first division, then within that cell to one of the $c$ cells of its subdivision, and so on until the individual is located in a particular cell of the final grid. A simple colonization rule applies to every colonizing individual at every round of division, thus generating a spatial landscape for the species. The shading on Fig. 1 illustrates the assignments that would lead a single individual into cell $(2, 3)$ of the final grid.

For given $(c, i)$, the spatial arrangement of the subdivisions is critical in generating cross-cell clustering in the bisection and quadrisection models. Consider a variant on the quadrisection model with $(c, i) = (4, 2)$. At the first round, divide the initial area vertically into four strips and at the second round divide each strip horizontally into four final cells. The result is again a $4 \times 4$ grid. However, the different subdivision pattern, when combined with the colonization rule below, would yield only vertical cross-cell clustering (vertically adjacent cells tending to have similar abundances) and not also horizontal cross-cell clustering. For both the bisection and quadrisection models, the particular subdivision assumptions of Fig. 1 are important.

Examples of more radically different division schemes might involve one-dimensional grids, three-dimensional grids, or non-rectangularly divided areas. The challenge is to find schemes, such as bisection and quadrisection, which work well in applications. Other schemes, if desired, could be analyzed by much the same techniques.

**Single-division model**

The single-division model divides the initial area into $c$ final cells, indexed $j = 1, 2, \ldots, c$ in some convenient order. Individuals are born, one by one, into the cells. The colonization rule is

$$\Pr \left( \text{next individual is born into cell } j \mid \text{already in cells } 1, 2, \ldots, c \right) = \frac{\phi n_j + (1 - \phi)}{\phi (n_1 + n_2 + \cdots + n_c) + (1 - \phi)c}.$$

Here, $n_1, n_2, \ldots, n_c$ denote the numbers of individuals already in the cells prior to the next individual's birth (the $n_j$ are updated after each individual colonization), and $\phi$ is a parameter obeying $0 \leq \phi \leq 1$. As discussed following Theorem 1, (1) is a Polya-Eggenberger urn scheme.

The total number of individuals in all $c$ cells, taken as given, is denoted $n_0$, the 0 subscript to indicate the $i = 0$ grid, or the total area. Colonization proceeds until all $n_0$ individuals are distributed over the cells. Eq. 1, together with the assignment of indices to cells, determines the spatial landscape. At the outset of colonization, with all $n_j = 0$, the location probability is $1/c$ for every cell; the first individual is born into a cell purely at random, as if by seed dispersal from outside the area. Thereafter, if $\phi$ is positive, a new individual is more likely to be born into a cell where there are more potential parents (higher $n_j$). Thus, a larger value of the aggregation parameter $\phi$ generates more within-cell aggregation.

At one extreme, if $\phi = 0$, Eq. 1 equals $1/c$ for every individual (not just the first) at every colonization step. Thus, $\phi = 0$ defines the random placement model. At the other extreme, if $\phi = 1$, only the first individual is born to a cell purely at random; thereafter individuals have probability one of birth to the same cell as the first individual. Thus, $\phi = 1$ leads to total aggregation in a randomly chosen cell. Since certainty of total aggregation is not ecologically interesting, we assume $0 \leq \phi < 1$ from here on, excluding the $\phi = 1$ case. (We also exclude the $\phi < 0$ case because it requires delicate restrictions on the exact value of $\phi$ in relation to $n_0$; see Harte et al. 2005:184.) In the intermediate range $0 < \phi < 1$ between random placement and total aggregation, a larger $\phi$ makes an individual more likely, but not certain, to be born close to previously colonized individuals. The midpoint $\phi = 1/2$ yields algebraic simplifications.

As a simple example of Eq. 1, suppose that there are $c = 4$ cells, that $\phi = 1/2$, and that the first nine individuals are distributed over cells as

$$\begin{pmatrix} 2 \\ 4 \\ 0 \\ 3 \end{pmatrix}.$$

Then the assignment probabilities for the 10th individual are

$$\begin{pmatrix} 3/13 \\ 5/13 \\ 1/13 \\ 4/13 \end{pmatrix}.$$

as computed from Eq. 1. The 10th individual is most likely to locate in the most populated cell and least likely to locate in the empty cell. The example suggests how the assumptions of the model generate skewed distributions, with many empty or sparsely populated cells and a few heavily populated cells. This pattern is common in nature.

Several pieces of notation are required; most apply to all the models, not just the single-division model. Colonization yields an array, or landscape, of cell abundances. Let

$$N = (\text{matrix of cell abundances after colonization is complete}).$$

The $c'$ elements of $N$ sum to $n_0$. The model makes predictions about cell abundances comprising the matrix $N$. We test the predictions against observed $N$ matrices from plant censuses. Since $N$ is stochastic, model predictions are summarized by the probability density
function, hereafter PDF, of \( N \). It is a multivariate PDF, defined as the probability of observing \( N \). Let

\[
P_i(N|n_0, c, \phi) = \text{PDF of } N. \tag{3}\]

Since \( i \) (the number of rounds of division) is the index of an important recursion in the models below, it is given visual prominence as a subscript on \( P \). A number of model predictions arise from the univariate, or marginal, PDF of a typical element \( n \) of \( N \), by symmetry the same for every element of \( N \). Let

\[
P_i(n|n_0, c, \phi) = \text{PDF of a single element } n \text{ of } N. \tag{4}\]

The same function name \( P_i(\ldots) \) is used for both Eq. 3 and Eq. 4, leaving the argument, \( N \) or \( n \), to distinguish the multivariate from the univariate case.

The matrix

\[
Q(n_0, c, \phi) = \begin{bmatrix}
P_1(0|0, c, \phi) & 0 & \cdots & 0 \\
P_1(0|1, c, \phi) & P_1(1|1, c, \phi) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
P_1(0|n_0, c, \phi) & P_1(1|n_0, c, \phi) & \cdots & P_1(n_0|n_0, c, \phi)
\end{bmatrix}
\]

arrays univariate PDFs for the single-division model \((i = 1)\). The last row arrays probabilities \( P_i(n|n_0, c, \phi) \) for the \( n_0 + 1 \) possible values of \( n \), namely \( n = 0, 1, 2, \ldots, n_0 \). The preceding row gives the probabilities that would apply if the total number of individuals was instead \( n_0 - 1 \), and so on up the rows. Elements above the diagonal equal \( 0 \) because a cell abundance \( n \) cannot exceed the total number of individuals. Since each row forms a univariate PDF, the elements of each row sum to one. This matrix \( Q \) from the single-division model is important in analyzing the bisection and quadrisecion models in the next two sections.

