Network Analysis of Cropping Practices and Injury Profiles in Irrigated Rice Agroecosystems

Sith Jaisong

August 30, 2015

Abstract

Here is the abstract...........

Introduction

The use of in-field surveys is a useful tool to develop ground-truth databases that allow one identify actual constraints due to pests in an agricultural productions system. These sorts of databases provide an overview of the complex relationships between the crop, its management, pest injuries, (and what else?). In turn this may lead to better management, and guide researchers the new research hypotheses (Mew et al., 2004; Savary et al., 2006).

Several previous studies (Savary et al., 2000b,a, 2005; Dong et al., 2010; Reddy et al., 2011) involved surveys that have been used to identify relationships in an individual production situation (a set of factors that determine agricultural production) and the injury profiles (combination of disease and pest injuries that may occur in a given farmer's field) using nonparametric multivariate analysis (Savary et al., 1997). Performing correspondence analysis, they characterized the relationships between categorized levels of variables: actual yield, production situations and injuries profiles. Their results led to the conclusions that observed injuries profiles are strongly associated with production situations(Mew et al., 2004). However, the step of transforming continuous data to categorized data by clustering approach is poor reproducibility and difficult interpretability of the individual cluster (Jiang et al., 2004; Avelino et al., 2006).

The components of production situation and injury profiles, such as the amount of fertilizers applied, occurrence of leaf blast are biologically related. The relationships will be more complex when the number of their components are increased. A way to systemically model and intuitively interpret such relationships is the depiction as a graph or network. This approach has been widely used and proven very useful in biological studies. Networks typically consist of nodes, usually representing components (), while links between the nodes depict their interactions. A correlation network is a type of network in which two nodes are connected if their respective correlation lies above a certain threshold. The construction of this network is obtained from pairwise correlation methods. By using appropriate correlation measure, correlation networks can capture biologically meaningful relationships, and discover valuable information in crop health surveys.

Selecting the suitable association methods for network construction is important because the method that can capture the relationships with true concordance often determined the type and amount of knowledge we can gain from survey data.

Wwe have limited prior knowledge (positive relation and negative relation) for comparing the efficiency of different association methods in discovering true functionally associated variables.

The main aim of this article is to evaluate correlation methods including Pearson, Spearman, Kandell, Biweight to associate the components of cropping practices and the components of injuries profiles. Furthermore, we applied network theory and model to illustrate the pairwise relationships. Thus we hope to provide the necessary elements for a better comprehension of the methods and also the choice of a suitable dependence test method based on practical constrains and goals.

Materials and Methods

We inferred an interaction network from survey project comprising five countries (India, Indonesia, Philippines, Thailand, and Vietnam), 420 lowland farmers' fields. Our study aimed to determine the co-occurrence patterns among the incidence of injuries caused by animal pests and diseases and the cropping practices, potentially indicative of their occurrence relations. We thus constructed the network from the surveys. The limitation of each measure are difference assumption and detach different patterns.

Survey datasets

Crop health survey data were collected through surveys comprising 420 farmers' fields from 2010 to 2012 for wet and dry seasons in different production environments across South and South East Asia. The survey protocol described in the IRRI publication, "A survey portfolio to characterize yield-reducing factors in rice", was used for data collection (Savary and Castilla, 2009). The variables collected included patterns of cropping practices, crop growth measurement and crop management status assessments, measurements of levels of injuries caused by pests, and direct measurements of actual yields from crop cuts. The data collected can be classified into three groups: cropping practices, injuries, and actual yield measurements.

Evaluation of association methods

The criteria of choosing correlation measures

Step one: Data exploration

There are three main properties to be determined before deciding the appropriate correlation measure for use in constructing the network.

Check data distribution This test can be achieved by significance test and visual methods. Each variable in survey dataset was tested normality using the Shapiro-Wilk test (Ghasemi and Zahediasl, 2012). The Shapiro-Wilk test is based on the correlation between the data and the corresponding normal scores.

