Research Article

Abbas Golestani* and Robin Gras

A New Species Abundance Distribution Model Based on Model Combination

Abstract: Species abundance distribution (SAD) is one of the important measures of biodiversity and one of the most significant concepts in ecology communities. Using this concept, the biologists can infer a lot of information from their collected data. In this article, we proposed a new method for predicting SAD. This method is based on the combination of several measures parameterized by machine learning techniques and decomposition of the model in sub-ranges having their proper combination. The goal is to use the combination of several individual models to design a better and more informative model. We show in this article by using many datasets representing different ecological situations that our new method is more robust and outperforms the predictive capacity of the other existing models.

Keywords: model combination, optimization, species abundance distribution, Fisher's logseries

*Corresponding author: **Abbas Golestani**, Department of Computer Science, University of Windsor, 401 Sunset Avenue, Windsor, ON N9B 3P4, Canada, E-mail: golesta@uwindsor.ca

Robin Gras, Department of Computer Science, University of Windsor, Windsor, ON, Canada, E-mail: rgras@uwindsor.ca

1 Introduction

Biologists measure diversity in ecosystems as the variety and abundance of species in a specific location and time [1]. There are many measures of biodiversity, which show the importance of different characteristics of ecological communities [2]. The species abundance distribution (SAD) describes the number of species with a particular abundance in a community; it also indicates how many species are present in a community. Moreover, SAD also has been used for depicting the uncommonness and commonness of species [3]. There are several definitions of SAD which has been proposed by different authors. McGill gave the following explanation about the SAD [4]:

A species abundance distribution is a description of the abundance (number of individuals observed) for each different species encountered within a community. As such, it is one of the most basic descriptions of an ecological community. When plotted as a histogram of number (or percent) of species on the y-axis vs. abundance on an arithmetic x-axis, the classic hyperbolic, "lazy J-curve" or "hollow curve" is produced.

A lot of models have been proposed to predict SADs. McGill provided a helpful survey of different models [4]. Unfortunately, most of the models that have been proposed contain some weakness mentioned by McGill. We have considered them carefully for proposing a new method.

1. For most SAD models, no comparison with other existing models has been proposed. In other words, for the existing models, there is not any comparison of how their predictions fit data to other models.
2. According to McGill, another weakness is that different inconsistent methods have been used to measure goodness of fit. Unfortunately, the different methods used for evaluation, all emphasize different facets of fit. For example, "By-class Good Fit" fits data to the logged-bin and emphasizes fitting rare species. Therefore, the "By-class Good Fit" method and lognormal family methods work on similar features. Thus, any claim of an exceptional fit must be robust by being superior on multiple measures.

3. Even when consistent methods are used, most of the new models will fit some datasets well and other datasets poorly. In other words, for most of the models that have been proposed, the authors have evaluated their method on specific datasets well designed for their methods. So, it could be helpful if we can test our method over several datasets.

In practice, one might come across a case where no single model can achieve an acceptable level of accuracy. In such cases, it would be better to combine the results of different models to improve the overall accuracy. Every model operates well on different aspects of the dataset. For example, the lognormal family methods emphasize fitting rare species more than other methods. As a result, assuming appropriate conditions and combining multiple models may improve prediction performance when compared with any single model [5, 6]. In this article, we propose a new method called Fisher Power Logistic Poisson (FPLP), based on the combination of existing models: Fisher's logseries, power-law, logistic-J and Poisson-lognormal. The main idea comes from a combination of techniques that are used in different fields. According to our knowledge, it is first time that a model based on the combination of other models using genetic algorithm has been proposed for the SAD problem. In this article, we evaluate several SAD models, including the FPLP model, with three different goodness-of-fit measures and applied to eight different datasets. By using the FPLP method, we would be able to investigate how the combination of different models' behavior is important in characterizing different aspects of SAD.

In Section 2, we describe current models that are used in prediction of SAD. In Section 3, we present the different measures of goodness of fit we used. In Section 4, we present the details of the FPLP model. In Section 5, we present the results of new model, and we compare them with results obtained with other models.

2 SAD models

SAD typically represents the way N individuals are partitioned into S species [2]. An example of how SADs can be represented graphically is given in Figure 1.

Besides Figure 1, there are different ways to plot SADs. The complete set of ways to plot SADs have been presented in McGill et al. [4]. The origin of SAD model points to 1932 in which the first model for prediction

Figure 1 There are different representations for SAD. (Left) The histogram is the observed SAD and red curve is the predicted SAD and (Right) a histogram with abundance on a log-scale.

Table 1 Different families of SADs.

Family	SAD
Statistical	Fisher's logseries [7, 8] Lognormal – Preston [9, 10]
Spatial distribution of individuals	Power-law [11, 12] Fractal distribution [13] Multifracta [14]
Population dynamics (metacommunity models)	Logistic-J [15] Neutral model [16–19]
Metacommunity models	Zero-Sum Multinomial (ZSM) [20] Poisson-lognormal [21]
Niche partitioning	Broken stick [22, 23] Sugihara [24]

of SAD was proposed. Since that time, many models have been proposed. We give here several of them among others that are representatives of different modeling families (see Table 1).

As mentioned above, they correspond to various methods based on different concepts which lead to variable results depending on the dataset. The main idea of this article is to use a combination of models belonging to different families of methods for SAD modeling in order to have more flexibility in the final model. In this section, we introduce the four basic models that we used in our model.

