Linearity and scaling of a statistical model for the species abundance distribution

Hector Garcia Martin and Nigel Goldenfeld
Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, Illinois 61801
(Received 5 July 2001; published 13 February 2002)

We derive a *linear* recursion relation for the species abundance distribution in a statistical model of ecology and demonstrate the existence of a scaling solution.

DOI: 10.1103/PhysRevE.65.032901 PACS number(s): 87.23.Cc, 89.75.Da

I. INTRODUCTION

Understanding the relationship between species richness in a biome and its corresponding area is a long-standing problem in ecology, providing important information about species richness, extinction of species due to habitat loss, and the design of reserves [1]. Among the most usually cited mathematical functions relating the number of different species (S) and the area they occupy (A) is the power law form of the species-area relationship: $S=cA^z$. In a paper by Harte *et al.* [2] this result was shown to be equivalent to assuming self-similarity in the distribution of species. Furthermore, the species abundance distribution $P_0(n)$, the fraction of species with n individuals, was found to satisfy a nonlinear recursion relation.

Banavar *et al.* went on to show that this model exhibits scaling data collapse in the same way as observed in the two-dimensional XY model and in the power fluctuations in a closed turbulent flow [3], a result that follows from hyperscaling [4]. Alternative models for taxon abundance and the species-area law have been proposed, and may be found in [5] and [6]. Our work here is a technical contribution to the specific model proposed by Harte *et al.*, and its conclusions are not germane to these other models.

The purpose of this paper is to show that the nonlinear recursion relation can be recast as a *linear* recursion relation for the species abundance distribution that is much easier to handle; indeed, since the equation governs a probability distribution, it is natural to expect that a linear equation is obeyed. By means of this recursion relation we derive the scaling function *assumed* by Banavar *et al.* [7].

II. THE MODEL AND THE NONLINEAR RECURSION RELATION

In the model proposed by Harte *et al.* [2] an area A_0 with a number of species S_0 is considered. The number of individuals in each species is described by $P_0(n)$, where $S_0P_0(n)$ is the expected number of species with n individuals. The area A_0 is chosen to be in the shape of a rectangle with its length being $\sqrt{2}$ times its width, such that by a bisection along the longer dimension it can be divided into two rectangles of shape similar to the original (see Fig. 1). $A_i=A_0/2^i$ is the area of the rectangle after the ith bisection. If a species is present in an area A_i, and nothing else is known about the species, there are three possibilities: it might be present *only* on the right subpartition of area A_{i-1} [probability $P(R'|L)$], *only* on the left one [$P(R|L')$], or in both [$P(R'|L')$]. By symmetry $P(R'|L)=P(R|L')$; and a is defined as $P(R'|L)\equiv 1-a$. The probability of finding a species on the right side, independently of what happens on the left side, is

$$P(R')=P(R'|L)+P(R'|L')$$
$$=1-a+2a-1=a=P(L') \text{ by symmetry.} \quad (1)$$

Self-similarity is introduced by stating that a is independent of i, that is, scale.

Two conclusions can be derived from this: a species-area relationship of the kind $S=cA^z$ with $a=2^{-z}$ and a recursion relation for $P_i(n)$ (the expected fraction of species with n individuals for an area A_i; see Fig. 1) [2]:

$$P_i(n)=xP_{i+1}(n)+(1-x)\sum_{k=1}^{n-1} P_{i+1}(n-k)P_{i+1}(k), \quad (2)$$

where $x=2(1-a)$. This recursion relation requires an initial condition. It is supplied by defining a minimum patch $A_m=A_0/2^m$, such that it contains on average only one individual

