Article

Forms and genesis of species abundance distributions

Evans O. Ochiaga, **Cang Hui**

Theoretical Ecology Group, Department of Mathematical Sciences, Stellenbosch University, Matieland 7602, and African Institute for Mathematical Sciences, Muizenberg 7945, South Africa

E-mail: evansochiaga@aims.ac.za

Received 25 July 2015; Accepted 3 August 2015; *Published online 1 December 2015*

Abstract

Species abundance distribution (SAD) is one of the most important metrics in community ecology. SAD curves take a hollow or hyperbolic shape in a histogram plot with many rare species and only a few common species. In general, the shape of SAD is largely log-normally distributed, although the mechanism behind this particular SAD shape still remains elusive. Here, we aim to review four major parametric forms of SAD and three contending mechanisms that could potentially explain this highly skewed form of SAD. The parametric forms reviewed here include log series, negative binomial, lognormal and geometric distributions. The mechanisms reviewed here include the maximum entropy theory of ecology, neutral theory and the theory of proportionate effect.

Keywords maximum entropy; lognormal; neutral theory; proportionate effect; biodiversity.

Computational Ecology and Software
ISSN 2220-721X
URL: http://www.iaees.org/publications/journals/ces/online-version.asp
RSS: http://www.iaees.org/publications/journals/ces/rss.xml
E-mail: ces@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Ecology studies the abundance and distributions of living organisms, as well as their relationships and feedbacks with the environment (Odum and Barrett, 1953). Species are not alone in an ecosystem; as such, ecological community is often considered a confined unit of interest, and how community patterns and functions emerge from species interactions has become the overarching question in ecology (Zhang et al., 2011; Hui et al., 2013; Minoarivelo et al., 2014; Nuwagaba et al., 2015). In particular, the number of individuals of species constituting a given ecological community is not evenly distributed, often with many rare species but only a few common ones (Bowler and Kelly, 2010; Hui et al., 2009, 2011), as also observed in fine scale occupancy frequency distributions (Hui and McGeoch, 2007a, 2007b, 2008, 2014). Factors affecting population size of a given species in a community and thus the frequency distribution of species abundance have become a focal point of discussion in ecology (Harte, 2011).

Species abundance distribution (SAD)can be an indicator of commonness and rarity in a community (Harte, 2011); it is a community-level metric, denoted by $\Phi(n)$ hereafter, that represents the number of species with n

individuals, or the probability of detecting an individual being from a given species of *n* individuals in the community (Sognnæs, 2011). SAD is the most straightforward community-level metric that deals with the collective pattern of individuals belonging to different species. In general, species abundance is measured within a confined landscape where multiple species coexist (McGill and Magurran, 2011). It often only depicts species of a similar taxonomic group in a relatively large piece of land with sufficient environmental heterogeneity to facilitate species coexistence (Harte, 2011). Basically, a typical SAD is right skewed and different approaches have been put in place for fitting this SAD. We discuss the parametric forms and mechanistic models for explaining the particular shape of SAD in the following sections.

2 Parametric Forms of SAD

Parametric forms are those approaches aimed at describing the shape of SAD precisely. They are mainly probability distributions that can be interpreted, under some constraints or assumptions, as the probability of finding an individual of a particular species in a random draw (McGill and Magurran, 2011). Ecologists normally build the SAD from data collected through random or systematic samples, by assuming that species are randomly distributed in the community. Under this assumption of random distribution, the number of individuals of species *j* in a sample will follow a Poisson distribution with the intensity parameter λ_i indicating the average number of individuals of species *j* in samples. As such, the probability of species *j* having *r* individuals in a sample can be given as follows (Pielou, 1977):

$$\Pr(r) = e^{-\lambda_j} \lambda_j^r / r! . \tag{1}$$

This is only the probability for species *j*. Species distributions can also follow other non-random patterns that cannot be depicted as a Poisson distribution (Hui, 2009; Hui et al., 2006, 2010), and here we will not venture into such complexity. Now suppose that we are interested in studying the SAD for the entire community, then the probability showed above needs to be modified. The modification of the probability is due to the variation of species densities or abundances in samples. Let us assume that the total number of species in the community is S^T, and the value of λ can be considered a number of S^T samples from a continuous probability distribution of λ that has a probability density distribution of $f(\lambda)$. Now, the probability of finding a species with *r* individuals in a sample is the expectation of the probability of a species from the distribution of $f(\lambda)$ having *r* individuals (Pielou, 1977):

$$\Pr(r) = \int \frac{\lambda^r e^{-\lambda}}{r!} f(\lambda) d\lambda . \tag{2}$$

Now suppose that n_r is the frequency of species with *r* individuals, we thus have $n_r = S^T \cdot \Pr(r)$ (Pielou, 1977). This serves as the basic parametric form of SAD. The first parametric form is Logarithmic distribution which is one of the most widely used approaches put in place for describing species abundance distribution. It is defined as follows (McGill and Magurran, 2011).