Some necessary combinatorial notation is

\[
\binom{m}{k} = \frac{m!}{k!(m-k)!} \quad \binom{m}{k_1, k_2, \ldots, k_r} = \frac{m!}{k_1!k_2!\cdots k_r!} \quad \text{(with } k_1 + k_2 + \ldots + k_r = m). \tag{6}\]

\[
F(a, n) = \frac{\Gamma(a + n)}{\Gamma(a)\Gamma(n + 1)} = \binom{a + n - 1}{n} \quad \text{for } n \geq 1
\]

\[
= 1 \quad \text{for } n = 0. \tag{7}\]

In Eq. 7, \( a \) is nonnegative, \( n \) is nonnegative integer, and \( \Gamma \) denotes the gamma function. In the theorems, \( a = (1 - \phi)/\phi \), which varies from 0 to \( \infty \) as \( \phi \) varies from 1 to 0. Eq. 8 defines an \((n_0 + 1) \times (n_0 + 1)\) triangular matrix \( U \), illustrated on the right of Eq. 8 for \( n_0 = 4 \). \( U \) can be recognized as a signed Pascal triangle for which \( U = U^{-1} \).

Finally, negative binomial notation is needed. In a common ecological model, each cell of a landscape is allocated a number of individuals \( n \) through an independent draw from the negative binomial PDF, \( \text{NB}(a) = (1 - p)^{a}p^{n} \Gamma(n+a)/\Gamma(a+1)\Gamma(k) \), where \( p \) and \( k \) are parameters obeying \( 0 < p < 1 \) and \( k > 0 \); \( k \) is the “clustering parameter.” The multivariate PDF for a landscape \( N \) of \( c \) cells is then \( \text{NB}(N) = \text{NB}(n_1)\text{NB}(n_2)\ldots\text{NB}(n_c) \). Since, in this paper, we are considering landscapes with given total abundance \( n_0 \), we will consider the model defined by conditioning \( \text{NB}(N) \) on the restriction \( n_1 + n_2 + \ldots + n_c = n_0 \). It will be called the “conditional negative binomial model.”

Theorem 1 is mostly a summary of selected results from the mathematical literature, gathered here in a common notation for ecological application, in particular for application to the bisection and quadrisecion models below. Unified proofs are presented as Appendix A.

**Theorem 1.** Assume the single-division model \((i = 1)\) defined by colonization rule (1) and the preceding notation. Let \( n_1, n_2, \ldots, n_c \) denote the elements of \( N \) arranged in any order, and let \( a = (1 - \phi)/\phi \). Then

1.1. Every colonization sequence resulting in a given \( N \) has the same probability of occurring.

1.2. Multivariate PDF: \( P_i(N|n_0, c, \phi) = F(a, n_1)F(a, n_2)\ldots F(a, n_c)F(c|a, n_0). \)

1.3. Univariate PDF: \( P_i(n|n_0, c, \phi) = F(a, n)F(c - 1|a, n_0 - n)F(c|a, n_0). \)

1.4. Matrix of univariate PDFs: \( Q(n_0, c, \phi) = UAU \), where \( A \) is a diagonal matrix with diagonal elements \( F(a, 0)F(c|a, 0), F(a, 1)F(c|a, 1), \ldots, F(a, n_0)F(c|a, n_0). \)

1.5. Equivalence to conditional negative binomial model: Let \( k = a \). Then the multivariate and univariate PDFs of the conditional negative
binomial model are the same as those of the single-division model (1.2 and 1.3). Further, as \( n_0 \) and \( c \) approach infinity such that \( n_0/c = k(1 - p)/p \), the univariate PDF 1.3 approaches NB\((n)\).

Since early events in a colonization may have complex implications for later events, Theorem 1.1 may seem surprising. It is a critical simplification flowing from the functional form of the colonization rule (1). The rule can be interpreted as a Poly-a-Eggenberger urn scheme, leading to the closed form PDFs of Theorems 1.2 and 1.3, which are Poly-a-Eggenberger distributions (Johnson et al. 1997: Chapter 40 and references therein). By inspection of 1.2, the multivariate PDF is unchanged if the elements of the landscape N are arbitrarily permuted. That is, the single-division model predicts that there is no systematic pattern in the way more and less abundant cells are located relative to each other. That prediction conflicts with data, which show systematic clustering of abundant cells. This is the main reason for proceeding to the bisection and quadrisecion models below. Theorem 1.4 is a key algebraic connection to the models below, one more context in which Pascal’s triangle \( U \) appears in an important role (see Edelman and Strang [2004] for other examples).

Theorem 1.5 notes that the single-division PDFs are identical to conditional negative binomial PDFs (as derived in Barton and David [1959] for example). The parameter \( p \) from the unconditional PDFs does not appear in the conditional PDFs because its role is taken by the added parameter \( n_0 \). An implication of 1.5 is that the colonization rule in Eq. 1 is a chance mechanism leading to a negative binomial model. Numerous other mechanisms leading to an (unconditional) negative binomial distribution are surveyed in Boswell and Patil (1970). A particularly interesting further example is the population model of Nachman (2000), which may be contrasted to the single-division model here. The Nachman model builds on detailed ecological mechanisms (births, deaths, migration, density dependence); and, under certain transitional conditions, it yields an unconditional negative binomial PDF over “island” patches. Since the patches are not explicitly located, the model cannot generate cross-patch clustering. Although the single-division model does not generate cross-cell clustering either, it is a building block for the bisection and quadrisecion models below, which do. Another contrast to the Nachman model is our treatment of the total population \( n_0 \) as a given. However, the concluding section below will explain how our models can accommodate a birth and death mechanism (more restrictive than Nachman’s), and thus a changing \( n_0 \).

As \( \phi \to 0 \), the PDF of Theorem 1.2 approaches a symmetric multinomial PDF, with each cell having the same probability \( 1/c \) of capturing each individual; and the PDF of Theorem 1.3 approaches a binomial PDF with probability \( 1/c \) that the cell will receive a given individual. At the parameter midpoint \( \phi = 1/2 \) (\( a = 1 \)), the multivariate PDF collapses to

\[
P_1(N|n_0, c, 1/2) = 1/\binom{n_0 + c - 1}{c - 1}
\]

which does not depend on \( N \), implying that all landscapes N are equally likely. Following Harte et al. (2005), this case will be labeled HEAP for hypothesis of equal allocation probabilities. Using “broken stick” logic (MacArthur 1957), the denominator,

\[
\binom{n_0 + c - 1}{c - 1}
\]

can be recognized as the number of ways of allocating \( n_0 \) objects over \( c \) cells. Suppose that \( n_0 + c - 1 \) objects are lined up evenly along a stick, that \( c - 1 \) of the objects are selected at random, and that each of these \( c - 1 \) objects is replaced by a break in the stick. The result is \( c \) stick pieces, each carrying a number (possibly 0) of the remaining \( n_0 \) objects. If these remaining objects are thought of as individuals, the result is a division \( (n_1, n_2, \ldots, n_c) \) of the \( n_0 \) individuals, which can be thought of as a landscape. All such landscapes are equally likely, and the number of ways of choosing the \( c - 1 \) breakpoints is

\[
\binom{n_0 + c - 1}{c - 1}
\]

Fig. 2a illustrates the qualitative character of the univariate PDF in relation to \( \phi \). (Fig. 2b and 2c, for the bisection and quadrisecion models, are discussed in the next two subsections.) Fig. 2a plots \( P_i(n|n_0, c, \phi) \) against \( n \) for \( \phi = 0, 0.25, 0.50, 0.75, \) assuming \( n_0 = 617 \) individuals and \( c = 256 \) grid cells. The horizontal axis goes only to \( n = 7 \) rather than to \( n = n_0 = 617 \) because the probabilities are very small after \( n = 7 \). The most humped plot is for \( \phi = 0 \) and the successively more skewed curves are for the successively higher values of \( \phi \). As the aggregation parameter \( \phi \) increases from zero toward one, individuals tend to cluster in fewer cells, increasing the probabilities of low and high abundance cells at the expense of intermediate cells. The value \( n_0 = 617 \) is the actual number of plants for one of the species, Phlox gracilis, in the serpentine grassland data set. The unconnected dots on the figure show the empirical frequencies for this species, where an empirical \( P(n) \) is defined as the number of cells holding \( n \) individuals divided by the total number of cells \( c \), here 256. Given variation in empirical frequencies, we don’t expect a close fit of the dots to any of the theoretical PDFs. However, the figure is typical in the sense that the prediction for the random placement value \( \phi = 0 \) is a poorer fit than the prediction for an appropriate positive \( \phi \).

