Article

A new model to describe the relationship between species richness and sample size

WenJun Zhang

School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China; International Academy of Ecology and Environmental Sciences, Hong Kong

E-mail: zhwj@mail.sysu.edu.cn, wjzhang@iaees.org

Received 15 December 2016; *Accepted 12 January 2017*; *Published 1 March 2017*

Abstract

In the sampling of species richness, the number of newly found species declines as increase of sample size, and the number of distinct species tends to an upper asymptote as sample size tends to the infinity. This leads to a curve of species richness *vs*. sample size. In present study, I follow my principle proposed earlier (Zhang, 2016), and re-develop the model, $y=K(1-e^{-rx/K})$, for describing the relationship between species richness (y) and sample size (x), where K is the expected total number of distinct species, and r is the maximum variation of species richness per sample size (i.e., *max* dy/dx). Computer software and codes were given.

Keywords sampling; species richness; sample size; curve; model.

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

In the sampling of species richness (Zhang and Schoenly, 1999; Zhang, 2011, 2012a, 2012b), the number of newly emerged species declines as increase of sample size, and the number of distinct species tends to an upper asymptote as sample size tends to the infinity, which results in a curve on the relationship between species richness and sample size. In present study, I follow my principle proposed earlier (Zhang, 2016) to re-develop a model for describing the relationship between species richness and sample size. Computer software and codes are given.

2 Methods

2.1 Model

As a general rule, the number of newly found species declines as increase of sample size, and the number of distinct species (species richness) tends to an upper asymptote as sample size tends to the infinity (∞), as illustrated in Fig.1 (Zhang and Schoenly, 1999). The upper asymptote, K, is the expected total number of

distinct species. Here I follow my principle proposed earlier (Zhang, 2016), and re-develop the model for describing the relationship between species richness and sample size. Firstly, the variation of species richness per sample size is defined as

$$c=\mathrm{d}y/\mathrm{d}x \qquad (1)$$

where y is the number of distinct species (i.e., cumulative number of distinct species), and x is the sample size (i.e., cumulative number of samples). It is obviously that a preceding species in the samples list is more likely a distinct species than its succedent species. Thus c declines (from the maximum variation of species richness per sample size (i.e., max dy/dx), r) to zero as increase of y until the expected total number of distinct species, K, is achieved. As the first-order approximation of equation (1), let

$$c=r(K-y)/K \qquad (2)$$

It leads to the following model

$$\mathrm{d}y/\mathrm{d}x=r(K-y)/K \qquad (3)$$

Solving equation (3), the mathematical model for the relationship between number of distinct species, y, and sample size, x, is obtained

$$y=K(1-\mathrm{e}^{-rx/K}) \qquad (4)$$

According to the model (4), y increases as increase of x, and tends to an asymptote, i.e., expected total number of distinct species, K (Fig. 1).

The expected total number of distinct species, K, and the maximum variation of species richness per sample size, r, can be obtained by using data fitting to the model (4).

Bootstrap procedures are used to produce y-x curve from sampling data of the form, $(d_{ij})_{m*n}$, where m is the number of distinct species found in all samples, n is the total number of samples. y-x curve plots the number of distinct species (y), defined as the number of distinct species found in the previous sample(s), and the sample size (x), defined as the number of samples taken so far. For the first sample, y is defined to equal its number of distinct species (Zhang and Schoenly, 1999). Here the columns of the sample-by-species (species, family, etc.) matrix are bootstrapped. Repeating this process many times (i.e., randomizations), generates a family of curves from which the mean number of distinct species (y) can be calculated for each sample size.

