%\section{Evaluation of correlation methods for co-occurrence network construction of rice crop health survey data}

\subsection\*{Introduction}

Rice is not threatened by one but by many pests in a season. A combination of injuries caused by diseases and rice pests can be thought of as a crop health syndrome. The combinations of injuries depend on the production situation (\textit{i.e.}, the cultural practices, inputs used to produce a rice crop) as a range of agroecosystem \citep{Savary\_2006\_Quantification}.

A characterization study based on survey data collected in South and South East Asia \citep{Savary\_2000\_Characterization} showed the patterns of injury profiles (combinations of injuries by rice pests, IN) were common and different across sites. The result indicated that sheath blight, brown spot and leaf blast are the important diseases were commonly found in some sites, causing yield loss between 1 to 10%,and among insect injuries, stem borer causing yield losses of 2.3%. The patterns of injury profiles were clustered in five groups. For example, cluster1 (IN1) was characterized by comparatively high incidence of stem rot, sheath blight, plant hopper, and whorl maggot injuries, but low brown spot, and absence of bacterial leaf blight, leaf blast, and neck blast.

Networks are ubiquitous systems in nature, technology and society (Newman, 2003). A network is defined as one or more sets of nodes connected by links in various ways. A node can represent the individual units depending on the context. Links or edges are the connections between nodes, which may be directed or undirected. Network models are now becoming increasingly interesting and useful in social science, biology, and ecology. The network applications relevant to plant pathology were also increasingly studied \citep{Moslonka\_Lefebvre\_2011}.

Network analysis is a promising tool frequently used to describe the pairwise relationships of a large number of variables. For example, association networks or correlation networks were represented by their association or correlation (adjacency) matrices, which rows and columns denote nodes, and matrix entries denote links. They were widely applied in biological studies \citep{Toubiana\_2013\_Net, Barabasi\_2004\_Network}

In this chapter, selecting the most suitable correlation methods for correlation network construction is important since different correlation measures lead to different network structure and provide different information. I evaluated four correlation methods, including Pearson, Spearman rank correlation, Kendall correlation \citep{Prokhorov\_2001\_Kendall} and Biweight mid-correlation, to associate rice injuries.

\subsection\*{Materials and methods}

\textbf{Survey data}

Survey data collected in 450 farmers’ fields in irrigated lowland rice growing areas across South and Southeast Asia Tamil Nadu, India (TMN); Odisha, India (ODS), West Java; Indonesia (WJV); Suphanburi, Thailand (SPB) and Mekong river delta, Vietnam (MKD) were collected from 2013 - 2016. The number of survey were summerize in Table.

\ref{table:Survey\_data}.

The survey procedure and data were based on a standardized protocol described in ``A survey portfolio to characterize yield-reducing factors in rice'' developed by \citet{Savary\_2009\_Survey}. Thirty rice injuries were collected

including the injuries caused by animal pests, and pathogens, which are harmful to rice plants, and importantly considered to reduce yield productivity. They were evaluated at booting and ripening stage according to survey procedure. They expressed in different unites depending on nature of injuries found to particular plant organs.

Injuries on leaves such as whorl maggot injury (WM), leaffolder injury (LF), bacterial leaf blight (BLB), bacterial leaf streak (BLS), leaf blast (LB), brown spot (BS), leaf miner injuries (LM), leaf scald (LS), narrow brown spot (NBS), rice hispa injury (RH), red stripe (RS), rice thrip injury (RTH) were determined as a proportion of injured leaves. Injuries on tillers or hills such as stem rot (SR), sheath rot (SHR), sheath blight (SHB), whitehead (WH), deadheart (DH), silver shoot (SS), false smut (FS), Neck blast (NB), Panicle mite injury (PM), Rice bug injury (RB), rat injury (RT) were determined as a proportion of injured tillers or panicles. Systemic injuries such as Bug burn (BB), grassy stunt (GS), hopper burn( HB) , ragged stunt (RGS), tungo (RTG) were determined as the percentage of area affected. The rice injury lists were showed in Table.