### Bisection model

In this model, a colonizing individual born into the \( i = 0 \) total area is viewed as locating in the left or right half.
Eq. 1, where the $\phi = 0$ extreme again defines the random placement model since Eq. 9 collapses to $1/2$ at every step. As $\phi$ increases within the interval $0 \leq \phi < 1$, a colonizing individual is more strongly attracted to the locations of previously colonized individuals, promoting aggregation.

Although colonization events are sequentially dependent, a certain reordering of events simplifies analysis without violating dependencies. Consider two sequences:

Sequence 1. Each individual colonizes to its final location before the next individual colonizes.

Sequence 2. All $n_0$ individuals colonize to their locations in the $i=1$ grid before any individual colonizes to its location in the $i=2$ grid. Then all individuals colonize to their $i=2$ locations before any individual colonizes to its $i=3$ location. And so on.

Sequence 1 is the natural ecological order. Sequence 2, though artificial, implies the same probabilities of final location. This is true because, for any two colonization events in Sequence 1 such that the probability of the latter depends on the outcome of the former, the two dependent events are still in the same order relative to one another under Sequence 2.

Sequence 2 allows us to analyze the final landscape of the bisection model as a recursion on $i$, where each step is an application of the single-division model with $c = 2$. This recursion approach is from Harte et al. (2005). To get from the $i=0$ grid to the $i=1$ grid, use the single-division model to allocate the original $n_0$ individuals between the two cells at $i=1$. To get from $i=1$ to $i=2$, use the single-division model twice, once to allocate the individuals on the left of the $i=1$ grid into its vertical half-cells, and a second time to allocate the individuals on the right of the $i=1$ grid into its vertical half-cells, thus making four cells at $i=2$. And so on. Going from the $i$ grid to the $i+1$ grid uses the single-division model $2^i$ times, implying $1 + 2 + 2^2 + \ldots + 2^{i-1} = 2^{i-1}$ applications of the single-division model in total.

At each of these Sequence 2 applications of the bisection model, a highly (or sparsely) populated “parent cell” is divided into two cells. The two cells are adjacent and are likely to be similar in abundance because of the common parentage. This is the mechanism that promotes cross-cell clustering in the final grid of the bisection model.

It follows from the Sequence 2 viewpoint that the multivariate PDF of the final landscape $N$ is a product of the component single-division PDFs. To construct that product, we need the list of all two-way divisions $(n_1, n_2)$ leading to the final landscape $N$:

\[
S(N) = \left[ \text{set of two-way divisions } (n_1, n_2) \right] \text{ leading to the final landscape } N. \tag{10}\]

The multivariate PDF is then the product

\[
P_i(N|n_0, 2, \phi) = \prod_{(n_1, n_2) \in S(N)} P_i[(n_1, n_2)|n_1 + n_2, 2, \phi]. \tag{11}\]

In Eq. 11, the typical factor $P_i[(n_1, n_2)|n_1 + n_2, 2, \phi]$ is the probability that the latter $n_2$ individual colonizes to its final location before the next individual colonizes.
the probability that \( n_1 + n_2 \) individuals will divide as \( (n_1, n_2) \), which is a single-division PDF \( (i = 1) \) for two cells \( (c = 2) \). The functional form of \( P_i[(n_1, n_2)|n_1 + n_2, 2, \phi] \) is given by Theorem 1.2. Fig. 3 illustrates the construction of \( S(N) \) and computation of \( P_i(N \ldots) \). Theorem 2 characterizes the bisection model.

**Theorem 2.** Assume the bisection model defined by colonization rule (Eq. 9). Then:

1. Every colonization sequence resulting in the same \( N \) has the same probability of occurring.
2. Multivariate PDF

\[
P_i(N|n_0, 2, \phi) = \prod_{(n_1, n_2) \in S(N)} \frac{F(a, n_1)F(a, n_2)}{F(2a, n_1 + n_2)}.\]

3. Univariate recursion

\[
P_i(n|n_0, 2, \phi) = \sum_{q=n}^{n_0} P_{i-1}(q|n_0, 2, \phi) \frac{F(a, n)F(a, q-n)}{F(2a, q)}.\]

4. Univariate PDF

\[
P_i(n|n_0, 2, \phi) = \left( \frac{n_0}{n} \right) \sum_{q=n}^{n_0} (-1)^{q-n} \left( \begin{array}{c} n_0 - n \\ 2a \end{array} \right) \frac{F(a, q)}{F(2a, q)}.\]

Derivations are in Appendix B. Theorem 2.1 is inherited from 1.1. Theorem 2.2 is Eq. 11 with \( P_i[(n_1, n_2)|n_1 + n_2, 2, \phi] \) replaced by its solution from Theorem 1.2. Theorem 2.3 is a generalized form of a recursion on the grid index \( i \) that was central to the analysis in Harte et al. (2005: Eq. 23). It gives univariate probabilities of grid \( i \) as a weighted sum of univariate probabilities of grid \( i-1 \). The summation index \( q \) runs over the possible values for the number of individuals in a typical cell of grid \( i-1 \). The typical weight \( F(a, q)F(a, q-n)F(2a, q) \) can be recognized as \( P_i(n|q, 2, \phi) \) from 1.3, which is a typical element of the \( Q \) matrix. In the HEAP special case \( \phi = 1/2 \), the weights simplify to \( 1/(q+1) \). The initial condition for the recursion is that all individuals are contained in the single cell at index \( i = 0 \) (the whole area) and thus that \( P_i(n_0|n_0, 2, \phi) = 1 \) and \( P_i(n|n_0, 2, \phi) = 0 \) for all \( n < n_0 \). The closed-form solution for the univariate PDF, stated in Theorem 2.4, is obtained by solving 2.3. In the derivation, the recursion on \( i \) is put in vector-matrix form and recognized as a Markov chain on the index \( i \) with transition probability matrix \( Q \). The chain is solved using the Pascal triangle relation from Theorem 1.4. From a computational viewpoint, the recursion of Theorem 2.3 is often easier to compute than the explicit solution of 2.4. Due to alternating signs and high order factorials, the sum in 2.4 can be overwhelmed by rounding error when \( n_0 \) is large.