The following are Matlab codes, Bootstrap.m, to produce y-x curve from sampling data of the form, $(d_{ij})_{m*n}$, where m is the total number of distinct species found in all samples, n is the total number of samples

```
samp=input('Input the excel file name of sampling data (e.g., raw.xls. Sampling data matrix is d=(dij)m*n, where m is the
number of distinct species in the network, n is the number of samples): ','s');
sm=input('Input the number of randomizations (e.g., 100, 500, etc.): ');
sampling=xlsread(samp);
m=size(sampling,1); n=size(sampling,2);
x=(1:n)';
for pool=1:n
u=0;
```

```
for sim=1:sm
ma=randperm(n);
s=zeros(m,1);
for i=1:pool
s=s+sampling(:,ma(i));
end
u=u+sum(s~=0);
end
ya(pool)=u/sm;
end
y=ya';
disp('        x        y')
[x y]
```

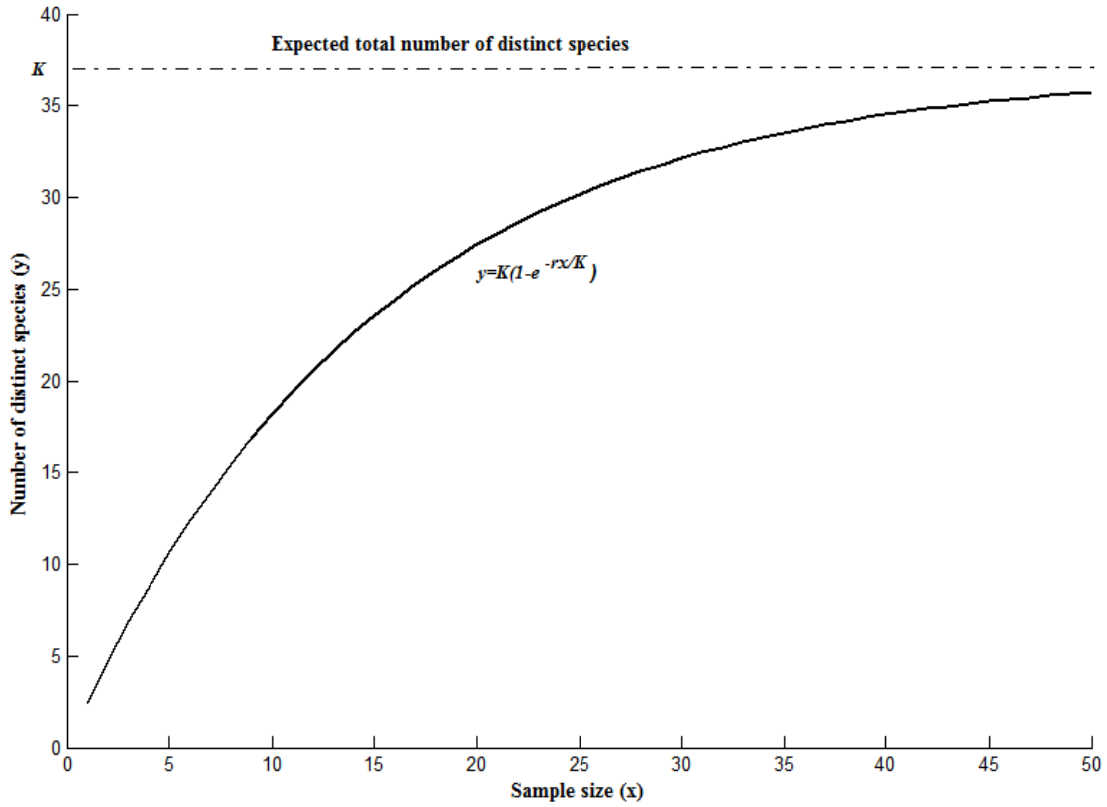

Fig. 1 Relationship between the number of distinct species (*y*; species richness) and sample size (*x*), revised from Zhang (2016b).

The following are Matlab codes, speRichModel.m, to numerically estimate *K* and *r* from *y*-*x* curve

```
xyy=input('Input the excel file name of x-y data (e.g., raw.xls. There are two columns in the file. The first is x (sample size) and the second is y (number of dictinct species).): ','s');
xy=xlsread(xyy);
x=xy(:,1); y=xy(:,2);
k=input('Input the estimated value of parameter K (e.g., 80): ');
r=input('Input the estimated value of parameter r (e.g., 5): ');
```

```
sig=input('Input the significanc level (e.g., 0.01): ');
beta=[k r];
[beta,R,J,SIGMA,MSE]=nlinfit(x,y,@predictfunction,beta);
K=beta(1)
r=beta(2)
deltabeta=nlparci(beta,R,J);
fitted=predictfunction(beta,x);
chi_square=sum((y-fitted).^2./fitted)
p=chi2cdf(chi_square,n-2)
if (p<sig) disp('The data fit model well at the given significanc level.');
else disp('The data is not able to fit model at the given significanc level.');
end
```

The following is the function predictFunction.m

```
function f=predictfunction(beta,x)
f=beta(1)*(1-exp(-beta(2)/beta(1)*x));
```

The example data and software of speRichModel and Bootstrap can be found in supplementary material of the present article.

