Article

Simple and easy estimation of network properties based on linear correlation analysis

Yanhong Qi
Sun Yat-sen University Libraries, Sun Yat-sen University, Guangzhou 510275, China
E-mail: qiyh@mail.sysu.edu.cn

Received 8 October 2015; *Accepted 3 November 2015*; *Published online 1 December 2015*

Abstract
An ecological network can be constructed by calculating the sampling data of taxon × sample type. A statistically significant Pearson linear correlation means an indirect or direct linear interaction between two taxa, and a statistically significant partial (net, or pure) correlation based on Pearson linear correlation means a candidate direct linear interaction between two taxa. In many cases, statistically significant partial correlations are not available, or we only need to estimate some of network properties. Based on sampling data of arthropods in different countries and periods, in present study I proved that the number of candidate direct linear interactions (y) increases with the number of indirect + direct linear interactions (x) calculated by Pearson linear correlation (y=-0.2757+0.5343x, r^2=0.859, p<0.00001), and the former is approximately half of the later. The proportion of candidate direct interactions in possible maximum interactions (y%) is approximately two-thirds of mean Pearson linear correlation (x) (y=1.9060+64.6084x, r^2=0.339, p=0.023). These conclusions are expected to provide simple and easy quantities to estimate some of network properties.

Keywords network; direction interactions; Pearson linear correlation; partial linear correlation; estimation.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

In previous studies (Zhang, 2007, 2011, 2012a, 2012b), a set of methodology for constructing ecological networks by correlation analysis of community sampling data were presented. In these methods, a statistically significant Pearson linear correlation means an indirect or direct interaction between two taxa, and a statistically significant partial (net, or pure) correlation based on Pearson linear correlation means a candidate direct interaction between two taxa. However in many cases, statistically significant partial correlations are not available or we only need to estimate some of network properties. Based on sampling data of arthropods in different countries and periods (Zhang, 2011), I tried to use measures based on Pearson linear correlation to derive some network measures based on partial linear correlation. The conclusions are expected to provide simple quantities to estimate some of network properties.

2 Material and Methods

2.1 Methods

A network is globally the linear network, quasi-linear network or nonlinear network. Furthermore, a network changed in a local domain (a short time, a small extent) can be approximated as a linear network (Zhang, 2011, 2012a, 2012b; Zhang and Li, 2015), i.e., in the local domain, all between-node (or -taxon, -component, etc) changes are treated as linear ones, and linear correlation measures can thus be used. Pearson linear correlation is a measure to reflect the linear dependence between two taxa. A statistically significant Pearson linear correlation represents a direct or indirect linear interaction between two taxa. Partial (net, or pure) linear correlation is based on Pearson linear correlation. It has eliminated the indirect effects produced by the remaining taxa. A statistically significant partial linear correlation represents a candidate direct linear interaction between two taxa (Zhang, 2007, 2011, 2012a, 2012b). For a weighted network, Pearson linear correlations or partial linear correlations can be treated as the weights of connections. Linear regression analysis was also conducted in present study.

Table 1 Network information based on Pearson linear correlation and partial linear correlation (Zhang, 2011).