2.1 Fisher's logseries

In the 1940s, researchers proposed different statistical models to describe patterns of species abundance [7], which still stimulate a great deal of interest today [25]. Given a sample of a community, Fisher has defined a series expressing the SAD of this sample. Let N and S be, respectively, the numbers of individuals and species in the sample. If n_i is the number of species that contain i individuals in the sample, then

$$\forall i \in N, n_i = \alpha x^i / i \quad [1]$$

The series is, thus, represented by

$$\text{SAD}_{\text{Fisher}} = n_1, n_2, \dots, n_k = \alpha x, \alpha \frac{x^2}{2}, \alpha \frac{x^3}{3}, \dots, \alpha \frac{x^k}{k} \quad [2]$$

where α and x , the two parameters of the model, satisfy the equations:

$$S = \alpha \ln \left(1 + \frac{N}{\alpha} \right) \text{ and } x = \frac{N}{N + \alpha} \quad [3]$$

Therefore, if N and S are known, α and x can be easily calculated. The first parameter, α , is constant for all samples from a given community (it is a characteristic of the community and not of the sample). α is correlated with the total number of species in the considered community and is called the “index of diversity” of the community [26].

2.2 Logistic-J

The logistic-J distribution arises from a dynamical, individual-based model of species [27]. The resulting probability density function (PDF) can be written as follows:

$$f(x) = \begin{cases} c\left(\frac{1}{x} - \delta\right) & \varepsilon \leq x \leq \Delta \\ 0 & \text{Otherwise} \end{cases} \quad [4]$$

where the abundance x runs from ε to a maximum Δ . The constants $\delta = 1/\Delta$ and ε are parameters of the distribution, and c is a constant of integration that gives a value of 1 to the area under the curve of the PDF. The constant c is a function of ε and δ . The parameters ε and Δ are called the inner and outer limits of the distribution, respectively (see Figure 2).

Figure 2 The PDF of the general logistic-J distribution.

The distribution function F of the logistic-J PDF f is obtained by multiplying it by R , the number of species in a sample or in a community ($F(x) = Rf(x)$) leading to the following prediction by logistic-J:

$$\text{SAD}_{\text{Logistic-J}} = F_1, F_2, \dots, F_k \quad [5]$$

2.3 Power-law

One of the best-known patterns in ecology is the power-law form of the species–area relationship. Such a general pattern is important not only for fundamental aspects of ecological theory but also for ecological applications such as the design of reserves and the estimation of species extinction [14]. We consider here a SAD which decays with the power-law from the minimum number of individuals [11], $x = 1$, to the maximum value, $x = X$ as

$$\begin{aligned} f(x) &= (S+1)\alpha x^{-(1+\alpha)} \\ \text{SAD}_{\text{power-law}} &= f_1, f_2, \dots, f_k \end{aligned} \quad [6]$$

The relation between the total number of species S and the maximum number of individuals X is obtained as follows:

$$X^\alpha = S + 1 \quad [7]$$

2.4 Poisson-lognormal

This model mixes the lognormal with the Poisson distribution. One possible way to generalize the univariate Poisson distribution is to use a variable that follows a univariate lognormal distribution. If the abundances, x , are lognormally distributed (which mean that $\log x$ is normally distributed) with mean M and variance V , then the compound Poisson-lognormal distribution is the probability function [21]:

$$P_r = \frac{(2\pi V)^{-1/2}}{r!} \int_0^\infty x^{r-1} e^{-x} e^{-(\log x - M)^2 / 2V} dy, \quad r = 1, 2, \dots \quad [8]$$

$$\text{SAD}_{\text{Poisson-lognormal}} = P_1, P_2, \dots, P_k$$

where r specifies the number of individuals. The distribution can be fitted to observed data by estimating the parameters, M and V , by the method of maximum likelihood.

3 Goodness of fit

The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. In this section, several criteria that we used to compare the observed abundance distribution and the calculated abundance distribution have been introduced.

3.1 Squared prediction error

Squared prediction error (SPE) is a frequently-used measure of the differences between values predicted by a model and the actual observed values [28]. SPE is a good measure of precision. The formula used is as follows:

$$\text{SPE} = \sqrt{\sum [g(x_i) - \hat{g}(x_i)]^2} \quad [9]$$

where $\hat{g}(x_i)$ and $g(x_i)$ are observed values and predicted (or calculated) values, respectively.

3.2 Acceptable fit

Modeled distribution provides an acceptable fit to an observed SAD if the absolute difference between the observed and the calculated values of n_1 is less than 15% of the observed value which means that $|n_1^{\text{obs}} - n_1^{\text{calc}}| \leq 0.15 n_1^{\text{obs}}$ [26]. n_1 is historically meaningful; it has always been an important statistic of a sample, and the SAD models had to give good approximations of n_1 to be validated [7]. Finally, from a practical standpoint, ignoring the role of n_1 can lead to unacceptable results of a goodness-of-fit test. For example, the distributions depicted in Figure 1 (left) do not present an acceptable fit, since the error on n_1 is 68% ($n_1^{\text{obs}} = 38$, $n_1^{\text{calc}} = 64$).

3.3 Basic good fit

Based on the notion of acceptable fit, we can define what we call a “basic good fit”. We say that the model distribution provides a basic good fit to an observed SAD, if it presents an acceptable fit, and if a basic X^2

test (chi-square test), applied on both distributions, gives a X^2 that is not significant (for example, 5% significance) [26, 27].

The X^2 test is then performed, calculating the observed and expected counts and computing the chi-square test statistic.

$$X^2 = \sum_{i=1}^N (O_i - E_i)^2 / E_i \quad [10]$$

where O_i are the observed counts and E_i are the expected counts. The statistic has an approximate chi-square distribution.

To make a direct comparison between test results possible, all the data in Section 5 includes a chi-square test with corresponding scores adjusted to a degrees of freedom equal to $S-1$ (S is number of species or number of classes).