2.2 Data sources

2.2.1 Dataset I

The data are from our field sampling (1 m^2 of each sampling unit) on arthropods and weeds around Pearl River delta and Zhuhai Campus of SYS University in 2008 (Zhang, 2014; Zhang et al., 2014). Arthropods data for different taxa and areas are represented by dataset names xygz, xyfampea, xyspepea, and weed data for different taxa and areas are represented by dataset names xyweedspepea, xyweedspezhu, xyweedfampea.

2.2.2 Dataset II

The data are from field sampling (0.16 m^2 of each sampling unit) on arthropods in 1996 (Zhang and Schoenly, 1999; Zhang, 2011). Arthropods data for different taxa and seasons are represented by dataset names ir0318family, ir0318species, ir0415family, ir0415species, ir0917family, ir0917species, ir1008family, and ir1008species.

3 Results

Using model (4) and the codes to fit taxa richness *vs*. sample size relationship of datasets I and II, the results are listed in Fig. 1, Fig. 2, Table 1, and Table 2. In most cases the fitting is statistically well. In Dataset I, Pearl River Delta has the greatest the maximum variation of taxa richness per sample size (r) for arthropod families and species compared to its sub-areas and weed taxa. For Dataset II, ir0415 has the greatest maximum variation of taxa richness per sample size (r) for arthropod families and species compared to other seasons.