Network (Data set)	Taxon type	Pearson linear corr. (PLC)/Partial PLC	Sample size	No. taxa	Total No. PLCs/Partial PLCs (*N*)	Mean of PLCs/Partial PLCs	Total No. statistically significant PLCs ($p \leq 0.01$) (*L*)/Partial PLCs ($p \leq 0.01$) (*LP*)	*L*/*N* (%)/*LP*/*N* (%)
CN-06sep	Func. Group	PLC	35	4	6	0.2791	2	33.3
CN-06sep	Func. group	Partial PLC	35	4	6	0.2032	1	16.7
CN-06sep	Func. group	PLC	54	4	6	0.2663	1	16.7
CN-06sep	Func. group	Partial PLC	54	4	6	0.1804	2	33.3
CN-06Oct	Func. group	PLC	60	4	6	0.1698	2	33.3
CN-06Oct	Func. group	Partial PLC	60	4	6	0.1285	2	33.3
CN-06Oct	Func. group	PLC	60	4	6	0.1791	0	0
CN-06Oct	Func. group	Partial PLC	60	4	6	0.1364	0	0
PH-Mar	Func. group	PLC	60	21	210	-0.0003	9	4.3
PH-Mar	Func. group	Partial PLC	60	21	210	0.0029	3	1.4
PH-Apr	Func. group	PLC	60	20	190	0.0977	22	11.6
PH-Apr	Func. group	Partial PLC	60	20	190	0.0151	12	6.3
PH-Sep	Func. group	PLC	60	21	210	0.0766	13	6.2
PH-Sep	Func. group	Partial PLC	60	21	210	0.0259	4	1.9
PH-Oct	Func. group	PLC	60	21	210	0.0416	11	5.2
PH-Oct	Func. group	Partial PLC	60	21	210	0.0178	3	1.4
PH-Mar	Macro func. group	PLC	60	7	21	-0.0045	1	4.8
PH-Mar	Macro func. group	Partial PLC	60	7	21	-0.0012	2	9.5
PH-Apr	Macro func. group	PLC	60	7	21	0.2427	7	33.3
PH-Apr	Macro func. group	Partial PLC	60	7	21	0.0952	2	9.5
PH-Sep	Macro func. group	PLC	60	7	21	0.169	6	28.6
PH-Sep	Macro func. group	Partial PLC	60	7	21	0.0963	1	4.8
PH-Oct	Macro func. group	PLC	60	7	21	0.0757	2	9.5
PH-Oct	Macro func. group	Partial PLC	60	7	21	0.0508	2	9.5
CN-06Sep	Family	PLC	54	23	253	0.0529	19	7.5
CN-06Sep	Family	Partial PLC	54	23	253	0.0276	12	4.7
CN-06Oct	Family	PLC	60	23	253	0.0374	16	6.3
CN-06Oct	Family	Partial PLC	60	23	253	0.0082	7	2.8
CN-06Oct	Family	PLC	60	27	351	0.032	24	6.8
CN-06Oct	Family	Partial PLC	60	27	351	0.0171	15	4.3

2.2 Sampling data

Temporal sampling is always time-cost. However, if the environmental conditions of sampling sites are the same, spatial sampling may be used to replace temporal sampling and dynamic interactions can thus be represented by spatial changes of interactions. In present study I used the spatial sampling data and the results of correlation analysis from Zhang (2011) (Table 1). These data on biological networks are different in countries, years, seasons, types of taxa, and number of taxa.

3 Results

Let ID numbers 1-9 in Table 2 be sample size (1), total number of taxa (2), possible maximum interactions (i.e., total Pearson linear correlations, *N*) (3), mean of Pearson linear correlations (PLCs) (4), number of statistically significant direct + indirect linear interactions derived from Pearson linear correlation ($p \leq 0.01$; *L*) (5), *L*/*N* (%) (6), mean of partial linear correlations (Partial PLCs) (7), number of statistically significant candidate direct linear interactions derived from partial linear correlation ($p \leq 0.01$; *LP*) (8), and *LP*/*N* (%) (9). The calculated Pearson linear correlations between various measures are indicated in Table 2.

Table 2 Pearson linear correlations between various measures.

	1	2	3	4	5	6	7	8	9
1	1	0.283	0.240	-0.527**	0.217	-0.420	-0.614**	0.136	-0.307
2		1	0.989***	-0.715***	0.905***	-0.559*	-0.773***	0.777***	-0.588**
3			1	-0.677***	0.909***	-0.544**	-0.711***	0.812***	-0.537**
4				1	-0.496*	0.739***	0.938***	-0.450*	0.582**
5					1	-0.327	-0.629**	0.927***	-0.467*
6						1	0.644***	-0.321	0.592**
7							1	-0.525**	0.648***
8								1	-0.268
9									1

*: $p \leq 0.1$; **: $p \leq 0.05$; ***: $p \leq 0.01$; $n=15$.

As indicated in Fig. 1, linear regression relationships can be found between some measures.

The linear regression shows that the number of candidate direct linear interactions (*y*) increases with the number of indirect + direct linear interactions (*x*) calculated by Pearson linear correlation measure

$$y=-0.2757+0.5343x,\ r^2=0.859,\ p<0.00001,\ n=15$$

The number of candidate direct interactions is approximately half of the number of indirect + direct linear interactions (*x*) calculated by Pearson linear correlation measure.

The proportion of candidate direct interactions in possible maximum interactions (*y*%) is approximately two-thirds of mean Pearson linear correlation (*x*), with the linear regression as the following

$$y=1.9060+64.6084x,\ r^2=0.339,\ p=0.023,\ n=15$$

For example, suppose there are 30 taxa (maximally 435 possible interactions), and there are 30 candidate

direct interactions, then y=6.89 (%).