H_0: sample distribution is normal.

H_a: sample distribution is not normal.

Thus if the p-value is less than the chosen alpha level, the null hypothesis is rejected and there is evidence that the data tested are not from a normally distributed population. In other words, the data are not normal. On the contrary, if the p-value is greater than the chosen alpha level, then the null hypothesis that the data came from a normally distributed population cannot be rejected. However, for small sample sizes, normality tests have little power to reject the null hypothesis, so a QQ (quantile–quantile plot) plot and the frequency distribution (histogram) are required for verification in addition to check normality visually.

The `R` function for Shapiro-Wilk Normality test is `shapiro.test` (package stats). The stats package can be downloaded from the `R` web page (http://www.r-project.org) (R Core Team, 2014).

Check for independence Performing the distance correlation t-test(Székely et al., 2007) to check independence aims to select the the pair of variables which are able to be detected correlation. The distance correlation of two random variables is obtained by dividing their distance covariance by the product of their distance standard deviations. The distance correlation is

$$\mathrm{dCor}(X,Y) = \frac{\mathrm{dCov}(X,Y)}{\sqrt{\mathrm{dVar}(X)\ \mathrm{dVar}(Y)}} \tag{1}$$

- $0 \leq \mathrm{dCor}_n(X,Y) \leq 1$ and $0 \leq \mathrm{dCor}(X,Y) \leq 1$
- $\mathrm{dCor}(X,Y) = 0$ if and only if X and Y are independent.
- $\mathrm{dCor}_n(X,Y) = 1$ implies that dimensions of the linear subspaces spanned by X and Y samples respectively are almost surely equal and if we assume that these subspaces are equal, then in this subspace $Y = A + b\,\mathbf{C}X$ for some vector A, scalar b, and orthonormal matrix $\mathbf{C}$.

This test was performed using the `fda.uss` package (Febrero-Bande and Oviedo de la Fuente, 2012).

Check linearity or non–linearity

Step two: identify the most appropriate method

Pearson's product-moment correlation coefficient

The Pearson's product-moment correlation or simply Pearson's correlation is a measure of linear dependence, as the slope obtained by the linear regression of Y by X is Pearson's correlation multiplied by that ratio of standard deviations. Let $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ be the means of X and Y, respectiverly, then the Peasons's corrlation coefficient $\rho_{Pearson}$ is defined as follows:

$$\rho_{Pearson}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2}$$

For joint normal distributions, Pearson's correlation coefficient under H_0 follows a Student's t-distribution with $n-2$ degrees of freedom. The t statistic is as follows:

$$t = \frac{\rho_{Pearson}(X,Y)\sqrt{n-2}}{\sqrt{1-\rho^2_{Pearson}(X,Y)}} \tag{3}$$

When the random variables are not jointly normally distributed, the Fisher's transformation is used to get an asymptotic normal distribution.

In the case of perfect linear dependence, we have $\rho_{Pearson} = \pm 1$. The Pearson correlation is $+1$ in the case of a perfect positive (increasing) linear relationship and -1in the case of a perfect negative (decreasing) linear relationship. In the case of linearly independent random variables, $rho_{Pearson} = 0$, and in the case of imperfect linear dependence, $-1 < \rho_{Pearson} < 1$. These last two cases are the ones for which misinterpretations of correlation are possible because it is usually assumed that non-correlated X and Y means independent variables, whereas in fact, they may be associated in a non-linear fashion that Pearson's correlation coefficient is not able to identify.

The `R` function for Pearson's test is `cor.test` with parameter method 'Pearson' (package stats). The stats package can be downloaded from the R web page (http://www.r-project.org).

Network Construction

Co-occurrence network construction

The matrix can be viewed as an adjacency matrix of a weighted network. The matrix contains the correlation coefficient between each node (i.e., the variable). Thus the matrix can be thought of as the population average of the network structure. Because we are looking at several specific links, we control for multiple testing by controlling the False Discovery Rate (FDR method) at 5%. The generated network structure can be visualized through the R package qgraph. Only connections that surpass the significance threshold are shown in the visual representation.