3.4 By-class good fit

Because of the problem of statistical invalidity of the X^2 test when applied on too small values, another solution consists in grouping the terms of a usual SAD into classes, in order to produce a grouped SAD [26].

Analyzing some geometrically varying data is more convenient and meaningful, if they are transferred on a logarithmic scale [29, 30]. The naive approach would be to use the base 2 for the logarithm, but this presents the disadvantage to violate the independence of data points [31]. Traditionally, a base 3 logarithm is used to transform the abundance data. This is done by grouping data into “ $\times 3$ classes” (C_k) $k \geq 0$: class C_k has its center at 3^k , and its edges at $3^k/2$ and $3^{k+1}/2$ [31]. When used with integer values, it gives the following classes: class C_0 contains only 1; class C_1 contains 2–4; class C_2 contains integers from 5 to 13; class C_3 contains integers from 14 to 40; class C_4 contains integers from 41 to 121; and so on. After log-scaling, we can use formulae [9] and [10] to measure difference between observed and predicted values.

4 The FPLP model

In this section, we describe the FPLP model. This model is based on a combination of other models and follows the stacking approach which uses a combination of models to generate performing predictors [5]. By combining models, we expect a more accurate prediction at the expense of an increased complexity of the final model.

Suppose we have a sample dataset Z and a number of different models with a good performance on Z . We can pick a single model as the solution, running onto the risk of making a bad choice for the problem. Instead of picking just one model, a safer option would be to use them all and “average” their outputs. The new model might not be better than the single best model but will diminish or eliminate the risk of picking an inadequate single model [5]. This approach will lead to a more robust predictive method. The proposed new method is based on the combination of four different models:

Fisher’s logseries,
Power-law,
Logistic-J, and
Poisson-lognormal distribution

The main reason of selecting these four models is that they represent different families of methods for SAD modeling and we tried to pick one method from each family. Each family has a specific approach for modeling the SAD, and we wanted to include most of these approaches to have enough flexibility to generate all possible predictions for modeling. We expect to enrich our global model, if we choose our base models from different families. Moreover, as a FPLP model uses a learned weighted combination of base models, it provides information on the relative importance of each base model on each sub-range of SAD.

The FPLP method can be viewed as a post-processing process that combines several fits, using a pre-computed weighted combination, to generate a new fit.

Another important point is that we combine the basic models in three equal sub-ranges of the whole range of values. We found out that the SAD pattern is so complex that it could not be modeled by a single formula. In addition, partial combination gives us more flexibility to use all aspects of combination of models. We noticed that every single base model that we used obtained good prediction levels in specific sub-ranges of SAD. Therefore, we chose to build our model using combinations of the basic models, one for each sub-range we considered. We have chosen three sub-ranges because of a trade-off between two extreme situations: having too many sub-ranges leads to the over-fitting problem and more sub-ranges enhance the flexibility capability of new model and therefore improve the quality of the match. A division in three sub-ranges seems to be a good initial compromise, but more study needed to be done to see what the impact of the number of sub-ranges and their positions is. Moreover, it has been shown in Dornelas and Connolly [3] that the community abundance distribution might have at least three modes and it could be another justification of having three sub-ranges in our combinatorial model. Since there is no limitation on sum of weights (see below) and we have three sub-ranges, it is possible to have multimodal patterns and even, varying the number of sub-ranges, the exact number of modes can be specified. Consequently, to predict a SAD for a specific community, we divide it into three sub-ranges and combine the four basic models in each sub-range independently leading to 12 weights to be learned. The process for making and evaluating the FPLP method is as follows: (1) For each dataset, knowing the population and the number of species, the parameters of each base model (Fishers,...) are computed. In the training part, we just computed the weights of the combination of functions for the FPLP model. (2) Then, we used the computed weights for the rest of datasets to predict the SAD by FPLP model. (3) We compared the accuracy of predicted SADs by base models with the FPLP model.

For the optimization process, we used the genetic algorithm [32] to estimate the weights; however, other optimization method could also be easily applied. Genetic algorithms are combinatorial optimization methods belonging to the class of stochastic search methods [33]. Whereas most stochastic search methods operate on a single solution of the problem at a time, genetic algorithms operate on a population of solutions. They can be viewed as a kind of hill-climbing optimization approach [34] applied simultaneously to a population of solutions. In our problem, a solution is a combination of the values of the weights associated with each base model that minimize the error between SAD's real values and predicted SAD values. For the learning process, as we try to minimize the error, we used the SPE method to evaluate the performance of each weight combination. More precisely, in our case, 12 weights are used: one for each three sub-ranges for each four base models (see Table 2).

The weights can be obtained by fitting the FPLP model to a dataset. In this article, we measured the weights with using two different datasets, and we evaluated the performance of this method to show how robust this method is and no matter what type of dataset is used for estimating the weights of the combination. As the weights only represent the relative importance of each of the four base models for each of the three sub-ranges, we can expect that the weights learned on a particular dataset are still valid in different conditions and can, therefore, be used on different datasets.

Problem of over fitting, which is traditional pitfall of learning methods, is bypassed by setting some stop criterion. For example, during the genetic algorithm process, the learning process was stopped in

Table 2 Interpretation of weights.

	W_1 (Fisher)	W_2 (logistic-I)	W_3 (Power-law)	W_4 (Poisson-log)
Sub-range 1	w_{11}	w_{12}	w_{13}	w_{14}
Sub-range 2	w_{21}	w_{22}	w_{23}	w_{24}
Sub-range 3	w_{31}	w_{32}	w_{33}	w_{34}

generation number 15. In other words, we do not allow the process to go through over and over until it perfectly matches on training dataset with a risk of loss of generality.