Mean Pearson linear (y_1) and partial linear (y_2) correlations decrease with sample size (x_1) and total number of taxa found in the sampling (x_2)

$$y_1=0.5572-0.0077x_1,\ r^2=0.278,\ p=0.044,\ n=15$$
$$y_1=0.2175-0.0077x_2,\ r^2=0.512,\ p=0.003,\ n=15$$
$$y_2=0.4327-0.0064x_1,\ r^2=0.377,\ p=0.015,\ n=15$$
$$y_2=0.1460-0.0059x_2,\ r^2=0.597,\ p=0.001,\ n=15$$

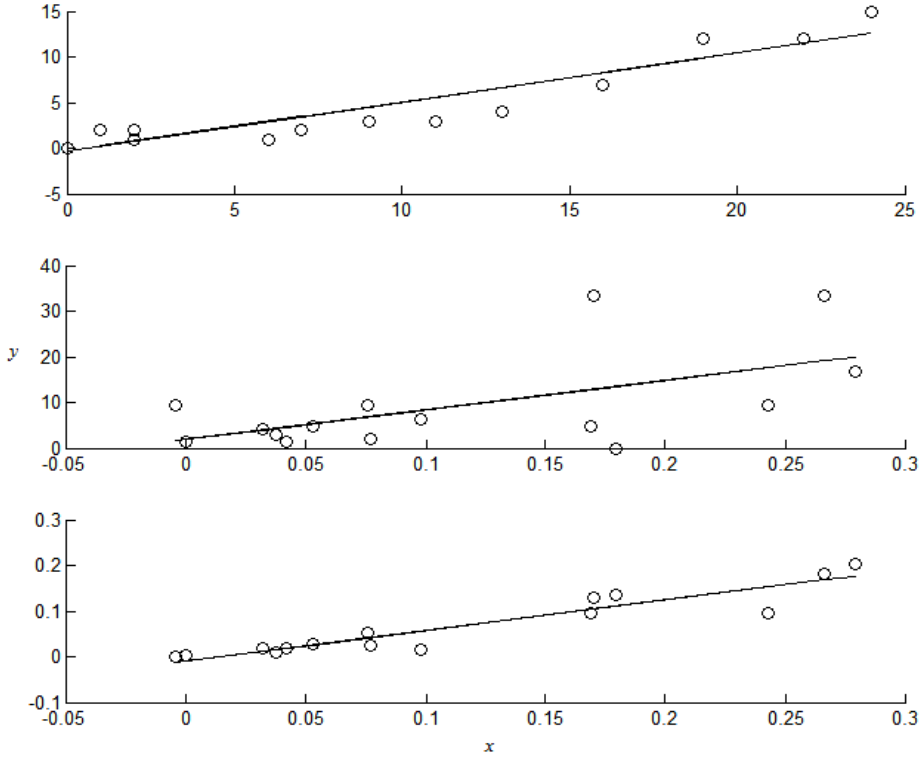

Fig. 1 Upper: linear relationship between the number of candidate direct linear interactions (y) and the number of indirect + direct linear interactions (x) calculated by Pearson linear correlation measure; Middle: linear relationship between the proportion of direct interactions in possible maximum interactions (y%) and mean Pearson linear correlation (x); Lower: linear relationship between partial linear correlation (y) and Pearson linear correlation (x).

Simplified from the results above, Table 3 summarizes the simple and easy quantities for future use.

Table 3 Simplified approximate quantities to estimate network properties.

	Pearson linear correlation (x)	Number of indirect + direct linear interactions calculated by Pearson linear correlation (x)
Number of candidate direct linear interactions (y)	-	1/2
Proportion of candidate direct interactions in possible maximum interactions (y)	2/3	-

4 Brief Discussion

(1) The conclusions may be applicable to ecological networks of terrestrial arthropods or even invertebrates. For other organisms, they should be further validated.

(2) For the situations where partial linear correlations are calculable, the full methodology of Zhang (2011) and Zhang and Li (2015) are suggested for following.

(3) The most important is that, the predicted candidate direct interactions should be tested against the direct interactions already confirmed by experiments or observations. The predicted, but not yet confirmed direct interactions can be further validated in future experiments or field observations.

Acknowledgment

I am thankful to the support of Discovery and Analysis of Important Biological and Social Networks (2015.6-2020.6), from Yangling Institute of Modern Agricultural Standardization, China. Also, I thank Prof WJ Zhang for providing data material.

References

Zhang WJ. 2007. Computer inference of network of ecological interactions from sampling data. Environmental Monitoring and Assessment, 124: 253-261

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, Li X. 2015. Linear correlation analysis in finding interactions: Half of predicted interactions are undeterministic and one-third of candidate direct interactions are missed. Selforganizology, 2(3): 39-45