$$p(X = x) = \frac{kc^x}{x}, \tag{3}$$

where $k = -1/\log(1-c)$. Logarithmic distribution is derived by assuming that the value of λ for different species sampled from a community follows a Pearson Type III distribution (Gamma distribution); that is, $f(\lambda)$ is defined as follows:

$$f(\lambda) = \frac{p^{-k}\lambda^{k-1}e^{-\frac{\lambda}{p}}}{\Gamma(k)},$$

(4)

where $\lambda \geq 0$ and $k, p > 0$. From equation (4), we can easily compute the probability that a species is represented by r individuals in the collection (Pielou, 1977):

$$pr = \int \lambda^r \frac{e^{-\lambda}}{r!} \frac{p^{-k}\lambda^{k-1}e^{\frac{-\lambda}{p}}}{\Gamma(k)} d\lambda . \quad (5)$$

The solution to equation (5) is nothing but a negative binomial distribution given by:

$$pr = \frac{\Gamma(k+r)}{r!\Gamma(k)}\left(\frac{p}{1+p}\right)^r \left(\frac{1}{1+p}\right)^k, \quad (6)$$

By further simplifying equation (6) and letting $X = p/(1+p)$, we have:

$$pr = \frac{\Gamma(k+r)}{r!\Gamma(k)}(1-X)^k X^r, \quad (7)$$

where $0 < X < 1$. Equation (7) represents the probability of species being represented by r individuals in the sample without ignoring zero class species (missing species in the sample). We now define truncated negative binomial distribution, which is the probability of a given species in the sample when the zero class species is ignored:

$$pr = \frac{pr}{1-p_0} = \frac{\Gamma(k+r)}{r!\Gamma(k)} \frac{X^r(1-X)^k}{\left[1-(1-X)^k\right]}, \quad (8)$$

where $r = 1,2,\ldots$ and p_0 is the probability of zero class species calculated from (7):

$$p_0 = \frac{\Gamma(k)}{0!\Gamma(k)} X^0 (1-X)^k = (1-X)^k .$$

By collecting like terms in equation (8) and substituting them by C, we have:

$$p_r = C\frac{\Gamma(k+r)}{r!}X^r, \quad (9)$$

where $C = (1- X)^k/((1 - (1 - X)^k) \times \Gamma(k))$. In equation (9), k measures variability in the densities of different species. Large and small k values imply small and large variability in species densities, respectively. However, since there are different species in any given ecological communities, this implies that there is a high variation in species abundance. As such, we could let $k \rightarrow 0$ in (9). In doing so, we have:

$$\lim_{k\to 0} pr = \lim_{k\to 0} C\frac{\Gamma(k+r)}{r!}X^r = r\frac{X^r}{r}, \quad (10)$$

where

$$r = \lim_{k\to 0} C\frac{\Gamma(k+r)}{r!} .$$

According to (10) we now have an SAD explained by the logarithmic distribution (Pielou, 1977):

$$\phi(n) = SXpr = S\alpha\frac{X^r}{r} = \alpha\frac{X^r}{r}. \tag{11}$$

Lognormal is another important parametric form of SAD. It has the following probability density function (Aitchison & Brown, 1957);

$$f(\lambda) = \frac{1}{\lambda\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}\ln\left(\frac{\lambda}{\mu}\right)^2\right], \tag{12}$$

where *ln* λ follows normal distribution with mean *ln* μ and variance σ^2. Suppose that the λ value sampled from the community follows a lognormal distribution, then probability of a species being represented by *r* individuals in the sample is given by (Pielou, 1977):

$$pr = \int_0^\infty \lambda^r \frac{e^{-\lambda}}{r!}\frac{1}{\lambda\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}\left(\ln\frac{\lambda}{\mu}\right)^2\right]d\lambda, \tag{13}$$

$$pr = \frac{1}{r!\sigma\sqrt{2\pi}}\int_0^\infty \exp\left[-\lambda + r\ln\lambda - \frac{1}{2\sigma^2}\left(\ln\lambda - \ln\mu\right)^2\right]\frac{d\lambda}{\lambda}, \tag{14}$$

where $\lambda>0$.We could further simplify the Poisson lognormal distribution by letting *ln* $\lambda = x$:

$$pr = \frac{1}{r!\sigma\sqrt{2\pi}}\int_0^\infty \exp\left[e^{-x} + rx - \frac{1}{2\sigma^2}\left(x - lm\mu\right)^2\right]dx \tag{15}$$

Basically, the probability distribution in (15) has no explicit expression for the integral. However, Frank W. Preston was the first to test this distribution against field observations, and proposed a theoretical lognormal frequency distribution to approximate observed frequencies. According to Preston, a given species is represented by its expected number of individuals in the sample, and this number is not affected by the variation during sampling. Preston grouped values of *r* into groups called octaves such that the midpoint of each group is double that of the proceeding one (i.e. r = 1, 2, 4, 8, 16, 32…). The grouping of *r* implies that if a species falls on the boundary of a group, for example 2^x individuals, then it is considered to contribute half a species to the octave ($2^{x-1}$to 2^x) and the other half to the octave (2^x to 2^{x+1}), similar to transforming species abundance when using $\log_2$. When the observation is grouped in this way, the histogram fits well by a symmetrical normal curve. This clearly explains that different values of λ for different species will generate a lognormal distribution of SAD.