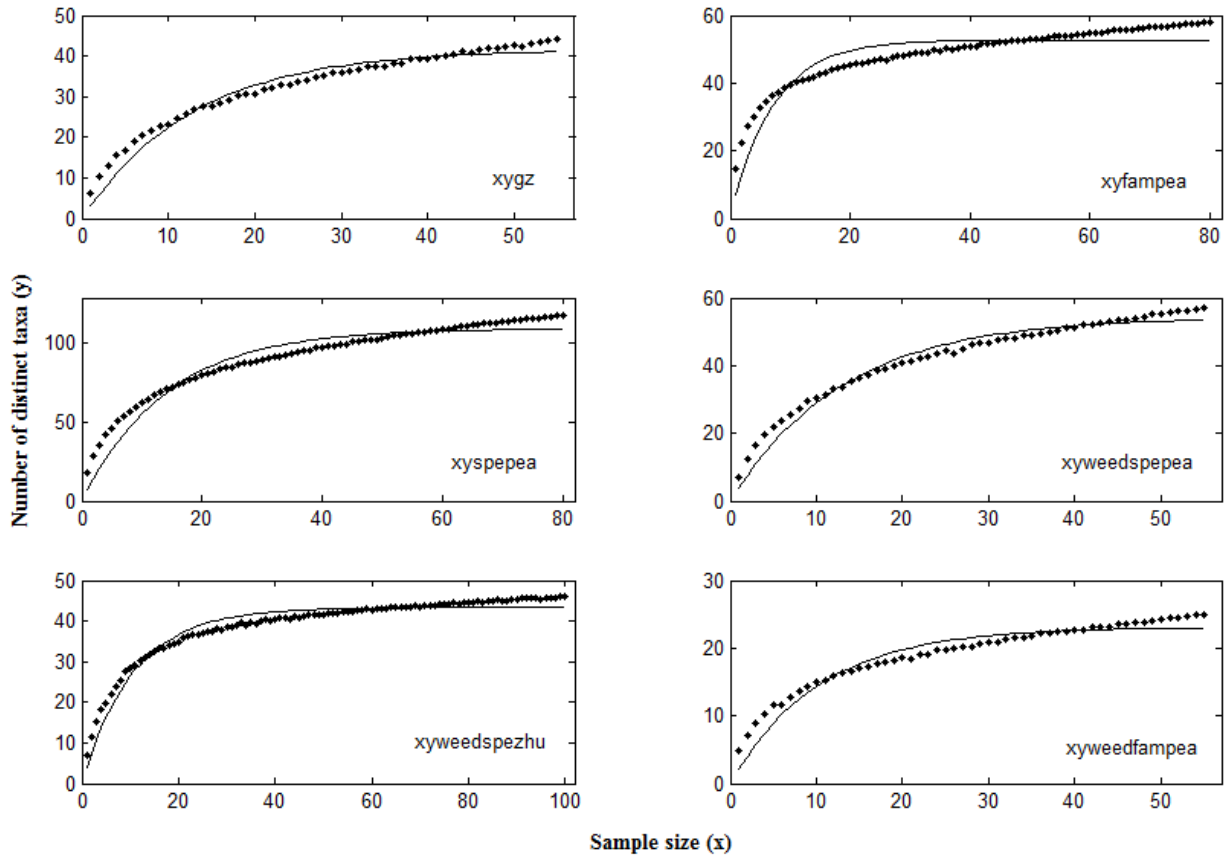

Fig. 1 Fitting results of model (4) to taxa richness *vs*. sample size relationship of Dataset I.

