Evaluation of correlation methods for construction of a co-occurrence network of rice crop health survey data

S. Jaisong , N.P. Castilla, C.T. Magculia, S. Savary, I.B. Pangga, and A.H. Sparks

Introduction

Surveys[1] conducted in rice fields of Asia have generated large data sets to characterize pests, pest injuries, and production situation.

Network analysis is widely applied in biological science and can be used to improve the efficiency of pest management. The construction of a network model is based on correlation coefficients. Methods are often selected based on the type of data to be analyzed.

Problem

Which correlation measure is suitable for collected data on pest and pest injuries?

Choosing the correlation measure

CHECK whether the data are normally distributed or not.

COMPARE the similarity of correlation coefficients from different correlation measures.

IDENTIFY which suitable correlation measure can capture biological relationships the most between variables.

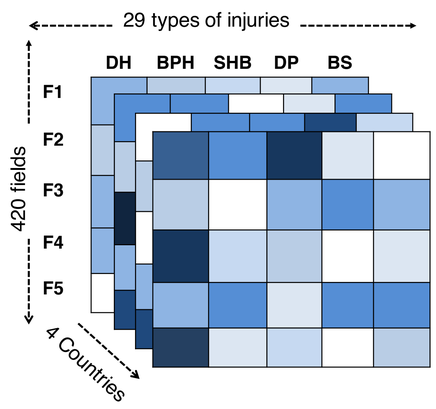
Conclusion

Spearman's rank correlation measure is the most suitable because:

- collected data are not normally distributed (Fig. 1),
- it yields the difference results of parametric correlation measures (Fig. 2), and
- it detects naturally occurring, non-linear relationships (Fig. 3).

Evaluation

A. Input (pests and pest injuries data)



CHECK Visual assessments and statistical tests were performed (Shapiro-Wilk test[2]) to check the normality of data distribution (Fig. 1).

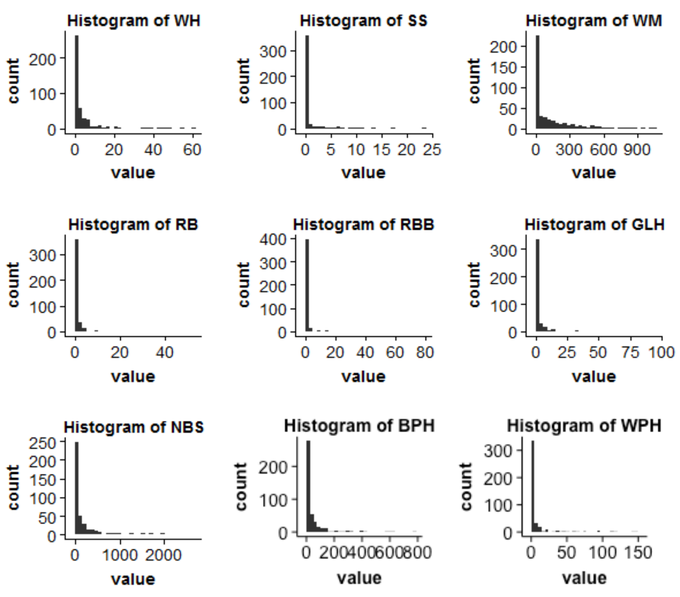
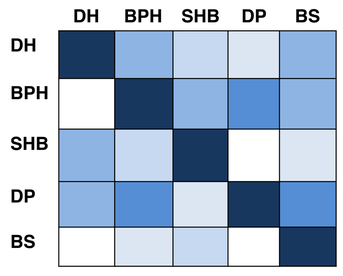


Fig. 1. Histograms of data on pests and pest injuries. The distribution of most variables are described as power law or long tail.

B. Correlation Matrix

Correlation matrix construction[4]

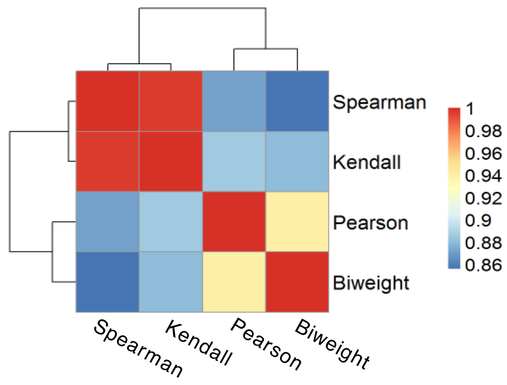


Correlation Matrix

Pairwise correlation between pests and pest injuries are computed using a suitable measure