Negative binomial distribution is another model that plays a key role in modelling SAD. It is defined as follows (Furbes et al., 1957),

$$p(y) = \frac{\Gamma(y)}{\Gamma(r)(y-r)}p^r q^{y-r}, \tag{16}$$

where $y = r, r+1, r+2, \ldots$ and $0 \le p \le 1$. Unlike the lognormal distribution which assumes that *ln* λ follows a lognormal distribution, a negative binomial distribution assumes that the parameter λ of species sampled follows a Pearson Type III distribution (Pielou, 1977),

$$f(\lambda) = \frac{1}{\Gamma(k)}p^k\lambda^{k-1}e^{\frac{-\lambda}{p}}. \tag{17}$$

Equation (17) implies that the expected frequency of a species having *r* individuals without letting $k \to 0$

follows a negative binomial distribution. Given a negative binomial distribution, the main interest is to investigate whether equation (17) is a monotonically decreasing function at $\lambda = 0$ or whether it has an internal mode at $\lambda > 0$. This is only possible by letting p in (17) to be a constant and then having the derivative of f with respect to λ. In doing so we have (Pielou, 1977),

$$\frac{df(\lambda)}{d\lambda} = \frac{1}{\Gamma(k)} p^{-k} \lambda^{k-2} e^{\frac{-\lambda}{p}} \left\{ k - 1 - \frac{\lambda}{p} \right\}, \tag{18}$$

It is clear that if $k>1$, then $f(\lambda)$ will have its maximum at $\lambda = p(k-1)$; this is a clear indication that if the SAD is fitted using the negative binomial distribution with $k>1$, then some intermediate species will be more frequent than the rare or common species. On the other hand, as $df(\lambda)/d(\lambda)$ is negative for all values λ when $0 \leq k \leq 1$, rare species are more frequent than more abundant species (Pielou, 1977).

Besides the above distributions, SAD can also be modelled by the geometric distribution which is a special case of the negative binomial distribution. First, we have,

$$p(y) = q^{y-1} p\,, \tag{19}$$

where $y = 1, 2, \ldots$ and $0 \leq p \leq 1$. Unlike above cases where the value of λ follows a Pearson Type III distribution and the SAD becomes the negative binomial distribution and log series distribution, respectively, based on the behaviour of k, here we let $k=1$ which gives us the following probability distribution function (Pielou, 1977),

$$f(\lambda) = \frac{1}{\Gamma(1)} p^{-1} \lambda^{0} e^{\frac{-\lambda}{p}}, \tag{20}$$

$$f(\lambda) = e^{-\lambda / p} / p\,, \tag{21}$$

Equation (21) is the probability density function of an exponential distribution. However, by substituting again $k = 1$ in (19) yields the geometric distribution which is the discrete case of an exponential distribution,

$$pr = \left(\frac{p}{1+p} \right)^{r-1} \left(\frac{1}{1+p} \right). \tag{22}$$

SAD can not only be modelled by the above discussed distributions, but also by other parametric models such as the broken-stick and shared sub-niche models. However, these two parametric models are beyond the scope of this article. An example of the fit of these parametric forms to an observed SAD is shown in Fig.1. Clearly, the lognormal distribution is the best in fitting and describing the observed SAD. As all curves decreases with the abundance ranking, it implies that rare species are more frequent than common species. In the next section, we examine a few mechanistic models that can give rise to the skewed SADs.

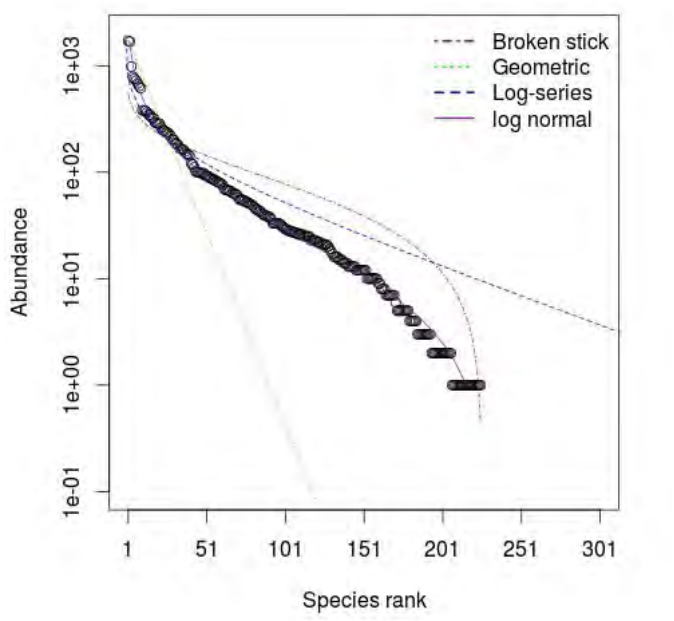

Fig. 1 The fit of four parametric models to the observed abundance ranking of trees on the Barro Colorado Island, data from 1982 (Condit et al., 2000; Zillio and He, 2010).

