Article

Estimation of node richness by sampling: Application of nonparametric methods

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 9 December 2015; *Accepted 17 January 2016*; *Published 1 March 2017*

Abstract
In the sampling of statistc networks (Zhang, 2011, 2012a, 2012b), the number of new nodes will decline as increase of sample size, and it tends to an upper asymptote as sample size tends to the infinity. However, in most cases our sampling is incomplete. Therefore, the exact number of nodes of a stastic network is unknown. We need to find some methods to estimate node richness in statistic networks. In this study, I use some of the known nonparametric methods to estimate node richness. Computer software and codes were given.

Keywords statistic network; node richness; estimation.

Selforganizology
ISSN 2410-0080
URL: http://www.iaees.org/publications/journals/selforganizology/online-version.asp
RSS: http://www.iaees.org/publications/journals/selforganizology/rss.xml
E-mail: selforganizology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

In the sampling of statistc networks (Zhang, 2011, 2012a, 2012b), the number of new nodes will decline as increase of sample size, and it tends to an upper asymptote as sample size tends to the infinity. For example, the number of new species decreases as increase of sample size in the studies of ecological networks (Zhang and Schoenly, 1999). In most cases, our sampling is incomplete, i.e., the samples we taken are limited. Thus the exact number of nodes of a stastic network is unknown. We need to find some methods to estimate node richness in statistic networks. So far a lot of methods on link prediction have been proposed (Zhang, 2015, 2016). But there are little studies on node richness estimation. In present study, I use some of known nonparametric methods (Burnham and Overton, 1978, 1979; Chao, 1984; Chao and Lee, 1992; Colwell and Coddington, 1994; Zhang and Schoenly, 1999) to estimate node richness. Software and codes are given.

2 Methods

2.1 Nonparametric methods

Six nonparametric estimators are used to estimate node richness of statistic networks (Colwell and Coddington, 1994; Zhang and Schoenly, 1999).

(1) Chao (1984) requires presence-absence data only (Zhang and Schoenly, 1999).

$$S = S_{obs} + L^2/(2M)$$

where L and M are the number of nodes that occur in only one and two samples, respectively.

(2) Jackknife 1 (Burnham and Overton, 1978, 1979) is the first-order jackknife estimator. It can be used to estimate total number of nodes from a sample (Zhang and Schoenly, 1999)

$$S = S_{obs} + L(n\text{-}1)/n$$

where L is the number of nodes found in only one sample, and n is the number of samples.

(3) Jackknife 2 (Burnham and Overton, 1978, 1979) is the second-order jackknife estimator (Zhang and Schoenly, 1999)

$$S = S_{obs} + L(2n\text{-}3)/n\text{-}M(n\text{-}2)^2/[n(n\text{-}1)]$$

where L, M, and n are the same as the above.

(4) Bootstrap (Smith and van Belle, 1984)

$$S = S_{obs} + \sum_{j=1}^{Sobs} (1\text{-}p_j)^n$$

where p_j is the proportion of samples containing node j.

(5) Two methods of of Chao and Lee (1992) are based on sample coverage (Zhang and Schoenly, 1999)

$$S = D/C + n(1 – C)/C*g^2 S = D/C + n\ (1 – C)/C*\beta^2$$

where

$C = 1–f_1/n$

$n=\sum i f_i g^2=\max\{D/C*\sum i(i–1)\ f_i\ /(n(n–1))–1,\ 0\}$

$\beta^2 = \max\{g^2[1+n(1–C)\sum i(i–1)\ f_i\ /...(n(n–1)C)],\ 0\}$

and $D =\sum f_i$ and f_i is the number of classes that have exactly i elements in the sample.

The following are Matlab codes, nodeEst.m, of the nonparametric methods to estimate node richness in a network

```
samp=input('Input the file name of sampling data (e.g., raw.xls, etc. Sampling data matrix is s=(sij)m*n, where m is the number of nodes (species, or objects, etc.) in the network, n is the number of samples): ','s');
ss=xlsread(samp);
m=size(ss,1); n=size(ss,2);
dat=ss';
Sobs=m;
v=sum(dat~=0);
L=sum(v==1); M=sum(v==2);
Chao=Sobs+L^2/(2*M)
Jackknife1=Sobs+L*(n-1)/n
Jackknife2=Sobs+L*(2*n-3)/n-M*(n-2)^2/(n*(n-1))
Bootstrap=Sobs+sum((1-v/n).^n)
for i=1:Sobs
ps(i)=0;
for j=1:n
if (dat(j,i)~=0) ps(i)=ps(i)+1; end
end; end
for i=1:n
w(i)=0;
for j=1:Sobs
```

```
if (ps(j)==i) w(i)=w(i)+1; end
end; end
gam1=0; gam2=0;
d=sum(w);
sp1=0; sp2=0;
for i=1:n
sp1=sp1+i*(i-1)*w(i);
sp2=sp2+i*w(i);
end
cbar=1-w(1)/sp2;
sp3=d/cbar*sp1/(sp2*(sp2-1))-1;
if (sp3>0) gam1=sp3; else gam1=0; end
sp4=gam1*(1+sp2*(1-cbar)*sp1/(sp2*(sp2-1)*cbar));
if (sp4>0) gam2=sp4; else gam2=0; end
if (cbar==0)
ChaoLee1=Sobs
ChaoLee2=Sobs
else
ChaoLee1=d/cbar+sp2*(1-cbar)/cbar*gam1
ChaoLee2=d/cbar+sp2*(1-cbar)/cbar*gam2
end
```

2.2 Data source

The data are from our field sampling (1 m^2 of each sampling site) 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.

3 Results

Table 1 lists estimated taxa richness in six arthropod and weed communities.

Table 1 Estimation of taxa richness for various ecological networks.

	xygz	xyfampea	xyspepea	xyweedspepea	xyweedspezhu	xyweedfampea
Chao	72	143	149	80	-	41
Jackknife 1	59	71	148	75	53	33
Jackknife 2	69	83	163	85	59	39
Bootstrap	50	63	131	65	49	28
Chao & Lee 1	74	71	176	92	53	42
Chao & Lee 2	101	77	218	113	55	56
Mean	71	85	164	85	54	40

Acknowledgment

We are thankful to the support of Discovery and Crucial Node 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

Burnham KP, Overton WS. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika, 65: 623-633

Burnham KP, OvertonWS. 1979. Robust estimation of population when capture probabilities vary among animals. Ecology, 60: 927-936

Chao A. 1984. Non-parametric estimation of the number of classes in a population. Scandinavian Journal of Statistics, 11: 265-270

Chao A, Lee SM. 1992. Estimating the number of classes via sample coverage. Journal of American Statistician Association, 87: 210-217

Cowell RK. 1992. Human aspects of biodiversity: an evolutionary perspective. In: Biological Diversity and Global Change, International Union of Biological Sciences, Monograph No. 8 (Solbrig OT, van Emden HM, van Oordt PGWJ, eds). 209-222, IUBS Press, Paris, France

Colwell RK, Coddington JA. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London B, 345: 101-108

Zhang WJ, Schoenly KG. 1999a. 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. 2015. Prediction of missing connections in the network: A node-similarity based algorithm. Selforganizology, 2(4): 91-101

Zhang WJ. 2016. A node degree dependent random perturbation method for prediction of missing links in the network. Network Biology, 6(1): 1-11

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