Before analysis, data were compacted over time during crop growth. Two types of data were computed, depending on the natures of injuries \citet{Savary\_2009\_Survey}. One is an area under injury progress curve (AUIPC) used for injury variables, which present on the leaves, and for weed infestation. Another is the maximum level at any of the two observations used for injury variables that can be observed on tillers, panicles, and hills, and insect pest count. The area under injury progress curve (AUIPC) \citep{Campbell\_1990\_Introduction} were calculated by the mid-point method using the following equation:

\begin{equation}

AUIPC = \sum{\frac{1}{2(X\_{i} + X\_{i-1})(T\_{i} - T\_{i-1})}}

\end{equation}

where $X\_i$ is percentage (\%) of leaves, tillers or panicles injured due to rice pests (e.g., leaf blast, leaf folder), or number of insects (e.g., plant hoppers, leaf hoppers) per quadrat, or percentage (\%) of weed infestation (ground coverage) at the $i$th observation, $T\_i$ is time in rice development stage units (dsu) on a 0 to 100 scale (10: seedling, 20: tillering, 30: stem elongation, 40: booting, 50: heading, 60: flowering, 70: milk, 80: dough, 90: ripening, 100: fully mature) at the $i$th observation and $n$ is total number of observations.

\text{Evaluation}

In this study, correlation measures including, Pearson, Spearman, Kendal and Biweight mid correlation were evaluated to discover the true functionally related variables in crop health survey data. The data will have to follow the assumption of correlation measures. The correlation measures will also be able to effectively capture biological relationships that are well published. I proposed three steps for correlation methods selection:

\begin{itemize}

\item Testing whether or not the data are normally distributed by visual assessments and statistical tests. I examined the distribution of values of rice injuries in crop health survey data, and tested the hypothesis hypothesis that the sample comes from a population which has a normal distribution by performed Shapiro-Wilk test.

\item Comparing correlation measures by testing similarity of correlation coefficients. I evaluated the similarity of correlation coefficients of different correlation measures by using the Euclidean distance, and perform clustering analysis.

\item Identifying the most suitable correlation measure that can capture biological relationships between variables confirm with the published relationships.

\end{itemize}

\subsection\*{Result}

\textbf{Distribution of crop health survey data}

To determine normality of the survey data, I presented the histograms showing distribution of value of rice injuries, calculated summary statistics, and performed Shapiro-Wilk test. The distribution of value of rice injuries were visualized in histogram in Fig. The histograms depicted the distribution of values of injuries. Apparently, histograms showed that values of injuries are skewed to the left. Common values of the injuries were 0. A few farmers’ fields presenting in high values of injuries were relatively low. The distribution of most injuries are described power low or long tails.

The summary statistics, and the result of Shapiro-Wick test of each injury were calculated and summarized in Table. As can be seen, from the previous observations that rice injuries histograms tend to be positively skewed, the median values of injuries were considered due to their insensitivity to outliers. Median of injuries were almost 0, except LF and WH, which were 76.74 and 1.67, respectively. According to \citet{Doane\_2011\_Measuring}, skewness and kurtosis values were more than 0, which mean that the values of injuries were asymmetrically distributed with a long tail to the right. The normality is defined as p value 0.01 in Shapiro-Wilk testing. That test indicates that values of injuries were not normally distributed.

\textbf{Comparing}

I performed pair-wise analysis between each of injuries using all four correlation methods (Pearson, Spearman, Kendall correlation, and Biweight mid-correlation). I examined the similarity of correlation coefficients and clustered according to hierarchical clustering using Euclidean distance. The result was shown in Figure. Two groups of correlation methods. The first is parametric correlation measures (Pearson correlation and Biweight mid-correlation) and the second group is nonparametric correlation measure (Spearman correlation and Kendall correlation). Next, output for each method was sorted by $p$-values in ascending order and cut off at $p$ value \< 0.05 (Table to). Spearman method could capture 193 pairwise relationships, following with Biweight-mid correlation Pearson method captures and Kendall, which could captured 128, 122, and 72, respectively.

\textbf{Identify the most suitable correlation measure}

According to the previous results, the group of parametric correlation measures was selected out because these measures required the data normally distributed. So I would consider the correlation measures in the group of rank based methods that do not required normality. Compared between Spearman and Kendall correlation, the pair-wise relationships of rice injuries were captured differently from each correlation method. Dirty panicle – brown spot relationship was captured by Spearman correlation methods, but not by Kendall methods. This relationship was report in many studies \citep{Ou\_1985\_Rice, Barnwal\_2013\_Review}

\subsection{Discussion}

Evaluation of four correlation measures, including Pearson, Spearman, Kendall correlation, and Biweight mid-correlation shows two groups clustered according to hierarchical clustering using Euclidean distance. The Spearman and Kendall correlation were grouped in rank-base methods, and another group is non-rank-based correlations including Pearson correlation and Biweight mid-correlation. When choosing a measure, Spearman correlation was selected because its robustness to noise and to outliers.