3 Mechanistic Models of SAD

3.1 Maximum Entropy Theory of Ecology (METE)

Maximum entropy is one of the most fundamental theories in physics and ecology. It is based on maximising the available information entropy of a system under certain constraints. Information entropy is the measure of confusion for a given system. In a thermodynamic system, entropy is described as the number of specific micro-ways in which the system can be arranged for observed macro-states. In ecology, this theory has been used for inferring SADs under constrains of the area occupied by species, the number of species, the number of individuals per species and the total rate of metabolic energy required by all the individuals in the system. This section discusses the application of Lagrange multiplier method in maximizing entropy information for deriving the SAD (Harte, 2011). In a thermodynamic system, the Maximum Entropy theory is defined for some state variables, and the same rationale is also applied in the maximum entropy theory of ecology (METE). State variables are properties of a given system that need to be specified in order to fully describe the system. In the METE we lay our calculations and arguments on the following state variables (Sognnaes, 2011).

- A_0: total area of ecological community.
- S_0: total number of species from any specified taxonomic category contained in areaA_0.
- N_0: total number of individuals of all species in the confined area.
- E_0: total rate of metabolic energy consumed by all the individuals in area A_0.

The above state variables have been chosen for a variety of reasons. For example, A_0has been chosen due to the fact that it is the metric used in measuring physical scale of a given system. Similarly,$S_0$is chosen due to the fundamental role played by species richness in any ecological communities. The total abundance of all species, N_0, and the total metabolic rate of individuals, E_0, are chosen since they set the carrying capacity of the system (Sognnaes, 2011). The METE can not only be used in deriving SADs, but it can also be used in

deriving different ecological metrics such as species area relation, energy distribution among species and the probability density of intra-specific metabolic energy (Harte, 2011).Deriving an SAD using the METE is based on some constraints defined by the state variables A_0, S_0, N_0, E_0 and the ecosystem structure function, $R(n,\varepsilon)$. Here, ecosystem structure function, $R(n,\varepsilon)$, is a joint conditional probability distribution function that is defined over species and individuals of a given species in a confined ecological community with an area A_0, with respect to the metabolic rate of each individual. Metabolic rate is defined as the rate of energy consumption at which the basal metabolism occurs in a living organism (Sognnaes, 2011). In this function n represents species abundance and ε the metabolic energy rate of each individual in a community. In addition, this function is discrete over species abundance n and continuous over metabolic energy rate ε (Harte, 2011). In the METE we set the minimum metabolic rate to 1 ($\varepsilon = 1$) which implies that the metabolic rate of any individuals cannot be less than one. The normalization condition on $R(n,\varepsilon)$ is given by(Harte, 2011),

$$\sum_{n=1}^{N_0}\int_{\varepsilon=1}^{E_0} R(n,\varepsilon \mid A_0,S_0,N_0,E_0)d\varepsilon = 1. \tag{22}$$

State variables further define two new constraints: the average species abundance, N_0/S_0 and the average total metabolic rate of individuals defined over species, E_0/S_0. These two constraints are expressed as follows (Harte, 2011).

$$\sum_{n=1}^{N_0}\int_{\varepsilon=1}^{E_0} n.R(n,\varepsilon \mid A_0,S_0,N_0,E_0)d\varepsilon = \frac{N_0}{S_0}, \tag{23}$$

$$\sum_{n=1}^{N_0}\int_{\varepsilon=1}^{E_0} n.\varepsilon.R(n,\varepsilon \mid A_0,S_0,N_0,E_0)d\varepsilon = \frac{E_0}{S_0}. \tag{24}$$

From (22) to (24) it is clear that A_0 has not played any role in defining new constraints; therefore it does not give much insight in the derivation of an SAD. As such, we ignore its effect and drop it at this stage as we will not use it in the rest of calculation. In addition, from now henceforth, we are going to denote our ecosystem structure function by $R(n,\varepsilon)$ instead of $R(n,\varepsilon|A_0,S_0,N_0,E_0)$. At this stage, we cannot assume anything about the species abundance; that is, we cannot claim that the species of interest have the same abundance. The only information we have is about the state variables. As a result, we are going to use this limited knowledge to help us to infer and describe the shape of an SAD. By applying the maximum entropy principle that maximizes the information entropy of $R(n,\varepsilon)$, we get (Sognnaes, 2011).

$$I_R = -\sum_{n=1}^{N_0}\int_{\varepsilon=1}^{E_0} R(n,\varepsilon)\log(R(n,\varepsilon)) \tag{25}$$

By maximizing (25) subjected to constraints (23) and (24) further taking into account of the normalization condition in (22)have the expression of $R(n,\varepsilon)$, after which an SAD can be easily obtained from $R(n,\varepsilon)$ by integrating the energy requirement of individuals as below(Harte, 2011).