In order to show how the FPLP model works based on the combination of other models, we showed the prediction of every model as below in which the indices 1–4 have been chosen for Fisher's logseries, logistic-J, power-law and Poisson-lognormal, respectively.

$$\begin{aligned} \text{SAD}_{\text{Fisher}} &= n_{11}, n_{12}, \dots, n_{1k} \\ \text{SAD}_{\text{Logistic-J}} &= n_{21}, n_{22}, \dots, n_{2k} \\ \text{SAD}_{\text{Power-law}} &= n_{31}, n_{32}, \dots, n_{3k} \\ \text{SAD}_{\text{Poisson-lognormal}} &= n_{41}, n_{42}, \dots, n_{4k} \end{aligned} \quad [11]$$

Based on the characteristics of every dataset, such as number of species and number of individuals, the SAD can be predicted by the base models (see formula [11]). Based on the combination of models for the FPLP model, we have

$$\begin{aligned} \text{SAD}_{\text{FPLP}} &= W_1 \cdot \text{SAD}_{\text{Fisher}} + W_2 \cdot \text{SAD}_{\text{Logistic-J}} + W_3 \cdot \text{SAD}_{\text{Power-law}} + W_4 \cdot \text{SAD}_{\text{Poisson-lognormal}} \\ &= (w_{11} \cdot \text{SAD}_{\text{Fisher}}^{\text{part1}} + w_{12} \cdot \text{SAD}_{\text{Logistic-J}}^{\text{part1}} + w_{13} \cdot \text{SAD}_{\text{Power-law}}^{\text{part1}} + w_{14} \cdot \text{SAD}_{\text{Poisson-lognormal}}^{\text{part1}}) \\ &\quad + (w_{21} \cdot \text{SAD}_{\text{Fisher}}^{\text{part2}} + w_{22} \cdot \text{SAD}_{\text{Logistic-J}}^{\text{part2}} + w_{23} \cdot \text{SAD}_{\text{Power-law}}^{\text{part2}} + w_{24} \cdot \text{SAD}_{\text{Poisson-lognormal}}^{\text{part2}}) \\ &\quad + (w_{31} \cdot \text{SAD}_{\text{Fisher}}^{\text{part3}} + w_{32} \cdot \text{SAD}_{\text{Logistic-J}}^{\text{part3}} + w_{33} \cdot \text{SAD}_{\text{Power-law}}^{\text{part3}} + w_{34} \cdot \text{SAD}_{\text{Poisson-lognormal}}^{\text{part3}}) \\ &= w_{11} \cdot n_{11} + w_{12} \cdot n_{21} + w_{13} \cdot n_{31} + w_{14} \cdot n_{41}, \dots, w_{21} \cdot n_{1\frac{2k}{3}} + w_{22} \cdot n_{2\frac{2k}{3}} + w_{23} \cdot n_{3\frac{2k}{3}} + w_{24} \cdot n_{4\frac{2k}{3}} \\ &\quad , \dots, w_{31} \cdot n_{1k} + w_{32} \cdot n_{2k} + w_{33} \cdot n_{3k} + w_{34} \cdot n_{4k} \end{aligned} \quad [12]$$

The information on the weights for each sub-range and for each model of the combination is given in Table 2.

As we have access to the value of the learned weights, we have some information about what model is important (higher weight) for a particular sub-range. It is very important in terms of interpretation of the results. It gives the possibility to discover some specific properties of particular sub-ranges. For example, seeing that the weight of the Fisher's logseries model is very high for the second sub-range can tell us that the set of species that have an average number of individuals have a distribution which closely follows the Fisher's logseries distribution. It also gives the possibility to compare, for each sub-range, the relative predictive capacity of each model used and, therefore, has a better understanding of its relative importance. As a consequence, our approach could be very helpful for a more precise analysis of the properties of the observed distributions and contributes to build a better ecological theory to explain the distribution patterns observed in a given community. In the results presented in the next section, we always use three combinations of the four basic models for the FPLP model.

5 Results

In this section, we make a comparison between our new model and other models. We compare them according to the several goodness of fit to see how general different methods are in modeling of SAD in different cases. The SADs are different in terms of environment and other ecological factors. This means that different datasets have different characteristics. In order to see how general SAD models are in modeling different datasets, we used a dataset for training the FPLP method (computation of weights) and then used the same weights to get the SAD for other datasets.

From our experiment, it appears that the Fisher's model makes better prediction than logistic-J, power-law and classic Poisson-lognormal, and therefore, we only use Fisher's logseries and two recent proposed methods: ZSM [20] and advanced Poisson-lognormal [35] for comparison purposes. An interesting thing that can also be seen in Figure 3 is that the combination of a Fisher's logseries model with other base models leads to a more accurate global model. In order to give a first visual comparison, the outputs for the four selected base models and FPLP model over Mudumalai [36] dataset are shown in Figure 3.

Figure 3 Prediction of Fisher's logseries model, logistic-J model, power-law model, classical Poisson-lognormal model and FPLP model on Mudumalai dataset.

In the following, we compare more in depth our model with Fisher's model, extended version of Poisson-lognormal model (PN) that was proposed in Izsák [35] and model based on Neutral theory (ZSM) which was proposed in McGill et al. [20]. We test these models over eight different datasets (see Table 3). With these comparisons, we can make a better judgment about the relative efficiency of the models (this is an important feature according to McGill et al. [4]).

Table 3 Characteristics of datasets.