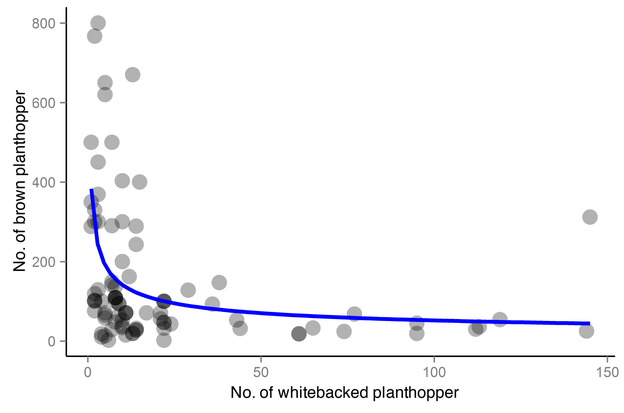
COMPARE Clustering analysis measures using Euclidian distance shows two groups of correlation methods (Fig. 2). The first is parametric correlation measures (Pearson and Biweight) and the second group is nonparametric correlation measure (Spearman and Kendall).

Fig. 2. Heatmap and dendrogram showing hierarchical unsupervised clustering analysis correlation measures of survey data.



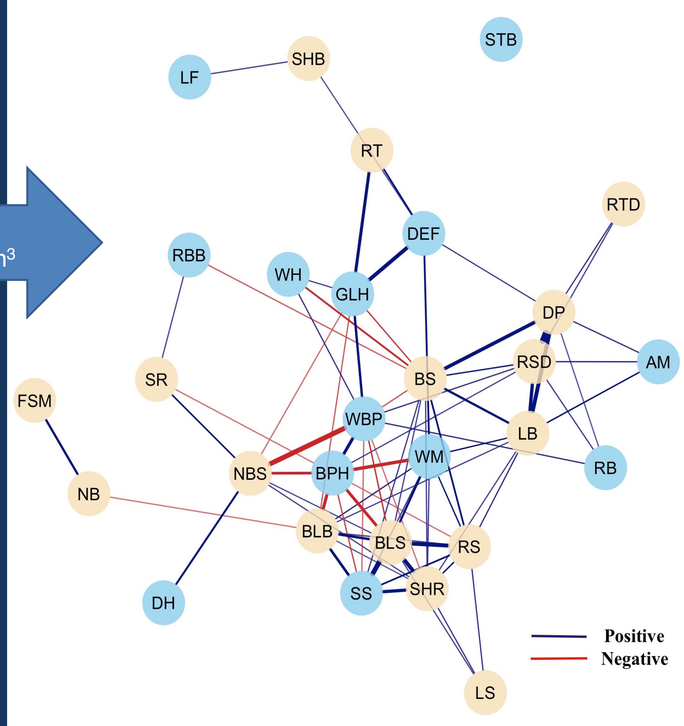
IDENTIFY The non-linear relationship between BPH and WBPH[5] was detected by Spearman correlation, but not by Pearson, Biweight, and Kendall correlations. Other correlation measures, which are able to capture non-linear correlation can be considered.

Fig. 3. Scatter plot illustrating the non-linear relationship between the brown planthopper and whitebacked planthopper.



C. Network output

Network visualization[3]



Co-occurrence network of pests - pest injuries based on Spearman's rank correlation

Blue and yellow nodes represent pest injuries and diseases, respectively. The layout of the network graph is based on the Fruchterman-Reingold algorithm, which places nodes with stronger or more connections closer to each other.

Represents animal pest group

AM = Army worm
BPH = Brown planthopper
DEF = Defoliators
DH = Deadheart
GLH = Green leafhopper
LF = Leaffolder
RT = Rat
RB = Rice bug
RBB = Rice black bug
SS = Silver shoot
STB = Stink bug
WMA = Whorl maggot
WBP = Whitebacked planthopper
WH = Whitehead

Represents disease group

BLB = Bacterial leaf blight
BLS = Bacterial leaf streak
BS = Brown spot
DP = Dirty panicle
FSM = False smut
LB = Leaf blast
LS = Leaf scald
NB = Neck blast
NBS = Narrow brown spot
RSD = Ragged stunt disease
RTD = Rice tungro disease
SHB = Sheath blight
SHR = Sheath rot
SR = Stem rot

Acknowledgments

Syngenta provided the financial support and collected the on-farm survey data as part of the SKEP (Scientific Know-How and Exchange Program) collaboration (IRRI-Syngenta).

References

[1]Savary S, Castilla NP. 2009. A survey portfolio to characterize yield-reducing factors in rice. An updated version of IRRI Discussion Paper No. 18. 32 p.
[2]R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org/.
[3]Epskamp S, Cramer AO, Waldorp LJ, Schmittmann VD, Borsboom D. 2012. qgraph: Network visualizations of relationships in psychometric data. Journal of Statistical Software 48:1-18.
[4]Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.
[5]Win SS, Muhamad R, Ahmad ZAM, Adam NA. 2011. Population fluctuations of brown planthopper, *Nilaparvata lugens* Stål. and whitebacked planthopper, *Sogatella furcifera* Horvath on rice. Journal of Entomology 8:183-190.

Find out more