$$\langle \phi(n) \rangle = \int_{\varepsilon=1}^{E_0} R(n,\varepsilon)d\varepsilon \tag{26}$$

In equation (26), it is clear that we need to have the expression of $R(n,\varepsilon)$ in order for the integration to be possible. This can be done by using the Lagrange multiplier function ($L(R,\lambda,\mu)$) given by,

$$L(R,\lambda,\mu)=\left\{\begin{array}{l} I_R-\lambda_1\left(\sum_{n=1}^{N_0}\int_{\varepsilon=1}^{E_0} n.R(n,\varepsilon)d\varepsilon-\frac{N_0}{S_0}\right)-\lambda_2\left(\sum_{n=1}^{N_0}\int_{\varepsilon=1}^{E_0} n.\varepsilon.R(n,\varepsilon)d\varepsilon-\frac{E_0}{S_0}\right)- \\ \mu\left(\sum_{n=1}^{N_0}\int_{\varepsilon=1}^{E_0} R(n,\varepsilon)d\varepsilon-1\right) \end{array}\right\}, \quad (27)$$

where λ_1 and λ_2 are Lagrange multipliers and μ some arbitrary constant (λ_0-1). By substituting for I_R and integrating (26) with respect to $R(n,\varepsilon)$ and setting the derivative to zero, we get,

$$\frac{dL(R,\lambda,\mu)}{dR(n,\varepsilon)}=\left\{\begin{array}{l} \left(-\sum_{n=1}^{N_0}\int_{\varepsilon=1}^{E_0} 1+\log(R(n,\varepsilon)d\varepsilon)\right)-\lambda_1\left(\sum_{n=1}^{N_0}\int_{\varepsilon=1}^{E_0} n.d\varepsilon\right)-\lambda_2\left(\sum_{n=1}^{N_0}\int_{\varepsilon=1}^{E_0} n.\varepsilon.d\varepsilon\right)- \\ (\lambda_0-1)\left(\sum_{n=1}^{N_0}\int_{\varepsilon=1}^{E_0} 1d\varepsilon\right) \end{array}\right\}=0 \quad (28)$$

After some simplifications of equation (28), we get the ecosystem structure function as follows (Harte, 2011),

$$R(n,\varepsilon)=\frac{1}{Z(\lambda_1,\lambda_2)}\exp(-\lambda_1 n-\lambda_2 n\varepsilon), \quad (29)$$

where $Z(\lambda_1,\lambda_2)$ is a partition function given by,

$$Z(\lambda_1,\lambda_2)=\sum_{n=1}^{N_0}\int_{\varepsilon=1}^{E_0}\exp(-\lambda_1 n-\lambda_2 n\varepsilon)d\varepsilon.$$

Accordingly, the SAD can be easily calculated from (26) by integrating out energy required by individuals as shown above (Harte, 2011),

$$\left\langle\Phi(n)\right\rangle=\frac{1}{\log\left(\frac{1}{\beta}\right)}\frac{\exp(-\beta n)}{n}, \quad (30)$$

where β and n are model parameter and abundance of species respectively. Equation (30) is actually following the log series distribution. Having discussed the METE, we move to the Neutral theory in the next section for modelling the SAD.

3.2 Neutral theory of SAD

The neutral theory also plays a fundamental role in generating species abundance distribution in local and meta-communities (He and Hu, 2005).A meta-community is a self-contained evolutionary biogeographic unit, within which most members of a given species originate, live, grow and die, whereas a local community is a unit embedded in the meta-community and is characterised by the exchange of migrants with other local communities (He, 2005). As the name suggests, the concept of neutrality in neutral theory implies that species share the same trophic level in the community are ecologically similar especially with respect to their population demography (i.e. death, birth, speciation and dispersal rates) (Chave, 2004). This theory is primarily based on four major population dynamics: birth, death, immigration and emigration. The stochastic dynamics of population growth can be given by a Markov chain model (He, 2005):

$$\frac{dP_{n,k}(t)}{dt}=P_{n+1,k}(t)d_{n+1,k}+P_{n-1,k}(t)b_{n-1,k}-P_{n,k}(t)\left(b_{n,k}+d_{n,k}\right), \quad (31)$$

where

- $P_{n,k}$ is the probability of species k having n individuals at time t.
- $b_{n,k}$ is the probability for the birth of an individual of species k.
- $d_{n,k}$ is the probability for the death of an individual of species k.