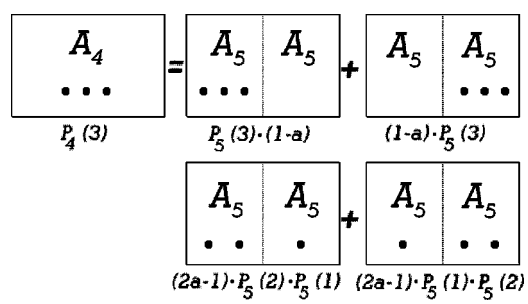

FIG. 1. Explanation of Eq. (2). Let us consider the case $i=4$ and $n=3$. Circles correspond to individuals of a particular species found in a patch. On the left side there are three individuals in the patch A_4; on the right side are all the possible ways in which those three individuals can be distributed in the two patches A_5. The probability of finding three individuals in the patch A_4 is then the addition of the probability that all the individuals are on one side (prob. $1-a$) times the probability that once all the individuals are on one side there are no individuals on one side and there are three individuals on the other side [prob. $1\times P_5(3)$] plus the probability that the species are present on both sides [prob. $2(1-a)$] times the probability that once the species are present on both sides there are two individuals on one side and one individual on the other [prob. $P_5(2)\times P_5(1)$]. Taking $x=2(1-a)$ and $1-x=2a-1$ we find $P_4(3)=xP_5(3)+(1-x)2P_5(2)P_5(1)$. This can be generalized to obtain Eq. (2). Figure taken from [2].

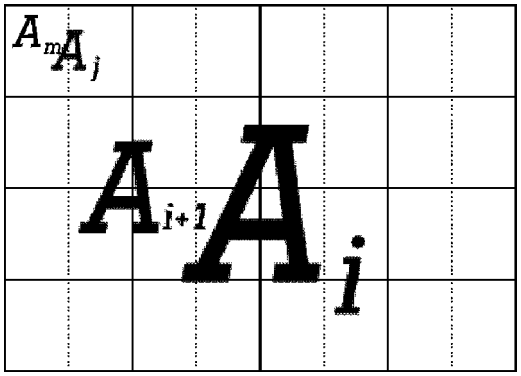

FIG. 2. A_m is the minimum patch. A_j in this case comprises two minimum patches, but it can be of any size. In Eq. (2) the contributions to $P_i(n)$ come from the two patches of size A_{i+1}, whereas in the case of the linear recursion relation they come from the 2^{j-i} patches of size A_j.

(see Fig. 2). Consequently, $P_m(n)=\delta_{n1}$. This also limits the maximum number of individuals that can be found in a patch A_i to 2^{m-i} so $P_i(n)=0$ for $n>2^{m-i}$.

III. THE LINEAR RELATION

Equation (2) is nonlinear, and thus inconvenient to handle efficiently. The purpose of this section is to derive the scaling relation for the probability distribution without making any assumptions about the existence of moments of distribution. As we will see, this can be accomplished by deriving an equivalent linear relation for the probability distribution. This derivation sums up multiple patches at once, rather than proceeding strictly hierarchically as in the original derivation.

We consider that the contributions to $P_i(n)$ come from several (2^{j-i}) patches of area $A_j=A_i/2^j$ ("boxes") instead of from two patches of area $A_{i+1}=A_i/2$ as before (see Fig. 2). The probability of finding n individuals in A_i is then the sum over the probabilities of finding r of these "boxes" with the species present $[R_j^i(r)]$, multiplied by the probability of finding a total of n individuals in these r boxes $[Q_j^i(r,n)]$:

$$P_i(n)=\sum_{r=1}^{2^{j-1}} R_j^i(r)Q_j^i(r,n). \tag{3}$$

Note that the index j is not summed over. It is arbitrary, indicating the size of the "box." For $j=i+1$ there are two boxes of area $A_i/2$ and the original result of Harte *et al.* is recovered, whereas for $j=m-1$ we will find a linear relation. But before establishing these results we explicitly calculate $R_j^i(r)$ and $Q_j^i(r,n)$.

$Q_j^i(r,n)$ is the probability of finding n individuals in r boxes of size $A_i/2^j$ in a total area A_i:

$$Q_j^i(r,n) = \left\{ \begin{array}{ll} \displaystyle\sum_{n_1\ldots n_n=1}^{2^{m-j}} \left(\prod_{l=1}^{r} P_j(n_l)\right)\delta\left(n-\sum_k n_k\right), & r\leqslant 2^{j-i}, \\ 0, \quad r>2^{j-i}. & \end{array} \right\} \tag{4}$$

This formula is the probability of finding n_1 individuals in the first box, n_2 in the second one, etc., while the Kronecker δ function limits the possibilities to those that add up to the total number of individuals n. 2^{j-i} is the maximum number of boxes and 2^{m-j} is the maximum number of individuals in each box.