Dataset	Description	N	S	Ref.(s)
Sherman	Trees of the Sherman 6 ha forest plot, Panama	22,000	230	Condit et al. [37]
Dirks	Sample of Lepidoptera, Maine, USA	55,539	349	Dirks [38] and Williams [31]
Fushan	Trees of the Fushan 25 ha forest plot, Taiwan	114,511	110	Su et al. [39]
HKK	Trees of the Huai Kha Khaeng 50 ha forest plot, Thailand	78,444	287	Bunyavejchewin et al. [40]
Bell	Bird community of lowland rainforest in New Guinea	27,112	165	Bell [41]
Thiollay	Birds in French Guiana, 1986	8,507	315	Thiollay [42]
Mudumalai 1988	Trees, Mudumalai 50 ha plot, 1988	25,551	70	Sukumar [36]
Malaysian Butterflies	Malaysian butterflies	9,029	620	Corbet [43]

Notes: S : number of species, N : number of individuals.

It is worth mentioning that the diversity measures can be affected by the sampling process. For example, rare species are less likely to be observed in small samples than in large samples. So, the sample size could be critical in estimation of species richness [3]. For this reason, we consider different datasets with different sizes of sample.

The eight datasets have been used for the evaluation process (see Table 3). For each dataset, there is the number α which reflects diversity of that community sample. In order to ensure a complete evaluation process, we trained the FPLP model on one dataset with low α value, computing the combination of weights, and then we compare the performance of all models on all datasets. We also repeat this experiment using a dataset with high α value for computing the combination of weights, and then, we compare the performance of all models on all datasets.

5.1 Learning with a low α value dataset

In this experiment, the FPLP model has been trained over the Fushan dataset, as it has been explained in Section 4, to compute the weights for combining the basic models in three sub-ranges of the whole range of values. We divided the Fushan dataset into three equal size sub-ranges and then learned the weights (w_1 , w_2 , w_3 , w_4) to combine the four basic models in each sub-range independently. We give here the 12 weights learned:

Sub-range 1:	Sub-range 2:	Sub-range 3:
$W_{\logseries} = 0.464$	$W_{\logseries} = 0.963$	$W_{\logseries} = 0.805$
$W_{\logistic-j} = 0.18$	$W_{\logistic-j} = 0.439$	$W_{\logistic-j} = 0.398$
$W_{power-law} = 0$	$W_{power-law} = 0.185$	$W_{power-law} = 0.116$
$W_{PN} = 0.282$	$W_{PN} = 0.02$	$W_{PN} = 0.356$

From the value of the weights, it can be deduced that the Fisher's logseries model has a much better predictive capacity than the other models. It also seems that the power-law is not a good predictor for these data and that the third sub-range is more complex to describe as it needs a more homogeneous combination of the four models to reach a high predictive level. Values in Table 4 indicate the prediction's error of the

Table 4 Different errors of the four selected models over eight various datasets.

Dataset	Model	Accepted fit (%)	SPE	Basic good fit	By-class (SPE)	By-class (chi-square)
Sherman 1996 $\alpha = 35.3709$	Fisher's logseries	26.12	10.85	19.66	11.27	4.28
	PN	6.08	20.54	127.14	47.60	71.64
	ZSM	47.51	22.57	67.99	45.96	67.30
	FPLP model	14.21	9.44	23.44	14.53	6.97
Driks $\alpha = 49.7198$	Fisher's logseries	30.72	21.32	25	32.99	16.40
	PN	5.46	21.66	60.83	35.82	18.08
	ZSM	80.24	55.35	177.42	107.04	179.25
	FPLP model	8.73	19.07	21.36	23.97	7.82
Fushan $\alpha = 12.0045$	Fisher's logseries	50.04	5.5	4.17	7.03	6.69
	PN	57.89	13.85	108.62	31.74	125.6
	ZSM	135.66	14.35	57.75	28.44	141.86
	FPLP model	15.44	5.16	6.41	11.43	9.28
HKK $\alpha = 37.5398$	Fisher's logseries	63.13	20.29	28.73	24.39	19.06
	PN	43.29	23.63	84.31	45.87	44.86
	ZSM	77.64	36.73	113.97	69.69	115.84
	FPLP model	17.28	12.98	18.55	9.62	2.67
Bell $\alpha = 23.3823$	Fisher's logseries	66.87	12.02	15.06	10.75	7.65
	PN	35.33	17.52	87.74	35.05	55.19
	ZSM	12.65	11.64	39.25	29.82	40.13
	FPLP model	16.71	7.63	13.81	7.95	3.11
Thiollay $\alpha = 64.4038$	Fisher's logseries	68.21	29.41	50.63	34.86	24.98
	PN	4.81	17.96	55.19	24.17	8.46
	ZSM	68.88	50.2	190.17	115.23	189.62
	FPLP model	5.67	16.27	36.14	24.94	8.49
Mudumalai 1988 $\alpha = 8.7754$	Fisher's logseries	46.20	7.85	6.41	10.88	11.1
	PN	33.97	10.35	20.25	14.12	33.15
	ZSM	363.8	23.84	31.7	30.9	110.4
	FPLP model	6.72	7.07	3.75	8.11	7.21
Malaysian butterflies $\alpha = 150.92$	Fisher's logseries	25.79	36.04	39.27	50.61	19.2
	PN	39.66	55.15	68.84	60.18	27.82
	ZSM	94.84	152.32	493.29	263.9	493.15
	FPLP model	24.11	40.67	44.8	50.19	18.41

Note: FPLP models are trained over the Fushan dataset.

different methods for every dataset and every model. The bold numbers are used in Table 4, when FPLP model outperforms all other methods. For all these results, lower the value is better the fit is.

According to Table 4, the FPLP combination model produces more accurate results for most of datasets, even in dataset with high value of α . When the FPLP method is not the best model, the accuracy of FPLP method is still reasonable and close to the best one which shows the robustness of this method. In this evaluation process, there are five goodness-of-fit methods and eight datasets. Therefore, 40 different comparison tests have been performed. FPLP method outperforms Fisher's logseries on 32 of these comparisons, PN on 35 and ZSM on 39. The average percentage of improvement for FPLP method is summarized in Table 5. For example, we computed the average percentage of improvement of FPLP on PN model for each dataset, and then we average the results over all datasets. The percentage values on different measures are not in same scale.