The population growth presented in (31) is subjected to the following boundary condition (McKane and Sole, 2000).

$$P_{-1,k} = d_{0,k} = 0\,, \tag{32}$$

The above boundary condition (32) implies that the probability of death in species k with no individual is zero and this is equal to the probability of species k containing -1 individual. By assuming that the community has a total of S species, we can define the SAD as follows (He, 2005):

$$\left\langle \Phi(n) \right\rangle = \sum_{k=1}^{S} P_{n,k}\,, \tag{33}$$

Equation (33) can be interpreted as the average number of species with n individuals (He, 2005). It only makes sense if the expression for stationary probability distribution $P_{n,k}$ is known. This is only possible by setting $dp_{n,k}(t)/dt$ to zero in (31):

$$p_{n+1,k}(t)d_{n+1,k} + p_{n-1,k}(t)b_{n-1,k} - p_{n,k}(t)b_{n,k} - p_{n,k}d_{n,k} = 0 \tag{34}$$

$$p_{n+1,k}(t)d_{n+1,k} - p_{n,k}(t)b_{n,k} = p_{n,k}(t)d_{n,k} - p_{n-1,k}(t)b_{n-1,k.} \tag{35}$$

Equation (35) is true for all values of n. For simplicity let us define J as follows;

$$p_{n,k}(t)d_{n,k} - p_{n-1,k}(t)b_{n-1,k} = J. \tag{36}$$

By imposing the boundary condition to (36) we have $J = 0$ and thus,

$$p_{n,k}(t)d_{n,k} = p_{n-1,k}(t)b_{n-1,k}\,, \tag{37}$$

$$p_{n,k}(t) = \frac{b_{n-1,k}}{d_{n,k}}\, p_{n-1,k}(t), \tag{38}$$

where $n = 0,1,..,N$ and $d_{n,k} > 0$. Based on Moran's (2008) argument, this then leads to;

$$p_{n,k}(t) = \frac{b_{n-1,k}b_{n-2,k}...b_{0,k}}{d_{n,k}d_{n-1,k}d_{1,k}} P_{0,k}(t), \tag{39}$$

where $n=1,\ldots,N$. Equation (39) can be further simplified to:

$$P_{n,k} = P_{0,k}\prod_{i=0}^{n-1}\frac{b_{i,k}}{d_{i+1,k}}, \tag{40}$$

Accordingly, the SAD is easily obtained from (40) by considering (33) as below,

$$\left\langle \Phi\left(n\right) \right\rangle = \sum_{k=1}^{S} P_{0,k}\prod_{i=0}^{n-1}\frac{b_{i,k}}{d_{i+1,k}}\,, \tag{41}$$

$$\left\langle \Phi(n) \right\rangle = SP_{0,k} \prod_{i=0}^{n-1} \frac{b_{i,k}}{d_{i+1,k}}, \tag{42}$$

In deriving an SAD using the neutral method, several assumptions can be made about death and birth processes in both local and meta communities. Here, we consider birth and death as linear processes. In doing so we have the following (He, 2005);

$$\left\langle \Phi(n) \right\rangle = \frac{Sp_0(v+\lambda)(b+\lambda)(2b+\lambda)...((n-1)b+\lambda)}{(d+\mu)(2d+\mu)...(nd+\mu)}, \tag{43}$$

where λ and μ are immigration and emigration rates respectively. Here, $b_0=(v+\lambda)$ implies that when species become extinct, it can only be replaced through speciation with rate v or via immigration with rate λ. Simplifying equation (43) further leads to the following equation for the SAD:

$$\left\langle \Phi(n) \right\rangle = \theta \frac{\alpha(1+\alpha)(2+\alpha)...((n-1)+\alpha)}{(1+\beta)(2+\beta)...(n+\beta)} (x)^n, \tag{44}$$

$$\left\langle \Phi(n) \right\rangle = \theta \frac{\Gamma(n+\alpha)\Gamma(1+\beta)}{\Gamma(\alpha)\Gamma(n+1+\beta)} x^n, \tag{45}$$

where $\theta=Sp_0(v/\lambda+1)$, $\beta=\mu/d$, $x=b/d$, $\alpha=\lambda/b$ are model parameters and $n=1,2,\ldots$ is species abundance. As the METE and the neutral theory are two major contending theories in ecology, we illustrate their fit to an observed SAD in Fig. 2.

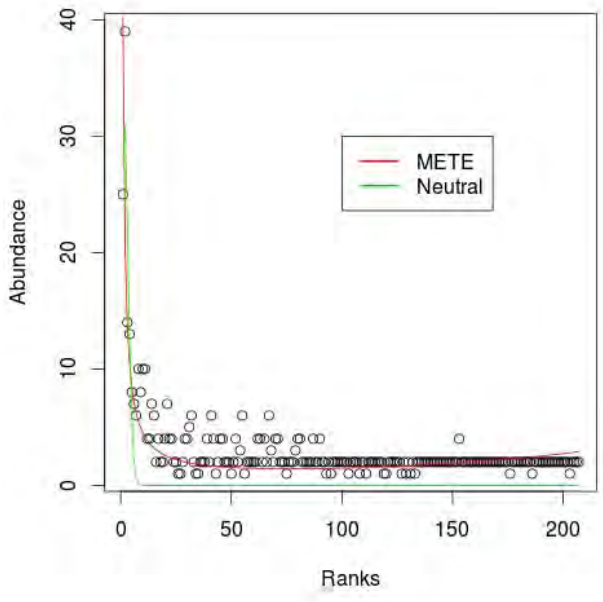

Fig. 2 The fit of two mechanistic models to the observed abundance ranking of trees on the Barro Colorado Island, data from 1982 (Condit et al., 2000; Zillio and He, 2010).