$R_j^i(r)$ is the probability of finding r boxes of size A_j in which the species is present, in a total area A_i. This is just

$$R_j^i(r)=P_{m+i-j}(r). \tag{5}$$

This follows because the reasoning expressed in Fig. 1 can be applied to find the same recursion relation for $R_j^i(r)$ as for $P_i(n)$:

$$R_j^i(r)=xR_j^{i+1}(r)+(1-x)\sum_{k=1}^{r-1} R_j^{i+1}(k)R_j^{i+1}(r-k). \tag{6}$$

The initial conditions do not change either, with $R_j^i(r)=\delta_{r1}$. The only difference from the derivation for $P_i(n)$ is that r refers to the number of boxes (not individuals) and that the recursion has to be applied $j-i$ times instead of $m-i$ times.

We can now check that for $j=i+1$ we find the same result as before:

$$P_i(n)=\sum_r R_{i+1}^i(r)Q_{i+1}^i(r,n)=R_{i+1}^i(1)Q_{i+1}^i(1,n) + R_{i+1}^i(2)Q_{i+1}^i(2,n). \tag{7}$$

Reading off from Eq. (4):

$$Q_{i+1}^i(2,n)=\sum_{k=1}^{n-1} P_{i+1}(k)P_{i+1}(n-k), \tag{8}$$

$$Q_{i+1}^i(1,n)=P_{i+1}(n), \tag{9}$$

$$R_{i+1}^i(1)=x, \tag{10}$$

$$R_{i+1}^i(2)=1-x. \tag{11}$$

Hence, we obtain

$$P_i(n)=xP_{i+1}(n)+(1-x)\sum_{k=1}^{n-1} P_{i+1}(k)P_{i+1}(n-k) \tag{12}$$

as announced previously. To obtain a linear relation we set $j=m-1$ and obtain

$$Q_{m-1}^i(r,n)=\sum_{n_1,\ldots n_r=1}^{2} \left(\prod_{l=1}^{r} P_{m-1}(n_l)\right)\delta\left(n-\sum_k n_k\right). \tag{13}$$

For $P_{m-1}(n_l)$ we have only the following possibilities:

$$P_{m-1}(n_l)=\begin{cases} x, & n_l=1,\\ 1-x, & n_l=2,\\ 0, & n_l\neq 1,2.\end{cases} \tag{14}$$

We find, denoting by $q\equiv n-r$ the number of boxes with two individuals [factors of $P_{m-1}(2)$ in the equation above]

$$g(n,r)\equiv Q^i_{m-1}(r,n)=\frac{r!}{(r-q)!q!}x^{r-q}(1-x)^q. \tag{15}$$

The first factor is the number of possible configurations in which there are q boxes with two individuals and $n-q$ with one individual. Finally, we obtain

$$P_i(n)=\sum_{r=1}^{2^{m-i-1}} P_{i+1}(r)g(n,r), \tag{16}$$

which is a *linear* relation involving $P_i(n)$ and $P_{i+1}(n)$.

IV. THE SCALING LAW

Equation (13) allows us to derive the scaling law that was assumed by Banavar *et al.* [7]:

$$P_i(n)=\frac{1}{n}f\left(\frac{n}{N_i^{\phi}}\right) \tag{17}$$

where $N_i(=2^{m-i})$ is the maximum number of individuals in an area A_i and $\phi=1-z$.