Fig. 2b shows univariate PDF plots for the bisection model. For a given \( \phi \), they are steeper than the single-division plots on Fig. 2a because aggregation in the bisection model is induced by cross-cell as well as within-cell clustering, allowing a given \( \phi \) to induce a greater degree of aggregation. In particular, the empty cell
probability \( P_i(0 | \ldots) \) is larger in the bisection than the single-division model, all else being equal.

**Quadrisection model**

In this model, a colonizing individual born into the \( i = 0 \) total area locates in one of the four quadrants (\( c = 4 \)) of the \( i = 1 \) grid, within that quadrant into one of its four sub-quadrants in the \( i = 2 \) grid, and so on until a final cell is reached at the final value of \( i \). The illustrative individual on Fig. 1c locates upper right at \( i = 1 \) and lower left at \( i = 2 \).

We study the quadrisection model to see if it fits the data better than the bisection model. Quadrisection has the appealing feature that it does not require cells to change shape or orientation at each division. In the bisection model, the long-to-short ratio remains constant with each division if and only if the total area has long-to-short ratio of \( \sqrt{2} \) to 1. Even so, the long dimension of cells changes orientation, vertical vs. horizontal, at each division. In contrast, quadrisection leaves all cells at every division with the same long-to-short ratio and orientation.

The colonization rule (Eq. 12) dictates which quadrant applies at each round of division

\[
\Pr(\text{direction } j) = \frac{\phi n_j + (1 - \phi)}{\phi(n_1 + n_2 + n_3 + n_4) + 4(1 - \phi) }. \tag{12}
\]

Here \( n_1, n_2, n_3, n_4 \) denote the numbers of individuals already in the four quadrants of the cell in question, where the indices \( j = 1, 2, 3, 4 \) label the quadrants. Eq. 12 can be interpreted as the \( c = 4 \) special case of the single-division rule (Eq. 1), or as the \( c = 4 \) variant of the bisection rule (Eq. 9). The aggregation parameter \( \phi \) has the same role as before, with \( \phi = 0 \) the random placement case, and with larger \( \phi \) in the interval \( 0 \leq \phi < 1 \) increasing the tendency toward aggregation.

As with the bisection model, the quadrisection model can be viewed in Sequence 1 or Sequence 2 terms, and we use the latter for analysis. To get from the \( i = 0 \) grid to the \( i = 1 \) grid, use the single-division model to allocate the original \( n_0 \) individuals over the four cells of the \( i = 1 \) grid. To get from \( i = 1 \) to \( i = 2 \), use the single-division model four times, once for each of the four cells of the \( i = 1 \) grid. And so on. Going from the \( i \) grid to the \( i + 1 \) grid uses the single-division model \( 4^i \) times, implying \( 1 + 4 + 4^2 + \ldots + 4^{i-1} = (4^i - 1)/3 \) applications of the single-division model in total.

It follows from this Sequence 2 viewpoint that the multivariate PDF of the final landscape \( \mathbf{N} \) is a product of the component single-division PDFs. To construct that product, the list of all four-way divisions \( (n_1, n_2, n_3, n_4) \) leading to the final landscape \( \mathbf{N} \) is needed. Thus, modifying the definition of \( \mathbf{S}(\mathbf{N}) \) from the bisection model, let

\[
\mathbf{S}(\mathbf{N}) = \left\{ \text{set of four way divisions } (n_1, n_2, n_3, n_4) \mid \text{leading to the final landscape } \mathbf{N} \right\}. \tag{13}
\]

The multivariate PDF then equals the product

\[
P_i(\mathbf{N}|n_0, 4, \phi) = \prod_{(n_1, n_2, n_3, n_4) \in \mathbf{S}(\mathbf{N})} P_i([n_1, n_2, n_3, n_4]|n_1 + n_2 + n_3 + n_4, 4, \phi) \tag{14}
\]

which is analogous to Eq. 11 for the bisection model. The typical factor on the right of Eq. 14 is the probability that \( n_1 + n_2 + n_3 + n_4 \) individuals will divide as \( (n_1, n_2, n_3, n_4) \), which is a single-division PDF \( (i = 1) \) for a division into four cells (\( c = 4 \)). The functional form of the probability is given by 1.2. Fig. 3 illustrates the construction of \( \mathbf{S}(\mathbf{N}) \) and computation of \( P_i(\mathbf{N}|\ldots) \). Theorem 3 characterizes the quadrisection model.

**Theorem 3.** Assume the quadrisection model defined by colonization rule (Eq. 12). Then:

1. Every colonization sequence resulting in a given \( \mathbf{N} \) has the same probability of occurring.
2. **Multivariate PDF**

\[
P_i(\mathbf{N}|n_0, 4, \phi) = \prod_{(n_1, n_2, n_3, n_4) \in \mathbf{S}(\mathbf{N})} \frac{F(a, n_1)F(a, n_2)F(a, n_3)F(a, n_4)}{F(4a, n_1 + n_2 + n_3 + n_4)}. \tag{15}
\]

3. **Univariate recursion**

\[
P_i(n|n_0, 4, \phi) = \sum_{q=0}^{n0} P_{i-1}(q|n_0, 4, \phi) \frac{F(a, n)F(3a, q - n)}{F(4a, q)}. \tag{16}
\]

4. **Univariate PDF**

\[
P_i(n|n_0, 4, \phi) = \left( \frac{n0}{n} \right) \sum_{q=0}^{n0} \left[ (-1)^{q+n} \binom{n0 - n}{q-n} \frac{F(a, q)}{F(4a, q)} \right]^i. \tag{17}
\]

This theorem is so closely parallel to Theorem 2 that Appendix derivations are not given. The \( c \) indices are raised from 2 to 4, and \( n_1, n_2, n_3, n_4 \) are added to Theorem 3.2.

Fig. 2c illustrates univariate PDF plots for the quadrisection model, in comparison to those for the single-division and bisection models on Fig. 2a and 2b. For a given \( \phi \), the quadrisection plots are steeper than the single-division plots because the quadrisection model reflects cross-cell as well as within-cell clustering, but less steep than the bisection plots because the quadrisection model involves fewer divisions and is thus closer to the single-division model.

**Mean, variance, and higher moments of the univariate PDFs**

For all three models, the mean of the univariate PDF is total abundance per cell, \( E[n] = n0/c \), which does not depend on \( \phi \). The variance and higher moments are more complicated, differ by model, and are very
sensitive to $\phi$. The $r$th factorial moment $E[n(n-1)\cdots(n-t+1)]$ is simple to present. It implies the ordinary moments $E[n^r]$. For example, $E[n^2] = E[n(n-1)] + E[n]$, and $\text{Var}(n) = E[n(n-1)] + E[n] - (E[n])^2$. Theorem 4 is derived in Appendix C.