From statistical point of view, we can investigate the result of Table 4 to see how significant is the difference between the FPLP model and the other models. For this reason, we used the t -test [44]. The t -test assesses whether the means of two groups are statistically different from each other. The results of applying t -test to the data of Table 4 are presented in Table 6.

Table 5 The average percentage of improvement of FPLP compared to each measure.

Criterion	FPLP vs Fisher's logseries (%)	FPLP vs PN (%)	FPLP vs ZSM (%)
Accepted fit (%)	352	112	1,109
SPE	28	75	182
Basic good fit	16	456	573
By-class (SPE)	30	157	334
By-class (chi-square)	131	745	2,005

Table 6 The p -value for the distance between error rates of different models for each measure.

Criterion	FPLP vs Fisher's logseries	FPLP vs PN	FPLP vs ZSM
Accepted fit (%)	0.0001	0.001	0.00001
SPE	0.04	0.03	0.0001
Basic good fit	0.06	0.0001	0.00001
By-class (SPE)	0.04	0.0001	0.00001
By-class (chi-square)	0.0001	0.00001	0.00001

Except when we consider the FPLP model with Fisher's logseries model for the "basic good fit", in all other cases the p -value is less than 0.05, which means that the differences between the FPLP model with the other single models are statistically significant.

5.2 Learning with a high α value dataset

In this experiment, the FPLP model has been trained over "Malaysian butterflies" dataset which has high value of α . We divided the Malaysian butterflies dataset into the same three sub-ranges as in Section 5.1 and then estimated the weights (w_1, w_2, w_3, w_4) to combine the four basic models in each sub-range independently. We give here the 12 weights learned:

Sub-range 1:	Sub-range 2:	Sub-range 3:	
$W_{\text{logseries}} = 0.464$		$W_{\text{logseries}} = 0.788$	$W_{\text{logseries}} = 0.805$
$W_{\text{logistic-j}} = 0.788$	$W_{\text{logistic-j}} = 0.429$	$W_{\text{logistic-j}} = 0.996$	
$W_{\text{power-law}} = 0$	$W_{\text{power-law}} = 0.558$	$W_{\text{power-law}} = 0.116$	
$W_{\text{PN}} = 0.337$	$W_{\text{PN}} = 0.02$	$W_{\text{PN}} = 0.325$	

The values obtained for these data are quite different from the previous ones. It seems that for these data the logistic-J is a much better predictor than for the previous dataset. The power-law is also quite important for the prediction of the middle sub-range distribution. The values in Table 7 indicate the prediction's error of the different methods for every dataset and every measurement method. The bold numbers are used in Table 7, when FPLP model outperforms all other methods.

According to Table 7, the FPLP combination method produces more accurate results in most of datasets even in dataset with low value of α . In this evaluation process, there are also five goodness-of-fit methods and eight datasets. Therefore, 40 different comparison tests have been performed. The FPLP method outperforms Fisher's logseries on 31 of these comparisons, PN on 39 and ZSM on 39. The average percentage of improvement for FPLP method is summarized in Table 8.

Table 7 Different errors of the four selected models over eight various datasets.

Dataset	Model	Accepted fit (%)	SPE	Basic good fit	By-class (SPE)	By-class (chi-square)
Sherman 1996 $\alpha = 35.3709$	Fisher's logseries	26.12	10.85	19.66	11.27	4.28
	PN	27.91	23.66	116.97	46.46	66.64
	ZSM	47.51	22.57	67.99	45.96	67.3
	FPLP Model	18.9	10.75	22.44	15.03	7.63
Driks $\alpha = 49.7198$	Fisher's logseries	30.72	21.32	25	32.99	16.4
	PN	43.64	28.52	62.38	40.98	27.56
	ZSM	80.24	55.35	177.42	107.04	179.25
	FPLP model	2.76	16.99	21.72	15.88	3.37
Fushan $\alpha = 12.0045$	Fisher's logseries	50.04	5.55	4.17	7.03	6.69
	PN	115.05	16.84	107.74	32.16	120.2
	ZSM	135.66	14.35	57.75	28.44	141.86
	FPLP model	24.38	6.2	5.59	9.3	8.58
HKK $\alpha = 37.5398$	Fisher's logseries	63.13	20.29	28.73	24.39	19.06
	PN	95.16	33.32	97.84	51.69	66.48
	ZSM	77.64	36.73	113.97	69.69	115.84
	FPLP model	25.39	13.5	26.57	11.45	3.84
Bell $\alpha = 23.3823$	Fisher's logseries	66.87	12.02	15.06	10.75	7.65
	PN	84.33	22.33	98.14	36.82	65.48
	ZSM	12.65	11.64	39.25	29.82	40.13
	FPLP model	24.37	8.81	19.57	11.95	6.89
Thiollay $\alpha = 64.4038$	Fisher's logseries	68.21	29.41	50.63	34.86	24.98
	PN	29.65	22.79	56.13	34.17	18.74
	ZSM	68.88	50.2	190.17	115.23	189.62
	FPLP model	11.06	15.42	35.53	19.94	5.65
Mudumalai 1988 $\alpha = 8.7754$	Fisher's logseries	46.2	7.85	6.41	10.88	11.1
	PN	82.47	12.37	31.33	16.67	48.16
	ZSM	363.8	23.84	31.7	30.9	110.4
	FPLP model	14.3	7.32	3.91	7.87	8.49
Malaysian butterflies $\alpha = 150.92$	Fisher's logseries	25.79	36.04	39.27	50.61	19.2
	PN	17.82	35.36	53.13	53.91	20.67
	ZSM	94.84	152.32	493.29	263.9	493.15
	FPLP model	20.7	35.2	38.93	39.97	11.87

Note: Models trained over Malaysian butterflies dataset.