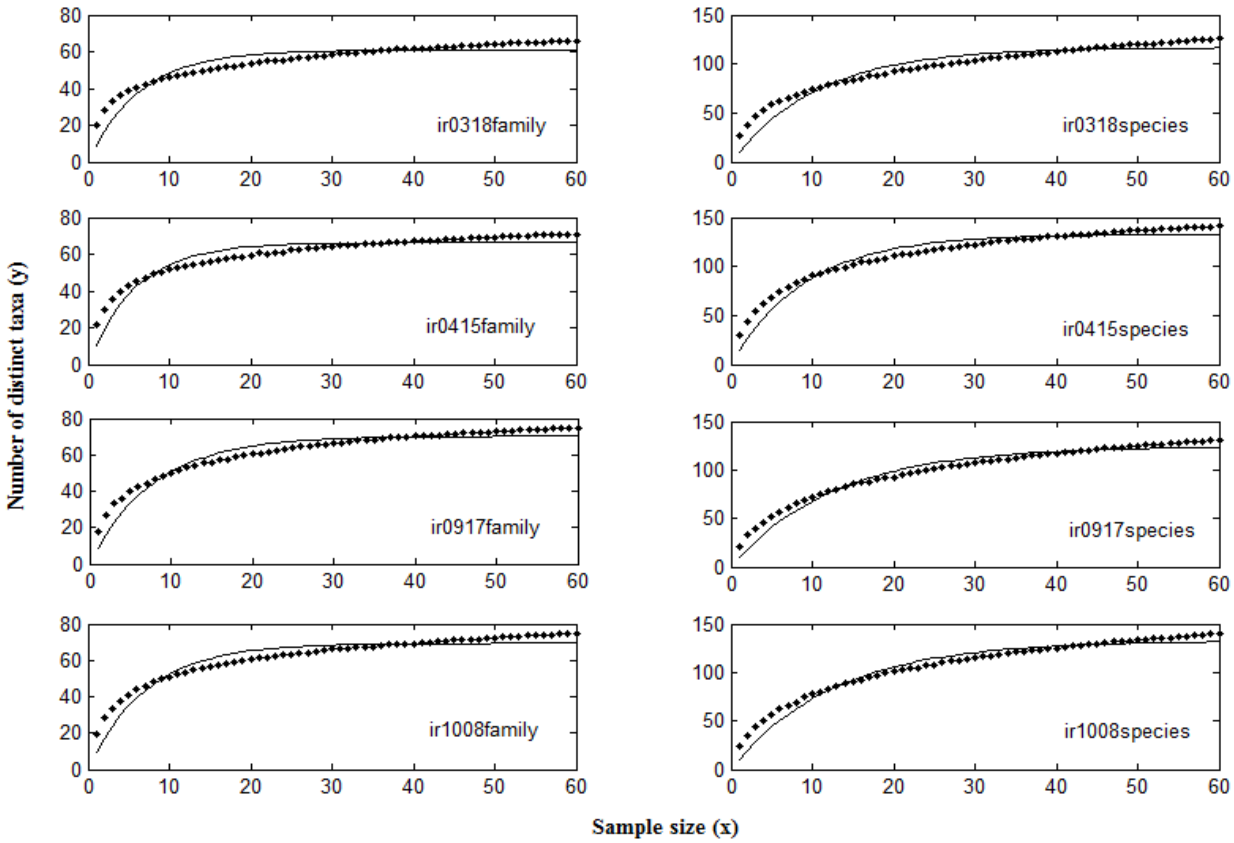

Fig. 2 Fitting results of model (4) to taxa richness *vs*. sample size relationship of Dataset II.

Table 1 Fitting results of model (4) to taxa richness *vs*. sample size relationship Dataset I.

	xygz	xyfampea	xyspepea	xyweedspepea	xyweedspezhu	xyweedfampea
Tot. No. Samples	55	80	80	55	100	55
No. Taxa	44 families	58 families	117 species	57 species	46 species	25 families
K	41.67	52.78	108.69	54.36	43.31	23.11
r	3.22	7.39	7.65	4.13	4.06	2.23
p	$6.3*10^{-8}$	$2.68*10^{-5}$	0.571	$7.96*10^{-8}$	$4.15*10^{-22}$	$1.02*10^{-9}$
Siginificance	<0.0001	<0.0001	-	<0.0001	<0.0001	<0.0001

Table 2 Fitting results of model (4) to taxa richness *vs*. sample size relationship of Dataset II.

	ir0318family	ir0318species	ir0415family	ir0415species
Tot. No. Samples	60	60	60	60
No. Taxa	66 families	126 species	71 families	141 species
K	60.89	116.77	66.22	132.66
r	9.58	10.88	11.37	14.57
p	0.038	0.989	0.003	0.583
Siginificance	<0.05	-	<0.01	-
	ir0917family	ir0917species	ir1008family	ir1008species
Tot. No. Samples	60	60	60	60
No. Taxa	75 families	131 species	75 families	140 species
K	70.22	124.29	69.34	132.70
r	8.99	9.81	9.77	10.57
p	0.004	0.467	0.006	0.684
Siginificance	<0.01	-	<0.01	-

4 Discussion

From Fig. 1 and 2, we may find that there is a systematic deviation between model (4) and observed curves. Generally, the deviation has two phase transitions. The model underestimates species richness in the first phase and proceeding into the second phase, it overestimates species richness. In the third phase, the model underestimates species richness again. How to find a mechanism of the deviation and improve the model is a future consideration.

Acknowledgment

We are thankful to the support of Discovery and Crucial Species Analysis of Important Biological and Social Networks (2015.6-2020.6), from Yangling Institute of Modern Agricultural Standardization, and High-Quality Textbook *Network Biology* Project for Engineering of Teaching Quality and Teaching Reform of Undergraduate Universities of Guangdong Province (2015.6-2018.6), from Department of Education of Guangdong Province, China.

References

Zhang WJ, Schoenly KG. 1999. IRRI Biodiversity Software Series. IV. EXTSPP1 and EXTSPP2: programs for comparing and performance-testing eight extrapolation-based estimators of total taxonomic richness. IRRI Technical Bulletin No.4. International Rice Research Institute, Manila, Philippines

Zhang WJ. 2011. Constructing ecological interaction networks by correlation analysis: hints from community sampling. Network Biology, 1(2): 81-98

Zhang WJ. 2012a. Computational Ecology: Graphs, Networks and Agent-based Modeling. World Scientific, Singapore

Zhang WJ. 2012b. How to construct the statistic network? An association network of herbaceous plants constructed from field sampling. Network Biology, 2(2): 57-68

Zhang WJ. 2014. Interspecific associations and community structure: A local survey and analysis in a grass community. Selforganizology, 1(2): 89-129

Zhang WJ. 2016. A mathematical model for dynamics of occurrence probability of missing links in predicted missing link list. Network Pharmacology, 1(4): 86-94

Zhang WJ, Wang R, Zhang DL, et al. 2014. Interspecific associations of weed species around rice fields in Pearl River Delta, China: A regional survey. Selforganizology, 1(3-4): 143-205