**Theorem 4.** Let $a = (1 - \phi)/\phi$. The $r$th factorial moments, means, and variances of the univariate PDF for the three models are as follows. For the single-division model

$$E[n(n-1)\cdots(n-t+1)] = n_0(n_0-1)\cdots(n_0+t-1)[F(a,t)/F(ca,t)]$$

$$E(n) = \frac{n_0}{c}$$

$$\text{Var}(n) = n_0^2 \left\{ \frac{1}{c[c-(c-1)\phi]} - \frac{1}{c^2} \right\} + n_0 \left\{ \frac{1}{c} - \frac{1}{c[c-(c-1)\phi]} \right\}.$$ 

For the bisection model

$$E[n(n-1)\cdots(n-t+1)] = n_0(n_0-1)\cdots(n_0+t-1)[F(a,t)/F(2a,t)]$$

$$E(n) = \frac{n_0}{2^3}$$

$$\text{Var}(n) = n_0^2 \left[ \frac{1}{2^3(2-\phi)} - \frac{1}{4^4} \right] + n_0 \left[ \frac{1}{2^3} - \frac{1}{2^4(2-\phi)} \right].$$ 

For the quadrisection model

$$E[n(n-1)\cdots(n-t+1)] = n_0(n_0-1)\cdots(n_0+t-1)[F(a,t)/F(4a,t)]$$

$$E(n) = \frac{n_0}{4^4}$$

$$\text{Var}(n) = n_0^2 \left[ \frac{1}{4^4(4-3\phi)} - \frac{1}{16^4} \right] + n_0 \left[ \frac{1}{4^4} - \frac{1}{4^5(4-3\phi)} \right].$$ 

For each model, the variance is a weighted sum of $n_0^2$ and $n_0$, with weights depending on $\phi$ and $i$. As $\phi$ rises from 0 toward 1, the coefficient of $n_0^2$ rises from 0 to its maximum value, and the coefficient of $n_0$ falls from the same maximum value to 0. Thus, the variance is strongly increasing in the value of $\phi$. This sensitivity is important, for example, in designing a census to estimate the abundance $n_0$ of a species. When variance is low, all cells have roughly the same number of individuals per cell; hence, relatively few cells need to be sampled for a good estimate of the total population. If the variance is high, numerous cells need to be sampled. The theory provides a framework for investigating this trade-off and also trade-offs involving the fineness of scale at which to sample.

**Spatial patterns across cells and an unrealistic feature of the model**

In an empirical grid for a species, cells which are closer together tend to have more similar abundances. Thus, the sample correlogram $r(d)$ typically declines with $d$, where $r(d)$ is the correlation computed over data points $(n_1, n_2)$ for all cell pairs that are a distance $d$ from each other. This pattern holds for most species in the three data sets studied below.

A problem arises in predicting a sample correlogram from a model’s theoretical correlogram. For a given model and parameters $(n_0, i, c, \phi)$, the theoretical correlogram, denote it $\rho(s)$, depends on an “analytical distance” $s$ between cells rather than on the physical distance $d$. To define $s$, consider the Sequence 2 interpretation of the models. Some cell pairs will be on the opposite sides of the first division; some will be on the same side of the first division but opposite sides of the second division; and so on. The distance $s$, which might be called a separation index, is the number of division levels after the two cells are first separated. For the bisection and quadrisection models (with $\phi > 0$), the theoretical correlogram $\rho(s)$ is strictly decreasing in $s$. For the random placement and single-division (negative binomial) models, $\rho(s)$ is constant (slightly negative since $n_0$ is fixed). A closed-form solution for $\rho(s)$ is available, but is too tedious for our purposes here.

The problem is that the physical distance $d$ and the analytical distance $s$, though roughly related across cell pairs, are not the same. The various cell pairs corresponding to a given physical distance $d$ may have different analytical distances $s$, and vice versa, creating ambiguity in predicting $r(d)$ using $\rho(s)$. For some choices of two cell pairs, $A$ and $B$, $d_A < d_B$ even though $s_A > s_B$, implying that the prediction $\rho_A < \rho_B$ contradicts the empirical fact $r_A > r_B$. Fortunately, such departures of theory from realism apply only to some cell pairs and have limited overall consequence. Across all cell pairs, $s$ and $d$ are closely enough correlated that landscapes $N$ generated by the bisection and quadrisection models (with $\phi > 0$) display generally realistic patterns of cross-cell clustering. In particular, the bisection and quadrisection models will, in the empirical comparisons below, greatly improve on the realism of the random placement and negative binomial models, which predict the same (slightly negative) $\rho$ for all cell pairs. Analogous departures from realism apply to other measures of cell-pair association, such as commonness (the probability that both cells of a pair are occupied). Maddux (2004) analyzed such problems in a closely related context, and Ostling et al. (2004) suggested “user rules” for relating theory to data in a way that mitigates the problems.

Each model might be thought of as generating a spatial landscape $N$ in two steps. (1) Divide $n_0$ individuals into $c^i$ groups. (2) Assign the groups to grid cells by some prespecified rule. Our assignment rules are as assumed on Fig. 1. It would be possible to combine a
EMPIRICAL COMPARISONS OF MODELS

To compare the models, we turn to the three data sets referenced in the introduction: serpentine grassland (24 species covering 37,182 individuals), moist tropical forest (305 species covering 235,308 individuals), and dry tropical forest (138 species covering 12,851 individuals). For a given data set, each species is represented by an observed landscape \( N \). For a given model (single-division, bisection, quadrisection) and parameter setting \((i, n_0, c, \phi)\), the multivariate PDF, or likelihood function, of \( N \) can be computed using the theorems. Since the colonization rule for a single species does not involve information about other species, it is implicit that the model applies independently to each species of a community. Therefore, the likelihood function for the entire community is the product of individual species’ likelihoods, and we can easily shift emphasis from individual species to whole communities. Since likelihood functions summarize all model predictions, and since community likelihoods account for every individual of every species, community likelihood comparisons are broad tests of model performance, the subject of the first subsection. The second subsection uses community data to test univariate PDFs. In both subsections, all landscapes are \( 16 \times 16 \), the finest grid size available for the serpentine census. Thus, \((c, i) = (256, 1)\) for the single-division model, \((c, i) = (2, 8)\) for the bisection model, and \((c, i) = (4, 4)\) for the quadrisection model, each model yielding \( c = 256 \) cells. Tests for separate species and other scales are considered in the next main section.

Community likelihood comparisons

Fig. 4 plots community log-likelihoods against the aggregation parameter \( \phi \) for each model and data set. On each graph, \( \phi = 0 \) describes random placement for all three models, hence they have the same vertical intercept, indicated by a large dot. The curves have four important features. First, the curves rise sharply from \( \phi = 0 \), indicating that models, over a range of positive \( \phi \) values, strongly outperform the random placement model. Second, the bisection and quadrisection curves are quite similar, indicating similar empirical performance. Third, these two multisection models achieve much higher likelihoods than the single-division (conditional negative binomial) model. Fourth, the curves have relatively flat plateaus before dropping sharply as \( \phi \) approaches one (the total aggregation extreme), indicating that a range of \( \phi \) values provides similar agreement with the data (roughly \( 0.1 \leq \phi \leq 0.5 \) for the two multisection models).

