**CHAPTER III**

**Evaluation of correlation methods for constructing a network of rice injuries of rice crop health survey data**

**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 (*i.e.*, the cultural prac- tices, inputs used to produce a rice crop) as a range of agroecosystem (Savary et al.,

2006).

A characterization study based on survey data collected in South and South East Asia (Savary et al., 2000b) 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

38

39 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. Net- work 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 (Moslonka-Lefebvre et al., 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 cor- relation networks were represented by their association or correlation (adjacency) ma- trices, which rows and columns denote nodes, and matrix entries denote links. They were widely applied in biological studies (Toubiana et al., 2013; Barabási and Oltvai,

2004)

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 meth- ods, including Pearson, Spearman rank correlation, Kendall correlation (Prokhorov,

2001) and Biweight mid-correlation, to associate rice injuries.

**Materials and methods**

The method for evaluating and selecting correlation methods to construct cor- relation network of rice injuries in crop health survey data was showed in Fig. Finally,

**Choosing the correlation measure**

29 types of injuries

**DH WH SHB DP BS**

**F1**

TEST COMPARE

IDENTIFY

whether data are normally distributed or not

the similarity of correlation coefficients from different correlation measures

which the most suitable correlation measure can capture biological

Correlation coefficient

**F2**

Value

**F3** low

relationships between variables

**DH**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Pearson correlation

**DH WH SHB DP BS**

low

**F4**

**F5**

high

Spearman correlation

Kendall’ correlation

**WH**

**SHB**

high

Biweight mid-correlation **DP**

Crop health survey data

**BS**

Correlation Matrix

Network model

Figure 3-1: Selecting of correlation methods for constructing network model. The plots on the left show sample survey data taken over 29 rice injuries at 5 production environments. The next step is to compute pair-wise correlations to obtain a cross-correlation matrix (the middle plot, rows and columns represent the nodes). Then, the matrix will be used for network analysis.

correlation metrics were used for the network analysis in next chapter.

**Survey data**

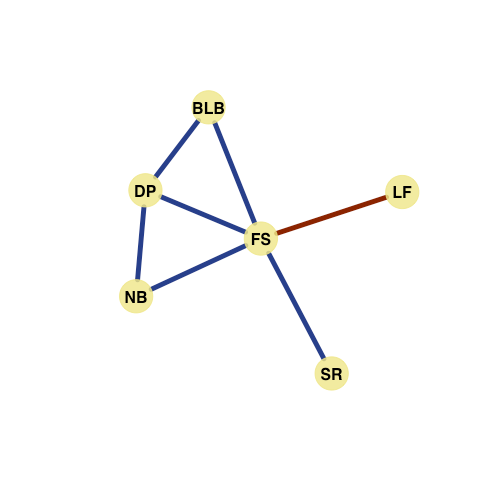
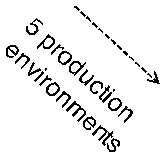
Survey data collected in 450 farmers’ fields in irrigated lowland rice growing areas across South and Southeast Asia, Central Plain; Thailand (CTP), Odisha; India (ODS), Red River Delta; Vietnam (RRD), Tamil Nadu; India (TMN), and West Java; Indonesia (WJV) were collected from 2013 - 2015. The number of survey were sum- merize in Table III.1.

Production environment

year

2013 2014 2015

Total



40

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | DS | WS | DS | WS | DS | WS |  |
| Central Plain, Thailand | 20 | 20 | 14 | 21 | 15 | 12 | 102 |
| Odisha, India | 15 | 12 | 15 | 16 | 15 | 15 | 88 |
| Red River Delta, Vietnam | 15 | 15 | 15 | 15 | 15 | 15 | 90 |
| Tamil Nadu, India | 15 | 15 | 15 | 14 | 15 | 15 | 89 |
| West Java, Indonesia | 15 | 15 | 14 | 15 | 15 |  | 74 |
| Total | 80 | 77 | 73 | 81 | 75 | 57 | 443 |

Table III.1: Number of farmers’ fields surveyed by production environment and year

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 Savary and Castilla (2009). Twenty-nine 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 bacterial leaf blight (BLB), bacterial leaf streak (BLS), brown spot (BS), leaf blast (LB), leaffolder injury (LF), leaf miner injuries (LM),leaf scald (LS), narrow brown spot (NBS), rice hispa injury (RH), red stripe (RS), rice thrip injury (RTH), and whorl maggot injury (WM) were determined as a proportion of in- jured leaves. Injuries on tillers or hills such as deadhead (DH), dirty panicle (DP), false smut (FS), neck blast (NB), panicle mite injury (PM), rice bug injury (RB), rat injury (RT), sheath rot (SHR), sheath blight (SHB), stem rot (SR),silver shoot (SS), and white- head (WH), were determined as a proportion of injured tillers or panicles. Systemic in- juries 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 III.2.