3.3 The theory of proportionate effect

This theory is based on the genesis of lognormal distribution proposed by Brown and Sanders (1981) which was also used for describing the distribution of income in Sydney human population. The theory is based on

partitioning an ecological community into groups of species with respect to different characteristics and then applying the law of proportionate effect in analysing the distribution of species abundance. Considering a given ecological community, since our main aim is to study species abundance in this community, we can partition the community into species groups S_i with N_i abundance based on the distinct characteristics of each species (e.g. based on their functional traits as for classifying morphological species). Then, we determine abundance distribution of each distinct species with respect to the elapse of time at the evolutionary time scale in the given community. This model is based on the following assumptions:

- No predator-prey interactions in the community.
- Abundance of species is jointly contributed by many trivial factors in the community.
- Change in abundance is proportional to the abundance of previous time step, i.e. an exponential growth.

Now suppose that species *i* have X_i abundance such that the abundance is a joint effect of a large number of mutually independent factors that contribute positively to its increase at some point in time. It follows that after some steps in time, changes in abundance of species *i* is proportional to a function $\Phi(X_{j-1})$, which is a function of species abundance in the previous time step. This proportionality is due to the fact that, we assume that species current abundance will depend on the previous abundance. Mathematically, this is to say the following (Aitchson and Brown, 1957):

$$X_j - X_{j-1} = \varepsilon_j \Phi\left(X_{j-1}\right), \tag{46}$$

where ε_j is mutually independent of X_j. The main interest is applying the law of proportionate effect to (46) which states that: *A variable subjected to a process of change is said to obey the law of proportionate effect if the change in the variable at any step of the process is a random proportion of the previous value of the variable.* In doing so, we have equation (46) reduced to (Aitchison and Brown, 1957);

$$X_j - X_{j-1} = \varepsilon_j X_{j-1}, \tag{47}$$

$$X_j = \left(1+\varepsilon_j\right) X_{j-1}, \tag{48}$$

Based on (Crow & Shimizu, 1988), equation (48) can be further simplified to:

$$X_n = X_0 \prod_{j=1}^{n} \left(1+\varepsilon_j\right). \tag{49}$$

By introducing the natural logarithm in (49), we have,

$$\log X_n = \log X_0 + \sum_{j=1}^{n} \log\left(1+\varepsilon_j\right). \tag{50}$$

Equation (50) can be easily approximated using Taylor expansion by assuming that the absolute value of $\varepsilon_j < 1$,

$$\log X_n = \log X_0 + \sum_{j=1}^{n} \varepsilon_j \,, \tag{51}$$

According to the additive form of the central limit theorem, $\log(X_n)$ is asymptotically normally distributed, implying that X_n is asymptotically log normally distributed (Aitchison and Brown, 1957). So far we have proved that the abundance distribution of species *i* is log-normally distributed with respect to time. Now assume that the total abundance of all the species in the ecological community is

$$N = \sum_{i=1}^{j} N_i \,, \tag{52}$$

where $i = 1,2,\ldots,j$ and N_i is abundance of species i. Since we have shown that the abundance of a single species is log-normally distributed, the abundance for a number of S_i species, where $i = 1,2,\ldots,j$ is also log-normally distributed. We now introduce a theorem that was used by Beaulie and Xie (2004) which states that the sum of identical log-normally distributed random variables is still lognormal. Consequently, the sum of abundance of all species shown in (52) is also lognormal, as explained by the theory of proportionate effect.

4 Conclusion

The performance of different parametric forms can be evaluated using the rank abundance curves (Fig. 1). It is clear that the lognormal distribution is the best fit, although each of these parametric models has its own merits as explained in Pielou (1969). In contrast, fitting mechanistic models can give good insights on the potential processes behind observed SADs (Fig. 2). Most ecological literature on SAD focus only on the six models discussed (lognormal, log series, negative binomial, geometric, METE and neutral), which can be compared by using nonlinear regression and related statistics. We also developed the theory of proportionate effect based on particular statistical assumptions which might lack ecological realism. It predicts the lognormal distribution as the null form of SAD, but these assumptions deserve further investigations.

Acknowledgments

This project is supported by the National Research Foundation of South Africa (grant nos. 81825 and 76912).

References

Aitchison J, Brown JA. 1957. The Lognormal Distribution. Cambridge University Press, USA

Beaulie NC, Xie Q. 2004. An optimal lognormal distribution approximation to lognormal sum distribution. IEEE Transaction on Vehicular Technology, 53: 479-489

Bell G. 2000. The distribution of abundance in neutral communities. American Naturalist, 155: 606-617

Bowler GM, Kelly CK. 2010. The general theory of species abundance distribution.ArXiv, 1002.5008

Brown G, Sanders JW. 1981. Lognormal genesis. Journal of Applied probability, 18: 542-547

Chave J. 2004. Neutral theory and community ecology. Ecology Letters, 7: 241-253

Cherukumilli K. 2012. Testing the maximum entropy theory of ecology for predicting species abundance distribution and species area relationship in control and warmed plots at the rocky mountain biological laboratory. University of California at Berkeley, USA