The superior spatial predictions of bisection and quadrisection over single-division occurs because the single-division model does not predict cross-cell clustering, whereas the multisection models do. This reasoning also explains why the maximum-likelihood estimate of \( \phi \), denoted \( \hat{\phi} \), is smaller for the bisection and quadrisection models than for the single-division model. Aggregation operates on clusters of cells as well as individual cells in the former two models, hence a smaller \( \phi \) will generate the observed degree of clustering. Since quadrisection, relative to bisection, is a step closer to single-division, the quadrisection \( \hat{\phi} \) is slightly higher than the bisection \( \hat{\phi} \) (visually apparent for the tropical data sets and true but not apparent for the serpentine data set).

These comparisons are supported by high statistical significance. For a given model, likelihood differences between models or \( \phi \) values are large. Using the
bisection model and serpentine data set as an example, the ratio of the likelihood at $\phi = 0.25$ to the likelihood at $\phi = 0$ is $1.1 \times 10^{10}$. Further, the sample sizes (numbers of individuals) are so large that even small likelihood differences are significant. Statistical significance is discussed in Appendix D.

**Univariate PDF comparisons**

For a particular data set, model, and species, let $P^{\text{obs}}(n)$ denote the empirical PDF, the fraction of the 256 cells which contain $n$ individuals and, for the remainder of this subsection, let the univariate theoretical PDF $P_i(n \mid n_0, c, \phi)$ be abbreviated $P(n \mid \phi)$. Since $P(n \mid \phi)$ predicts $P^{\text{obs}}(n)$, departure from perfect prediction can be measured as the sum of squared errors between $P^{\text{obs}}(n)$ and $P(n \mid \phi)$ over all values of $n$, namely $n = 0, 1, 2, \ldots, n_0$. Summing over species for the data set gives a community measure of univariate prediction error for the model at given $\phi$ values:

$$\text{SSE}(\phi) = \sum_{\text{species}} \sum_{n=0}^{n_0} [P^{\text{obs}}(n) - P(n \mid \phi)]^2.$$  

For easier interpretation, we normalize SSE(\phi) by defining

$$\text{GF}(\phi) = 1 - \frac{\text{SSE}(\phi)}{\text{SSE}(0)}$$

which measures the goodness of predictive fit of the model relative to random placement prediction. If GF(\phi) is less than, equal to, or greater than 0, the model predicts worse, the same, or better than random placement, respectively, with perfect prediction corresponding to GF(\phi) = 1. Since none of the components of GF(\phi) reflect the locations of more and less abundant cells relative to each other, GF(\phi) incorporates within-cell clustering only.

Fig. 5 plots GF(\phi) against \phi for the three data sets and three models. For each data set, the three curves have almost equal GF values at their peaks, indicating that the models perform equally well in univariate comparisons. In contrast, Fig. 4 indicated inferiority of the single-division model in the more inclusive multivariate (likelihood) comparisons. The contrast confirms the conclusion that the overall superiority of the bisection and quadrisection models results from their prediction of cross-cell as well as within-cell clustering. On Fig. 5, the moist and dry tropical curves peak at about \phi = 0.25, and the serpentine curve at about \phi = 0.50, in rough agreement with maximum-likelihood estimates from Fig. 4. Also of interest, Fig. 5 shows more peaked curves than Fig. 4; for univariate prediction alone, the intervals of well fitting \phi values are narrower.

For a given model and data set, the $\text{Var}[n \mid \phi]$ expressions from Theorem 4 predict the empirical cross-cell variance $\text{Var}^{\text{obs}}[n]$ for each species. An SSE(\phi) over species and a corresponding GF(\phi) can be computed and plotted against \phi, providing a variance analog to Fig. 5. The variance plots (omitted to save space) are quite similar to those of Fig. 5.

**Testing the Colonization Rule**

In this section, we go beyond the comparison of various models and look for violations of underlying assumptions. In particular, we assess whether \phi is invariant, as assumed, across grid scales, species, and individual cell divisions. Although various tests are possible, we consider three which range from less to more detailed. We expect to find violations since a simple model cannot match the complexity of nature. The question will be whether the violations are severe enough to disqualify the model as a reasonable approximation.

**Variation in \phi across scale**

For a given model and data set, the community maximum-likelihood estimate \phi can be computed for various values of the grid index $i$, thus generating data points ($\phi, i$). We illustrate using the bisection model,
where the number of cells, or scale, is $2^i$. The three curves on Fig. 6 plot the data points ($\phi$, $i$) for the three data sets. For the serpentine data set, $i = 8$ (256 cells) is the finest grid available. For the two tropical data sets, grids up to $i = 13$ (8192 cells) are considered. The theory assumes invariance of $\phi$ to $i$, and the three curves of Fig. 6 are indeed relatively flat. All curves lie well above 0, confirming that a positive $\phi$ accounts for the data better than does the random placement value $\phi = 0$, regardless of scale. If a single value of $\phi$ had to be selected for universal use, Fig. 6 suggests roughly $\phi = 0.25$.

**Variation in $\phi$ across species**

Harte et al. (2005) noted that species often appear less aggregated (closer to random placement) at higher abundance, as if the aggregation parameter $\phi$ were typically smaller for more abundant species. Fig. 7 illustrates this pattern using community likelihood plots (left) and univariate goodness of fit plots (right) for two serpentine species, using the bisection model at scale $c' = 256$ cells. Fig. 7a and b are similar in shape and content to Figs. 4 and 5, respectively, but for two individual species rather than whole communities. Fig. 7 shows larger maximum-likelihood and best-univariate-fit values of $\phi$ for *Phlox gracilis* with abundance $n_0 = 617$ ($\phi = 0.45$ and 0.25) than for *Bromus madritensus* with $n_0 = 6990$ ($\phi = 0.25$ and 0.20).

The pattern generalizes to multiple species. For any model and data set, a maximum-likelihood estimate $\hat{\phi}$ can be computed and paired with its abundance $n_0$ to yield a data point $(\hat{\phi}, n_0)$. The points for that model form a scatter which can be inspected for a systematic association between $\phi$ and $n_0$. Fig. 8 shows one scatter for each data set, this time for the quadrisection model, at scale $c' = 256$ cells. Only species with $n_0 \geq 10$ are included.