\citet{Doane\_2011\_Measuring} mentioned that values far from zero suggest a non-normal (skewed) population.

A distribution more peaked than a Gaussian distribution has a positive kurtosis,

Although principles of statistical operation play a key role in determining the efficiency of different methods, we should not ignore the biological models underpinning each data set that can make a statistical method less efficient than another. An example for this is Pearson and Spearman methods. Charles Spearman proposed rank correlation in 1904 [37], a non-parametric version of the conventional Pearson correlation.

ก่อนที่เรา จะเข้าประเด็นตรงนี้ เราน่า

Among the four correlation methods, Spearman and Kendall are nonparametric rank-based methods. This class of methods uses ranks for correlation and therefore provides a robust measure of a monotonic relationship between two continuous random variables. They are also useful with ordinal data and are generally more robust to outliers. For this reason, they are particularly suitable for identifying key genes that increase or decline in monotonic fashions in expression data collected during a biological process or developmental stage. In a previous study, the efficiency of the Kendall test and Spearman’s rho test in detecting monotonic trends in time series data are compared [30] and the conclusion is that the two methods have similar powers that depend on the pre-assigned significance level, magnitude of trend, sample size, and the variation within a time series. That is, the bigger the absolute magnitude of trend, the more powerful is the test; as the sample size increases, the test becomes more powerful; and as the amount of variation increases within a time series, the power of the test decreases. When a trend is present, the power is also dependent on the distribution type and the skewed nature of the time series.

Biweight midcorrelation has been shown to be more robust in evaluating similarity in gene expression networks.

However, Newson [30] has argued for the superiority of Kendall’s t over Spearman’s correlation rho as a rank-based measure of correlation because confidence intervals for Spearman’s rho are less reliable and less interpretable than confidence intervals for Kendall’s t-parameters. According to Fujita et al [16], the Hoeffding’s D measure may be used to infer both nonlinear and non-monotonic relationships between gene expression profiles with full control of type I error. Theil-Sen and Rank Theil-Sen methods are regression-based methods. Theil-Sen estimator is a median of the slopes determined by all pairs of sample points, and it provides accurate estimate and confidence intervals even when the data are non-normal and heteroscedastic. Pearson’s correlation is a measure of the linear relationship between two continuous random variables, and it assumes a bivariate normal distribution. Only when the sample size is large enough will the data be close to bivariate normal

Although we can opt for a method based on its principle of statistical operation without paying attention to the biological models in a given data set, this may not lead to a coordination network that will reveal biological knowledge.

High dimensional biological data from microarray or high throughput sequencing data often contain at least a few hundred different biological processes. There is no statistical method that is suitable for all of them. Identification of the most efficient method for knowledge discovery of a specific biological process demands concrete pre- diagnostic analyses. Based on our study and our empirical knowledge, we would suggest the following procedure for identifying the most appropriate gene association method for a specific biological theme in a given data set: (1) Evaluate the prior knowledge of biological processes of one’s interest, and select a few known genes involved in these processes; (2) Use the R codes from this study to perform a genome-wide coexpression analysis to obtain the top 100 or 500 genes that are most closely associated to the selected known genes; (3) Perform an evaluation of these 100 or 500 genes by examining which methods can associate the more functionally relevant genes to the selected genes. This can be achieved by examining gene annotation or performing GO term enrichment analysis: and (4) Choose the best method for the data. However, if prior knowledge of biological theme of one’s interest is lacking, we suggest the most stable gene association method. Generally speaking, Spearman or Rank Theil-Sen is recom- mended for constructing co-expression network, and Hoeffding, Kendall or Spearman for pathway gene analysis.

The appropriate correlation measures for studied data should closely associate to the prior knowledge of biological correlation. For this study, nonlinear relationship between the abundances of brown plant hopper and white backed plant hopper was detected by a Spearman correlation but not by a Pearson correlation.

However, Pearson correlation coefficient is sensitive to outliers. Biweight midcorrela- tion is considered to be a good alternative to Pearson correlation since it is more robust to outliers [23Wilcox R: Introduction to Robust Estimation and Hypothesis Testing. Academic Press, San Diego; 1997.].

\subsection{Conclusions}

The analyses I have performed clearly demonstrate the distinct and common performance of four correlation methods.