Chisholm RA, Linchstein JW. 2009. Linking dispersal, immigration and scale in the neutral theory of biodiversity. Ecology Letters, 12: 1385-1393

Condit R, et al. 2000. Spatial patterns in the distribution of tropicaltree species. Science, 288: 1414-1418

Crow EL, Shimizu K. 1988. Lognormal Distribution: Theory and Application. Marcel Dekker Inc, USA

Furbes C, Evans M, Hastins N, Peacock B. 1957. Statistical Distribution. John Wiley and Sons Inc, USA

Harte J. 2011. Maximum Entropy and Ecology: A Theory of Abundance, Distribution and Energetics. Oxford University Press, UK

He F. 2005. Deriving a neutral model of species abundance from fundamental mechanism of population dynamics. Functional Ecology, 19: 187-193

He F, Hu XS. 2005. Hubbell's fundamental biodiversity parameter and the Simpson diversity index. Ecology Letters, 8: 386-390

Hubbell SP. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, USA

Hui C. 2009. On the scaling pattern of species spatial distribution and association. Journal of Theoretical Biology, 261: 481-487

Hui C, McGeochMA. 2007a. A self-similarity model for occupancy frequency distribution. Theoretical Population Biology, 71: 61-70

Hui C, McGeoch MA. 2007b. Modeling species distributions by breaking the assumption of self-similarity. Oikos, 116: 2097-2107

Hui C, McGeoch MA. 2008. Does the self-similar species distribution model lead to unrealistic predictions? Ecology, 89: 2946-2952

Hui C, McGeoch, MA. 2014. Zeta diversity as a concept and metric that unifies incidence-based biodiversity patterns. American Naturalist, 184: 684-694

Hui C, McGeoch MA, Warren M. 2006. A spatially explicit approach to estimating species occupancy and spatial correlation. Journal of Animal Ecology, 75: 140-147

Hui C, McGeoch MA, Reyers B, le Roux PC, Greve M, Chown SL. 2009. Extrapolating population size from the occupancy-abundance relationship and the scaling pattern of occupancy. Ecological Applications, 19: 2038-2048

Hui C, Veldtman R, McGeoch MA. 2010. Measures, perceptions and scaling patterns of aggregated species distributions. Ecography, 33: 95-102

Hui C, Foxcroft LC, Richardson DM, MacFadyen S. 2011. Defining optimal sampling effort for large-scale monitoring of invasive alien plants: a Bayesian method for estimating abundance and distribution. Journal of Applied Ecology, 48: 768-776

Hui C, Richardson DM., Pyšek P, Le Roux JJ, Kučera T, Jarošík V. 2013. Increasing functional modularity with residence time in the co-distribution of native and introduced vascular plants. Nature Communications, 4: 2454

McGill BJ, Magurran AE. 2011. Biological Diversity, Frontiers in Measurement and Assessment. Oxford University Press, UK

McKane A, Alonso D, Sole DA. 2000. A mean field stochastic theory for species-rich assembled communities. SFI Working Paper 2000-10-056, Santa Fe Institute, USA

Minoarivelo HO, Hui C, Terblanche JS, Kosakovsky Pond SL, Scheffler K. 2014. Detecting phylogenetic signal in mutualistic interaction networks using a Markov process model. Oikos, 123: 1250-1260

Moran PA. 2008. Random process in genetics. Proceedings of the Cambridge Philosophical Society, 54: 60-71

Nekola JC, Sizling AL, Boyer AG, Storch D. 2008. Artifictions in the log-transformation of species abundance distributions. Folio Geobot, 43: 259-268

Nuwagaba S, Zhang F, Hui C. 2015. A hybrid behavioural rule of adaptation and drift explains the emergent architecture of antagonistic networks. Proceedings of the Royal Society B: Biological Sciences, 282: 20150320

Odum EP, Barrett GW. 1953. Fundamental of Ecology (fifth edition). Thomson Books, Canada

Pielou EC. 1969. An introduction to Mathematical Ecology.Wiley-Interscience, USA

Pielou EP. 1977. Mathematical Ecology. Wiley-Interscience, USA

Sognnaes IA. 2011. Maximum entropy and maximum entropy production in macroecology. Dissertation, Norwegian University of Science and Technology, Norway

Volkov I, Banavar JR, He F, Hubbell SP, Maritan A. 2005. Density dependence explains tree species abundance and diversity in tropical forest. Nature, 438: 658-661

Volkov I, Banavar JR, Hubbell SP, Maritan A. 2003. Neutral theory and relative species abundance in ecology. Nature, 424: 1035-1037

Wackerly DD, Mendenhall W, Scheaffer RL. 2008. Mathematical Statistics with Applications. Thomson Books, Canada

Zhang F, Hui C, Terblanche JS. 2011. An interaction switch predicts the nested architecture of mutualistic networks. Ecology Letters, 14: 797-803

Zillio T, He F. 2010. Modeling spatial aggregation of finite populations. Ecology, 91(12): 3698-3706