In order to achieve this, the following steps have to be taken:

First, find the *continuum limit* for $g(r,n)$. Since $g(r,n)$ is just a binomial distribution, it tends to a Gaussian for large n:

$$\begin{aligned} g(n,r)&=\frac{r!}{(r-q)!q!}x^{r-q}(1-x)^q\\ &\approx\frac{1}{\sqrt{2\pi r}}\frac{1}{\sqrt{x(1-x)}}\exp\left(-\frac{1}{2}\frac{(q-r(1-x)^2}{rx(1-x)}\right)\\ &=\frac{1}{\sqrt{\pi}\epsilon_{a,r}}\frac{1}{2a}\exp\left(-\frac{(r-n/2a)^2}{\epsilon_{a,r}^2}\right), \end{aligned} \tag{18}$$

$$\epsilon_{a,r}=\sqrt{4(2a-1)(1-a)r/(2a)^2}. \tag{19}$$

$g(n,r)$ is the probability of finding n individuals in r boxes. This probability is highly peaked around $n=2ar$, since $2a[=1(1-a)+1(1-a)+2(2a-1)]$ is the average of individuals per box. The more boxes there are (bigger r) the sharper the peak. This means that for large r the only relevant values of n are those near $n=2ar$ and the expression given above for $g(n,r)$ is valid for large r (which implies large n).

Second, rewrite everything in terms of a new variable x and a new probability density $\overline{P}_i(x)$. x replaces n and is the fraction of the total number of species: n/N_i (which varies from 0 to 1). $\overline{P}_i(n)$ is the density probability $P_i(x)/(1/2^{m-i})$, where $1/2^{m-i}$ is the distance between two points in the new variable x. In this way all $P_i(n)$ can be compared with each other on equal terms.

In terms of these new variables, the recursion relation can now be written as

$$\overline{P}_i(x)=2\sum_{y=1/2^{m-i-1}}^{1} g(2^{m-i}x,2^{m-i-1}y)\overline{P}_{i+1}(y). \tag{20}$$

The continuum limit is found by taking m (and consequently the number of points $N_{i+1}=2^{m-i-1}$) to an arbitrarily large value and using the continuum limit of $g(r,n)$ as defined above. The fact that the approximation for $g(r,n)$ is not a very good one for small values of n or r is of little importance in the limit of large m:

$$\begin{aligned} \bar{P}_i(x)&=22^{m-i-1}\sum g(2^{m-i}x,2^{m-i-1}y)\bar{P}_{i+1}(y)\underbrace{1/2^{m-i-1}}_{\Delta y}\\ &=\int_0^1 g^*(x,y)\bar{P}_{i+1}(y)dy, \end{aligned} \tag{21}$$

where $g^*(x,y)=22^{m-i-1}g(2^{m-i}x,2^{m-i-1}y)$ and is equal to $(1/a)\delta(y-x/a)$ in the limit of large m:

$$\begin{aligned} g^*(x,y)&=\frac{1}{\sqrt{\pi}}\frac{1}{2a}\frac{1}{\epsilon_{y,a}}\underbrace{\frac{1}{2^{(m-i-1)/2}}}_{\delta}\exp\left[\frac{(y-x/a)^2}{\epsilon_{y,a}^2}\underbrace{2^{m-i-1}}_{\delta^2}\right]\\ &=\frac{1}{\sqrt{\pi}}\frac{1}{2a}\frac{1}{\epsilon_{y,a}\delta}\exp\frac{(y-x/a)^2}{(\epsilon_{y,a}\delta)^2}, \end{aligned} \tag{22}$$

which is a standard representation of the Dirac δ function [8] in the x/a variable for $\epsilon_{y,a}\delta\to\infty$ (or $m\to\infty$):

$$\begin{aligned} &\lim_{m\to\infty}\int g^*(x,y)f(x)dx=f(ay)\\ &\Rightarrow\lim_{m\to\infty} g^*(x,y)=\frac{1}{a}\delta(y-x/a). \end{aligned} \tag{23}$$

This implies that

$$\overline{P}_i(x)=\frac{1}{a}\overline{P}_{i+1}(x/a) \tag{24}$$

which is, in terms of n and $P_i(n)$,

$$P_i(n)=\frac{1}{2a}P_{i+1}(n/2a). \tag{25}$$

Since $a=2^{-z}$ and $\phi=1-z$, multiplying the above equation by n and writing the explicit dependence of $P_i(n)$ on N_i as $P_i(n)=P(n,N_i)$, we obtain