The Spearman, Kendall methods performed very well with some minor discrepancies, which are rooted in their similar principles of operation. The Pearson and methods are distinct and generally are less valuable for identifying biologically or functionally associated genes.

Unfortunately, the efficiency of different methods indeed varies with the biological processes. For this reason, identification of the best method for a specific biological process requires some pre-analyses to be done first, which can be facilitated by the R programs we provided.

Pearson has been widely used in most co-expression analyses \citep{Zhang\_2005\_General}. However, Pearson correlation also limit to capture the linear relationships.

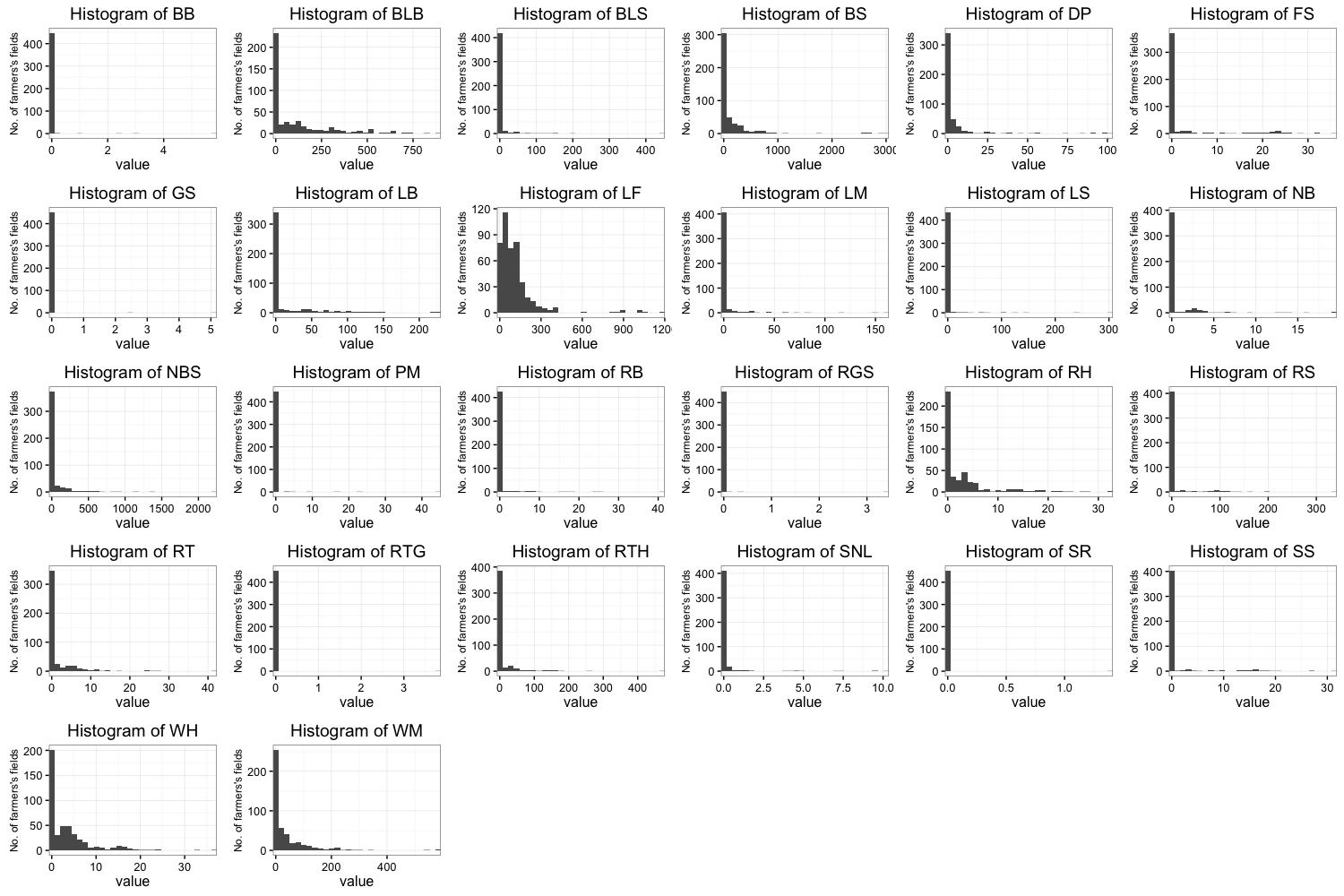


Fig1. Histograms showing the distribution of value of rice injuries in crop health survey data.