We also applied the t -test to compare how significance the difference is between the FPLP model and the other models (see Table 9).

Like in our previous experiment, except when we consider the FPLP model with Fisher's logseries model for the "basic good fit", in all other cases the differences between the FPLP model with the other single models are statistically significant.

As it can be seen in Tables 4 and 7, Fisher's logseries generally outperforms the recent methods PN and ZSM. It can be due to the fact that these methods have been developed for very specific cases, and they are not robust enough for general cases. The results presented in Tables 5 and 8 show clearly an important improvement of our new model compared with the three others. The improvement in average quality of prediction is very large compared with the two recent methods. The average improvement of our model is always positive for all measures and against all other tested models. What is also very important to notice is that our approach seems to be quite robust and works well even to make a prediction on distributions that are very different from the ones used to learn the parameters of our model.

Table 8 The average percentage of improvement of FPLP compared to each measure.

Criterion	FPLP vs Fisher's logseries (%)	FPLP vs PN (%)	FPLP vs ZSM (%)
Accepted fit (%)	280	381	861
SPE	25	97	181
Basic good fit	9	487	573
By-class (SPE)	37	173	371
By-class (chi-square)	144	754	2,429

Table 9 The p -value for the distance between error rates of different models for each measure.

Criterion	FPLP vs Fisher's logseries	FPLP vs PN	FPLP vs ZSM
Accepted fit (%)	0.0001	0.0001	0.00001
SPE	0.03	0.02	0.001
Basic good fit	0.08	0.0001	0.0001
By-class (SPE)	0.02	0.001	0.0001
By-class (chi-square)	0.001	0.0001	0.00001

6 Conclusions

In this article, the new SAD model (FPLP model) has been proposed. The FPLP model is based on the combination of several other base models. In response to the criterion defined in the McGill's survey, we have performed a large experimental comparison protocol with our model and the best existing and promising models. We also used eight different datasets with various characteristics corresponding to very different SADs. We also used five different criteria for evaluating the quality of fit of the models. We have shown that our model outperforms the Fisher's logseries model which itself outperforms the two recent models PN and ZSM for all criteria used. The improvements obtained are impressive and statistically significant.

These results show that the approach based on the combination of learned models is very promising and leads to robust and accurate predictors. One important point for these kinds of method is the choice of the base models. The main factor for the efficiency of the resulting global model is the diversity of prediction of the base models. Another important component of our approach is the decomposition of the range of the distribution in three sub-ranges. It seems that this concept is very important, because different sub-ranges of the distribution have different characteristics which can hardly be represented by one unique model.

To be able to conceive ecological theory from an observed SAD, it is very important to clearly understand what distribution this SAD follows. Because we use a weighted combination of models, we know the relative importance of each basic model in each sub-range. In other words, we have a global model which is a combination of four other basic models and, because we have the weights associated with each of them, we can deduce how close to each model, for every sub-ranges, the observed SAD is. For this reason, the combination distribution significantly outperforms other approaches based on a single model as a descriptor of abundances in communities. Obviously, the weights are computed for a given community (due to the training phase), and therefore they are a good instrument to discover specific properties of a given community. However, we have also shown that the predictor we build on a specific community is still a good predictor for a large range of other different communities, outperforming every single model approaches for this task. From our experiment, we have observed that the predictive capacities of the four basic models we used vary a lot depending on the value of the α parameter. For low α values, the Fisher's logseries seems to be a much better descriptor than the other models. But for high α values, the

logistic-J model seems to be more important. We have also observed high variations of the relative importance of these models depending on the sub-ranges considered.

These considerations lead to several interesting future works that we will investigate. First, we will be interested to test other base models that can be used in the combination. It should be interesting to know which models are useful and how many such models should be used. To find optimal combination of base models, there are interesting discussions about oracle inequalities that could be used [45, 46]. A connection between our results and statistical oracle inequalities can be studied more into details. Establishing sharp oracle inequalities in FPLP model and providing theoretical guarantees of optimality could be an interesting future work.

Second, we will investigate the decomposition concept. We will evaluate how many sub-ranges should be considered and what should be the ideal respective size of each sub-range to obtain the most accurate and robust possible predictor. It could be an interesting idea to study the existence of multiple modes in the SAD pattern by testing different sub-ranges. It gives us the ability to identify the exact number of modes (and corresponding ranges) which is needed for different kinds of datasets. It will be also very interesting to study the relative importance of each basic model depending on the sub-ranges and on characteristic of the observed community. Should it be possible to derive some general rules about how good descriptor each basic model is for each sub-range? How the relative importance of the basic model will be useful to characterize a community? This is some of the open questions that we will try to investigate in our future works.