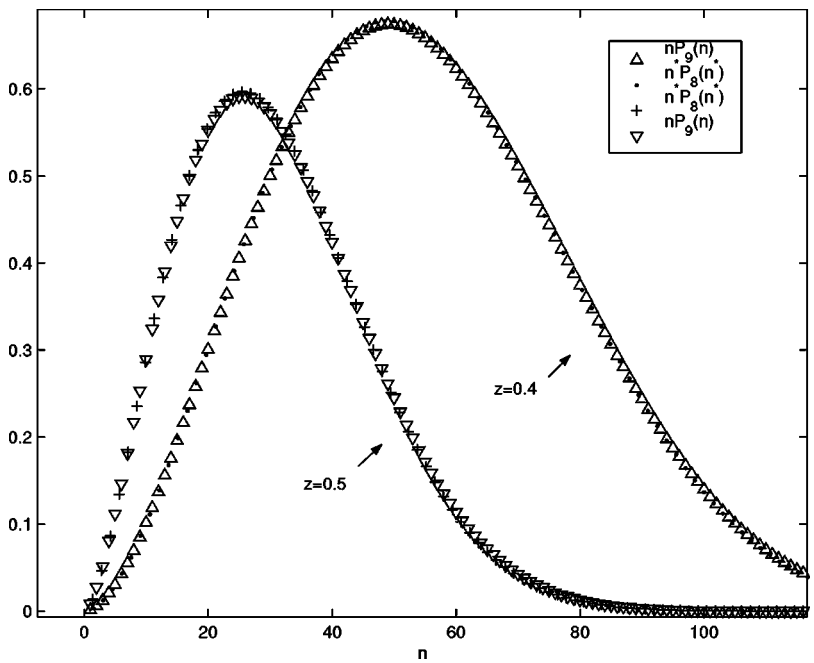

FIG. 3. Scaling function $nP_0(n)=f(n/N_i^{\phi})$ for $m=8$ and $m=9$, and $z=0.4$ and $z=0.5$. $n^* \equiv n/2^{\phi}=n/2a$.

$$nP(n,N_i)=\frac{n}{2a}P(n/2a,N_{i+1})=\frac{n}{2a}P(n/2a,N_i/2)\Rightarrow f(n,N_i)$$

$$\equiv nP(n,N_i)=\frac{n}{2^{\phi}}P(n/2^{\phi},N_i/2), \tag{26}$$

which is by definition $f(n/2^{\phi},N_i/2)$. Since N_i is equal to a power of 2 this means that $nPi(n)$ is a function only of n/N_i^{ϕ}:

$$P_i(n)=\frac{1}{n}f\left(\frac{n}{N_i^{\phi}}\right). \tag{27}$$

As can be appreciated from the results above, a constant a (not dependent on i) is necessary to obtain the scaling law: otherwise ϕ would depend on i. In Fig. 3 we exhibit the scaling function for several z and demonstrate the scaling law.

ACKNOWLEDGMENTS

The authors thank John Harte and Annette Ostling for helpful discussions. This work was supported by the National Science Foundation through Grant No. NSF-DMR-99-70690.

[1] M. Rosenzweig, *Species Diversity in Space and Time* (Cambridge University Press, Cambridge, England, 1995).
[2] J. Harte, A. Kinzig, and J. Green, Science **284**, 334 (1999).
[3] S. Bramwell, P. Holdsworth, and J. Pinton, Nature (London) **396**, 552 (1999).
[4] V. Aji and N. Goldenfeld, Phys. Rev. Lett. **86**, 1007 (2001).
[5] J. Chu and C. Adami, Proc. Natl. Acad. Sci. U.S.A. **96**, 15 017 (1999).
[6] J. D. Pelletier, Phys. Rev. Lett. **82**, 1983 (1999).
[7] J. Banavar, J. Green, J. Harte, and A. Maritan, Phys. Rev. Lett. **83**, 4212 (1999).
[8] C. Cohen-Tannoudji, B. Diu, and F. Laloe, *Quantum Mechanics* (John Wiley & Sons, New York, 1977), Vol. 2, p. 1470.