If there were perfect agreement with the theory and perfect estimation, the points would form a flat scatter. Some variation across species in “true” $\phi$ values would be acceptable in a simple model approximation, and sample variation in the $\hat{\phi}$ estimates would also occur, with greater vertical variation on the left of the scatters since greater sampling variation is expected at smaller $n_0$ values. However, Fig. 8 also shows a systematic negative relation between the estimates of $\phi$ and the abundance $n_0$. The scatters can be summarized by the following regressions of $\hat{\phi}$ on $\ln n_0$ for the serpentine (Eq. 17), moist tropical (Eq. 18), and dry tropical (Eq. 19) data sets, respectively (heteroscedasticity-robust standard errors [Stock and Watson 2003: section 4.9] are shown in parentheses under the coefficients; the sample sizes of the regressions [numbers of species included] are 17, 244, and 94, respectively):

$$\hat{\phi} = 0.4156 - 0.0557 (\ln n_0 - \ln \bar{n}_0) \quad (R^2 = 0.2880)$$

$$\hat{\phi} = 0.3274 - 0.0658 (\ln n_0 - \ln \bar{n}_0) \quad (R^2 = 0.2523)$$

$$\hat{\phi} = 0.3465 - 0.0612 (\ln n_0 - \ln \bar{n}_0) \quad (R^2 = 0.1258).$$

![Fig. 6. Community maximum-likelihood $\phi$ estimates at various values of $i$ for the bisection model. For each $i$, the number of cells is $2^i$. Beyond $i = 8$, to economize on computation, only odd values of $i$ were considered. Odd values of $i$ are emphasized because the long-to-short ratio for the overall moist tropical site is 2 to 1, implying that odd $i$-values yield square cells.

![Fig. 7. Log-likelihood curves (top panel) and goodness-of-fit (GF) curves (bottom panel) for the serpentine species *Phlox gracilis* (abundance $n_0 = 617$) and *Bromus madritensus* (abundance $n_0 = 5989$). For each graph, note that the maximizing $\phi$ values (dots) are smaller for the higher abundance species.](image-url)
Since the explanatory variable ln $n_0$ is expressed as a deviation from its mean (denoted $\overline{\ln n_0}$), the constant term in each regression equals the average $\phi$ across the species of that data set, well above 0 in each case. The regression coefficients are about $-0.06$, implying that doubling $n_0$ would lead to an estimated change in $\phi$ of $-0.06 \ln 2 = -0.042$. This negative association likely reflects density dependence. More abundant species spread more evenly to avoid local crowding, leading to lower estimated $\phi$ values (discussed further in the conclusion). Uriarte et al. (2004) and Wills et al. (2006) present more detailed evidence of density dependence.

**Variation in $\phi$ across abundance, scale, and division size**

As a more detailed test, we can investigate the constancy of $\phi$ across individual Sequence 2 divisions. We illustrate with the bisection model. Recall from Fig. 3 the set $\mathcal{S}(N)$ of all two-way divisions $n \rightarrow (n_1, n_2)$ needed to get from the initial area to the final landscape $N$: $1 + 2 + 2^2 + \ldots + 2^{i-1} = 2^i - 1$ total divisions to get to a final grid index $i$. Here we used $i = 8$, implying $2^i - 1 = 255$ divisions. The likelihood of a single one of these divisions (recall Theorem 2.2) is $F(a, n_1)F(a, n_2)/F(2a, n)$, where $n = n_1 + n_2$. Maximizing this likelihood with respect to $\phi$ for each division yields a separate estimate $\hat{\phi}_i$ specific to the division. Appendix E describes $\phi$ calculation. Associated with each $\hat{\phi}_i$ is the grid index $i$ at which the division takes place, the number of individuals divided $n$, and the total abundance $n_0$ of the species, yielding a data point $(\hat{\phi}_i, i, n, n_0)$. Although there are 255 divisions per species, many divisions do not provide usable $\phi$ values because $n$ is zero or too small. We set the cutoff at $n \geq 10$, yielding 1480, 16266, and 2062 data points for the serpentine, moist tropical, and dry tropical data sets, respectively.

For each data set, the objective is to see whether $\phi$ is influenced by the explanatory variables $(i, n, n_0)$. The following semilog regressions were computed for the serpentine (Eq. 20), moist tropical (Eq. 21), and dry tropical (Eq. 22) data sets (the grid index $i$ is entered in cell-area terms as $A_i = 2^{-i}$, which assumes the total area is normalized to $A_0 = 1$; heteroscedasticity-robust standard errors are in parentheses under coefficients):

$$\hat{\phi} = 0.5281 + 0.0942 \ln A_i - 0.0646 \ln n + 0.0345 \ln n_0$$

$$\left( R^2 = 0.1223 \right).$$

$$\hat{\phi} = 0.3470 - 0.0187 \ln A_i + 0.0123 \ln n - 0.0451 \ln n_0$$

$$\left( R^2 = 0.0469 \right).$$

$$\hat{\phi} = 0.4981 - 0.0261 \ln A_i + 0.0189 \ln n - 0.0826 \ln n_0$$

$$\left( R^2 = 0.0720 \right).$$

Perfect faithfulness to the single-$\phi$ assumption would require zero regression coefficients on all three explanatory variables. In fact, eight of the nine coefficients are statistically significant at a 1% level or better by a two-tailed $t$-test (the exception is $\ln n$ in the dry tropical regression). On the other hand, we do not expect a simple, one-parameter model to be perfectly faithful. To disqualify model predictions, we might expect strong patterns in the regressions. On the contrary, we see small $R^2$’s, and the coefficients in the serpentine regression, which are the largest, have exactly opposite signs to those in the other two regressions.

**Discussion**

Building on Harte et al. (2005), we have described a class of simple, broadly applicable models for analyzing the spatial distribution of a species over a grid of
arbitrary fineness. For each model, individuals are born to locations through a stochastic colonization rule. A single adjustable parameter $\phi$ makes a colonizing individual’s location sensitive to the locations of previously colonized individuals. The colonization rule moves the individual through successively finer subdivisions of the total area. Colonization continues until a given total abundance $n_0$ is reached. The given abundance $n_0$ is to be viewed as the result of ecological forces not explicitly modeled. Structuring colonization on subdivision allows closed form PDF solutions at arbitrary scale (Theorems 1–4) and generates more realistic spatial clustering than in the frequently used random placement and negative binomial models.

Using rectangular grid geometry, different models in the class are indexed by the number $c$ of sub-cells created at each division, by the number $i$ of rounds of subdivisions, and by the spatial arrangement of the subdivisions. The single-division, or conditional negative binomial, model ($c$ arbitrary, $i = 1$) serves as a building block for other models. Though various other models are possible, two were analyzed in detail, the bisection model ($c = 2, i$ arbitrary) and the quadrisection model ($c = 4, i$ arbitrary). The parameter $\phi$ varies from 0, corresponding to random placement, toward an upper bound of 1, corresponding to total aggregation. The midpoint $\phi = 1/2$, the case emphasized in Harte et al. (2005), yields algebraic simplifications and provides reasonably good empirical predictions for many species.

Empirical comparisons of the single-division (conditional negative binomial), bisection, and quadrisection models were based on three large data sets. Community level tests led to several main conclusions. First, random placement ($\phi = 0$) is strongly rejected in favor of the hypothesis that a colonizing individual’s location is sensitive to the locations of previously colonized individuals ($\phi > 0$). Second, the three models (and hence the conditional negative binomial model) perform equally well in predicting observed univariate PDFs. Third, by incorporating cross-cell clustering, the quadrisection and bisection models outperform the single-division model in the more demanding task of predicting multivariate PDFs. Fourth, the bisection and quadrisection models perform equally well overall. As discussed at the end of the theory section, there is a degree of unrealism in the cell correlograms of the bisection and quadrisection models. However, we note that these correlograms are much more realistic than the flat correlograms of the random placement and negative binomial benchmarks.