References

1. Magurran AE. Species abundance distributions: pattern or process? *Funct Ecol* 2005;19:177–81.
2. Magurran AE. Measuring biological diversity. Oxford: Blackwell Science, 2004.
3. Dornelas M, Connolly SR. Multiple modes in a coral species abundance distribution. *Ecol Lett* 2008;11:1008–16.
4. McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benetcha HK, et al. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol Lett* 2007;10:995–1015.
5. Kuncheva LI. Combining pattern classifier. New York: Wiley, 2004.
6. Van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. New York: Springer, 2011.
7. Fisher RA, Corbet AS, Williams CB. The relation between the number of species and the number of individuals in a random sample from an animal population. *J Anim Ecol* 1943;12:42–58.
8. Boswell MT, Patil GP. Chance mechanisms generating the logarithmic series distribution used in the analysis of number of species and individuals. In: Statistical ecology, volume i, spatial patterns and statistical distributions, edited by Patil GP, Pielou EC and Waters WE. University Park, PA: Pennsylvania State University Press, 1971:99–130.
9. Preston FW. The commonness and rarity of species. *Ecology* 1948;29:254–83.
10. Hubbell SP. A unified theory of biodiversity and biogeography. Princeton, NJ: Princeton University Press, 2001.
11. Irie H, Tokita K. Species-area relationship for power-law species abundance distribution, 2006. q-bio/0609012.
12. Powell CR, McKane AJ. Predicting the species abundance distribution using a model food web. *J Theor Biol* 2008;255:387–95.
13. Harte J, Kinzig AP, Green J. Self-similarity in the distribution and abundance of species. *Science* 1999;284:334–6.
14. Borda-de-Agua L, Hubbell SP, McAllister M. Species-area curves, diversity indices, and species abundance distributions: a multifractal analysis. *Am Nat* 2002;159:138–55.
15. Dewdney AK. A dynamical model of communities and a new species-abundance distribution. *Biol Bull* 2000;198:152–65.
16. Caswell H. Community structure: a neutral model analysis. *Ecol Monogr* 1976;46:327–54.
17. Hubbell SP. Tree dispersion, abundance and diversity in a tropical dry forest. *Science* 1979;203:1299–309.
18. Bell G. The distribution of abundance in neutral communities. *Am Nat* 2000;155:606–17.
19. Bell G. Neutral macroecology. *Science* 2001;293:2413–18.
20. McGill B, Maurer BA, Weiser MD. Empirical evaluation of neutral theory. *Ecology* 2006;87:1411–23.
21. Bulmer MG. Fitting the poisson lognormal distribution to species-abundance data. *Biometrics* 1974;30:101–10.
22. MacArthur R. On the relative abundance of bird species. *Proc Natl Acad Sci* 1957;43:293–5.
23. MacArthur R. On the relative abundance of species. *Am Nat* 1960;94:25–36.
24. Sugihara G. Minimal community structure: an explanation of species-abundance patterns. *Am Nat* 1980;116:770–87.

25. Chave J. Neutral theory and community ecology. *Ecol Lett* 2004;7:241–53.
26. Devaux D, Gras R. Species abundance patterns in an ecosystem simulation studied through Fisher's logseries. *Simulation Model Pract Theory* 2010;18:100–23.
27. Dewdney AK. The stochastic community and the logistic-J distribution. *Int J Ecol Acta Oecologica* 2003;24:221–9.
28. Potts JM, Elith J. Comparing species abundance models. *Ecol Model* 2006;199:153–63.
29. Kempton RA, Taylor LR. Log-series and log-normal parameters as diversity discriminants for the Lepidoptera. *J Anim Ecol* 1974;43:381–99.
30. Taylor LR, Kempton RA, Woiwod IP. Diversity statistics and the log-series model. *J Anim Ecol* 1976;45:255–72.
31. Williams CB. Patterns in the balance of nature and related problems in quantitative ecology, Vol. 64. London and New York: Academic Press, 1964:1116–26.
32. Periaux J, Winter G. Genetic algorithms in engineering and computer science. Chichester: John Wiley, 1995.
33. Spall JC. Introduction to stochastic search and optimization. New Jersey, NJ: Wiley, 2003.
34. Russell SJ, Norvig P. Artificial intelligence: a modern approach. New Jersey, NJ: Prentice Hall, 2003.
35. Izsák R. Maximum likelihood fitting of the poisson lognormal distribution. *Environ Ecol Stat* 2008;15:143–56.
36. Sukumar R, Suresh HS, Dattaraja HS, John R, Joshi NV. Mudumalai forest dynamics plot, India, 1988. Available at: <http://www.ctfs.si.edu/doc/plots/mudumalai/>
37. Condit RS, Aguilar S, Hernandez A, Perez R, Lao S, Angehr G, et al. Tropical forest dynamics across a rainfall gradient and the impact of an El Nino dry season. *J Trop Ecol* 2004;20:51–72.
38. Dirks CO. Biological studies of Maine moths by light trap methods. Maine Agricultural Experiment Station Bulletin 389, 1937.
39. Su SH, Chang-Yang CH, Lu CL. Fushan 25-ha Plot Data Book, Taiwan Forest Research Institute, Fushan forest dynamics plot, Taiwan, 2007. Available at: <http://www.ctfs.si.edu/doc/plots/fushan/>
40. Bunyavejchewin S, Baker PJ, LaFrankie JV, Ashton PS. Huai Kha Khaeng forest dynamics plot, Thailand, 1992. Available at: <http://www.ctfs.si.edu/doc/plots/hkk/>
41. Bell HL. A bird community of lowland rainforest in New Guinea I: composition and density of the avifauna. *Emu* 1982;82:24–41.
42. Thiollay JM. Structure comparee du peuplement avien dans trois sites de foret primaire en Guyane. *La Terre et la Vie – Revue d'Ecologie* 1986;41:59–105.
43. Corbet AS. The distribution of butterflies in the Malay Peninsula. *Proc Roy Entomol Soc Lond* 1942;A-16:101–16.
44. Kish L. Statistical design for research. New York: John Wiley and Sons, 1987.
45. Kivinen J, Warmuth MK. Averaging expert predictions. *Lecture Notes in Computer Science* 1572, 1999:153–67.
46. Van der Laan MJ, Polley E, Hubbard A. Super learner. *Stat Appl Genet Mol Biol* 2007;6:1–21.