Network analysis

Important information about a network can be gained by analyzing its global structure, for example by looking at the relative centrality of different nodes. In a centrality analysis, nodes are ordered in terms of the degree to which they occupy a central place in the network. Global descriptors of the modules were obtained using package qgraph in R. The neighborhood of a given node n is the set of its neighbors. The connectivity is the size of its neighborhood. The average number of neighbors indicates the average connectivity of a node in the network. A normalized version of this parameter is the network density. Density ranges between 0 and 1. It shows how densely the network is populated with edges. A network which contains no edges and solely isolated nodes has a density of 0. In contrast, the density of a clique is 1. Another related parameter is the network centralization. Networks whose topologies resemble a star have a centralization close to 1, whereas decentralized networks are characterized by having a centralization close to 0.

In undirected networks, the clustering coefficient is the number of connected pairs between all neighbors of the network. The clustering coefficient of a node is always a number between 0 and 1. The network clustering coefficient is the average of the clustering coefficients for all nodes in the network. Nodes with less than two neighbors are assumed to have a clustering coefficient of 0. We then determined network centralities on the modules obtained from network analysis. Centralities were assessed using package in R. We calculated Degree centrality and Betweenness centrality.

Results

Discussion

References

Avelino, J., Zelaya, H., Merlo, A., Pineda, A., Ordoñez, M., and Savary, S. (2006). The intensity of a coffee rust epidemic is dependent on production situations. *Ecological Modelling*, 197(3-4):431–447.

Dong, K., Chen, B., Li, Z., Dong, Y., and Wang, H. (2010). A characterization of rice pests and quantification of yield losses in the japonica rice zone of yunnan, china. *Crop Protection*, 29(6):603–611.

Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28.

Ghasemi, A. and Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 10(2):486.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 16(11):1370–1386.

Mew, T. W., Leung, H., Savary, S., Vera Cruz, C. M., and Leach, J. E. (2004). Looking ahead in rice disease research and management. *Critical Reviews in Plant Sciences*, 23(2):103–127.

R Core Team (2014). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Reddy, C. S., Laha, G. S., Prasad, M. S., and Krishnaveni, D. (2011). Characterizing multiple linkages between individual diseases, crop health syndromes, germplasm deployment, and rice production situations in India. *Field Crops.*

Savary, S. and Castilla, N. (2009). A survey portfolio to characterize yield-reducing factors in rice. *IRRI Discussion Paper No 18*, page 32.

Savary, S., Castilla, N. P., Elazegui, F., and Teng, P. S. (2005). Multiple effects of two drivers of agricultural change, labour shortage and water scarcity, on rice pest profiles in tropical asia. *Field Crops Research*, 91(2):263–271.

Savary, S., Elazegui, F., Pinnschmidt, H., Castilla, N., and Teng, P. (1997). A new approach to quantify crop losses due to rice pests in varying production situations. *IRRI discusion paper series no20. International Rice Research Institute, PO Box*, 933.

Savary, S., Teng, P. S., Willocquet, L., and Nutter, F. W. (2006). Quantification and modeling of crop losses: a review of purposes. *Annual Review of Phytopathology*, 44(1):89–112.

Savary, S., Willocquet, L., Elazegui, F. A., Castilla, N. P., and Teng, P. S. (2000a). Rice pest constraints in tropical Asia: quantification of yield losses due to rice pests in a range of production situations. *Plant disease*, 84(3):357–369.

Savary, S., Willocquet, L., Elazegui, F. A., Teng, P. S., Van Du, P., Zhu, D., Tang, Q., Huang, S., Lin, X., and Singh, H. M. (2000b). Rice pest constraints in tropical Asia: characterization of injury profiles in relation to production situations. *Plant Disease*, 84(3):341–356.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.