For the bisection model, the three data sets yield maximum-likelihood values of $\phi = 0.27$ (serpentine), $\phi = 0.20$ (moist tropical), and $\phi = 0.24$ (dry tropical). For the quadrisection model, in comparison, the figures suggest a modestly higher $\phi$, on the order of 0.05 higher. To predict spatial distribution in an area for which detailed data are not available (a primary context for which a simple model might be intended), a consensus setting might be $\phi = 0.25$ for the bisection model and $\phi = 0.3$ for the quadrisection model. The exact value is not critical since the likelihood curves are relatively flat around their maxima.

Given the grid geometry, the central premise of each model is that a single colonization rule with a single $\phi$ applies across all subdivisions for all individuals of all species. Violations of this invariant $\phi$ assumption were investigated empirically. First, a separate community maximum-likelihood estimate $\phi$ was computed for each data set at different values of the grid index $i$. The variation of $\phi$ with $i$ was only modest in degree, especially for the two tropical data sets (Fig. 6). Second, a separate $\phi$ was computed for each species of each data set, revealing a systematic negative association between $\phi$ and $n_0$ (Fig. 8). More abundant species tend to display less aggregation. Third, more detailed violations were considered. For a given model, if a data set is analyzed from the group division viewpoint, a separate $\phi$ can be computed for every species, at every grid index $i$, and for every division (involving enough individuals $n$ to make a sensible estimate). These data allow regression estimation of a function $\phi = \phi(n_0, i, n)$ for all arguments. However, the $R^2$ is small and the coefficient patterns unstable (for example, there is complete reversal of coefficient signs between the serpentine and the two tropical data sets).

Violations are expected. The class of models here builds on strong simplification, whereas nature is complex. Given the sample sizes, the noted violations would easily pass tests of statistical significance. However, statistical significance is not the same as ecological importance. It is a matter of judgment whether the models are well placed on the tradeoff between operational simplicity and predictive accuracy. The most substantial violation was the negative sensitivity of $\phi$ to abundance $n_0$, likely a symptom of density dependence. Fortunately, this violation would be easiest to mitigate; $\phi$ could vary by species with $n_0$ and still be consistent with Theorems 1–4. Allowing $\phi$ to vary by $i$ or $n$, however, would invalidate the theorems, and closed form solutions for PDFs would likely be impossible.

The models can be combined with a birth and death process to form a more general model to which the theorems still apply. For a single species, recall the models’ property that every colonization sequence leading to a given landscape $N$ has the same probability of occurring (Theorems. 1.1, 2.1, and 3.1). Imagine that a landscape $N$ experiences one random death followed by a replacement birth governed by the colonization rule. The death and the birth leave $n_0$ the same. Because every colonization sequence has the same probability of occurring, the effect of the death and birth on the PDF of $N$ is the same as if the most recent colonization was cancelled and replaced by a new one. Although the observed $N$ would typically change, since the new
individual is unlikely to locate in the same cell as the individual that died, the PDF of \( N \), which dictates model predictions, does not change. This argument may be repeated for any number of further deaths and replacement births, leaving \( n_0 \) unchanged. This reasoning explains why the model might be applied either to initial colonization or to mature communities. Further, the reasoning suggests an easy extension of the model to allow stochastic change in \( n_0 \) over time. Let \( P(n_0, t) \) denote the PDF of \( n_0 \) at time \( t \), as determined by some dynamic birth and death process not depending on the spatial distribution of \( n_0 \). We can combine the dynamic \( P(n_0, t) \) with the static \( P(N | n_0) \) from one of the models to form a dynamic spatial model with PDF \( P(N, t) = P(N | n_0)P(n_0, t) \).

For this dynamic generalization, as for the original static model, the parameter \( \phi \) must be assumed constant if the theorems are to hold. Therefore, neither model can accommodate true density dependence, which would require a species’ \( \phi \) value to change over time as abundance changed due to colonization or to continuing births and deaths. However, as suggested just above, density dependence could be approximated by allowing \( \phi \) to vary by species according to \( n_0 \) (in the static model) or according to some long run typical value of \( n_0 \) (in the dynamic model).

Harte et al. (2005) use the bisection model to predict species area relationships and other ecological patterns. We conclude here with three further illustrations of how the bisection, quadrisection, and related models might be used. The first example is estimation of the probability of a species’ extinction under habitat loss. If a species with a minimum viable population of \( n_{mv} \) loses half its area, the probability of extinction (the probability that the remaining half will have abundance less than \( n_{mv} \)) can easily be computed from the multivariate PDF of the bisection or quadrisection model. The probability of extinction increases rapidly with \( \phi \). For example, for the bisection model with \( n_0 = 800 \) and \( n_{mv} = 100 \), the probability of extinction is close to 0 under random placement (\( \phi = 0 \)), but is up to 1/8 under HEAP (\( \phi = 1/2 \)). Greater aggregation raises the probability that a critically large number of individuals will be in the lost area. A next step in extinction analysis would be to add a birth and death process for \( n_0 \).

A second illustrative application of the model, already discussed following Theorem 4, is to census design. How many grid cells need to be sampled to estimate abundance \( n_0 \) to desired accuracy? The answer depends critically on the variance expressions given in Theorem 4. Low variance, resulting from small \( \phi \) values, would require that relatively few cells be sampled. High variance, resulting from large \( \phi \) values, would require that relatively many cells be sampled.

A third illustrative application concerns estimation of \( n_0 \) from presence–absence data sets. The problem has been addressed theoretically and empirically (He and Gaston 2000, 2003, Holt et al. 2002, Harte et al. 2005: Fig. 9). These studies have taken the number of grid cells occupied as the single observed variable from which to estimate \( n_0 \). However, data on spatial clustering of occupied cells provide further information. Since the bisection and quadrisection models incorporate cross-cell clustering, they provide a framework for exploiting the further information.

In summary, a new class of simple spatial models, based on a one-parameter colonization rule, yields closed form expressions for spatial distribution functions. Within the class, the best empirical performers are the bisection and quadrisection models, which account for cross-cell and within-cell aggregation; they outperform the random placement and single-division (conditional negative binomial) models. These multisection models have potential application to census design, extinction risk, and abundance prediction from presence–absence data.

**Acknowledgments**

We thank three anonymous referees, Susan Carey, Adam Smith, and Uthara Sriniwasan for suggesting major improvements to the manuscript. We also thank Jessica Green for permission to use the serpentine data, Richard Condit and the Smithsonian Tropical Research Institute for permission to use the Barro Colorado Island data, and the National Science Foundation for financial support.

**Literature Cited**


**APPENDIX A**

Derivation of Theorem 1 (Ecological Archives M077-008-A1).

**APPENDIX B**

Derivation of Theorem 2 (Ecological Archives M077-008-A2).

**APPENDIX C**

Derivation of Theorem 4 (Ecological Archives M077-008-A3).

**APPENDIX D**

A discussion of the statistical significance of results (Ecological Archives M077-008-A4).

**APPENDIX E**

Division-specific estimate of aggregation parameter (Ecological Archives M077-008-A5).