42

Table III.2: Variables describing individual fields in surveys of rice injuries in five production environments in tropical Asia

|  |  |  |  |
| --- | --- | --- | --- |
| Injury variables | Acronym | Description | Unit |
| Bug burn | BB | maximum percentage of bugburn in a one-sqm area | % |
| Bacterial leaf blight | BLB | area under the progress curve of the mean percentage of leaves with bacterial leaf blight | % dsu |
| Bacterial leaf blight | BLS | area under the progress curve of the mean percentage of leaves with bacterial leaf streak | % dsu |
| Brown spot | BS | area under the progress curve of the mean percentage of leaves with brown spot | % dsu |
| Deadheart | DH | maximum percentage of tillers with deadheart | % |
| Dirty panicle | DP | maximum percentage of panicles with dirty panicle | % |
| False smut | FS | maximum percentage of panicles with false smut | % |
| Grassy stunt | GS | maximum percentage of grassy stunt disease in a one-sqm area | % |
| Hopper burn | HB | maximum percentage of hopperburn in a one-sqm area | % |
| Leaf blast | LB | area under the progress curve of the mean percentage of leaves with leaf blast | % dsu |
| Leaffolder | LF | area under the progress curve of the mean percentage of leaves with leaffolder injury | % dsu |
| Leafminer | LM | area under the progress curve of the mean percentage of leaves with leaf miner injury | % dsu |
| Leaf scald | LS | area under the progress curve of mean percentage of leaves with leaf scald | % dsu |
| Neck blast | NB | maximum percentage of panicles with neck blast | % |
| Narrow brown spot | NBS | area under the progress curve of the mean percentage of leaves with narrow brown spot | % dsu |
| Panicle mite injury | PM | maximum percentage of tillers with panicle mite injury | % |
| Rice bug injury | RB | maximum percentage of panicles with rice bug injury | % |
| Ragged stunt | RGS | maximum percentage of grassy stunt disease in a one-sqm area |  |
| Rice hispa | RH | area under the progress curve of the mean percentage of leaves with rice hispa injury | % dsu |
| Rat injury | RT | maximum percentage of tillers with rat injury | % |
| Red stripe | RS | area under the progress curve of mean percentage of leaves with red stripe | % dsu |
| Rice tungro | RTG | maximum percentage of tungro in a one-sqm area | % |
| Rice thrip injury | RTH | area under the progress curve of the mean percentage of leaves with rice thrip injury | % dsu |
| Sheath blight | SHB | maximum percentage of tillers with sheath blight | % |
| Sheath rot | SHR | maximum percentage of tillers with sheath rot | % |
| Stem rot | SR | maximum percentage of tillers with stem rot | % |
| Silver shoot | SS | maximum percentage of tillers with silvershoot | % |
| Whitehead | WH | maximum percentage of panicles with whitehead | % |
| Whorl maggot injury | WM | area under the progress curve of the mean percentage of leaves with whorl maggot injury | % dsu |

Before analysis, data were compacted over time during crop growth. Two types of data were computed, depending on the natures of injuries Savary and Castilla (2009). 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, pan- icles, and hills, and insect pest count. The area under injury progress curve (AUIPC) (Campbell et al., 1990) were calculated by the mid-point method using the following

equation:

*AU I P C* = ∑ 1

2(*Xi* + *Xi−*1)(*Ti − Ti−*1 )

(III.1)

where *Xi* 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,

*Ti* 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.

Evaluation In this study, correlation measures including, Pearson, Spearman, Kendal and Biweight mid-correlation (Wilcox, 2012) 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:

• **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.

• **Comparing** correlation measures by testing similarity of correlation coefficients.

I evaluated the similarity of correlation coefficients of different correlation mea- sures by using the Euclidean distance, and perform clustering analysis.

• **Identifying** the most suitable correlation measure that can capture biological re- lationships between variables confirm with the published relationships.

**Result**

**Checking and testing the distribution of crop health survey data**

To determine normality of the survey data, I presented the histograms show- ing 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 3-2. The histograms depicted the distribution of values of injuries. Apparently, histograms showed that values of injuries were 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.

45

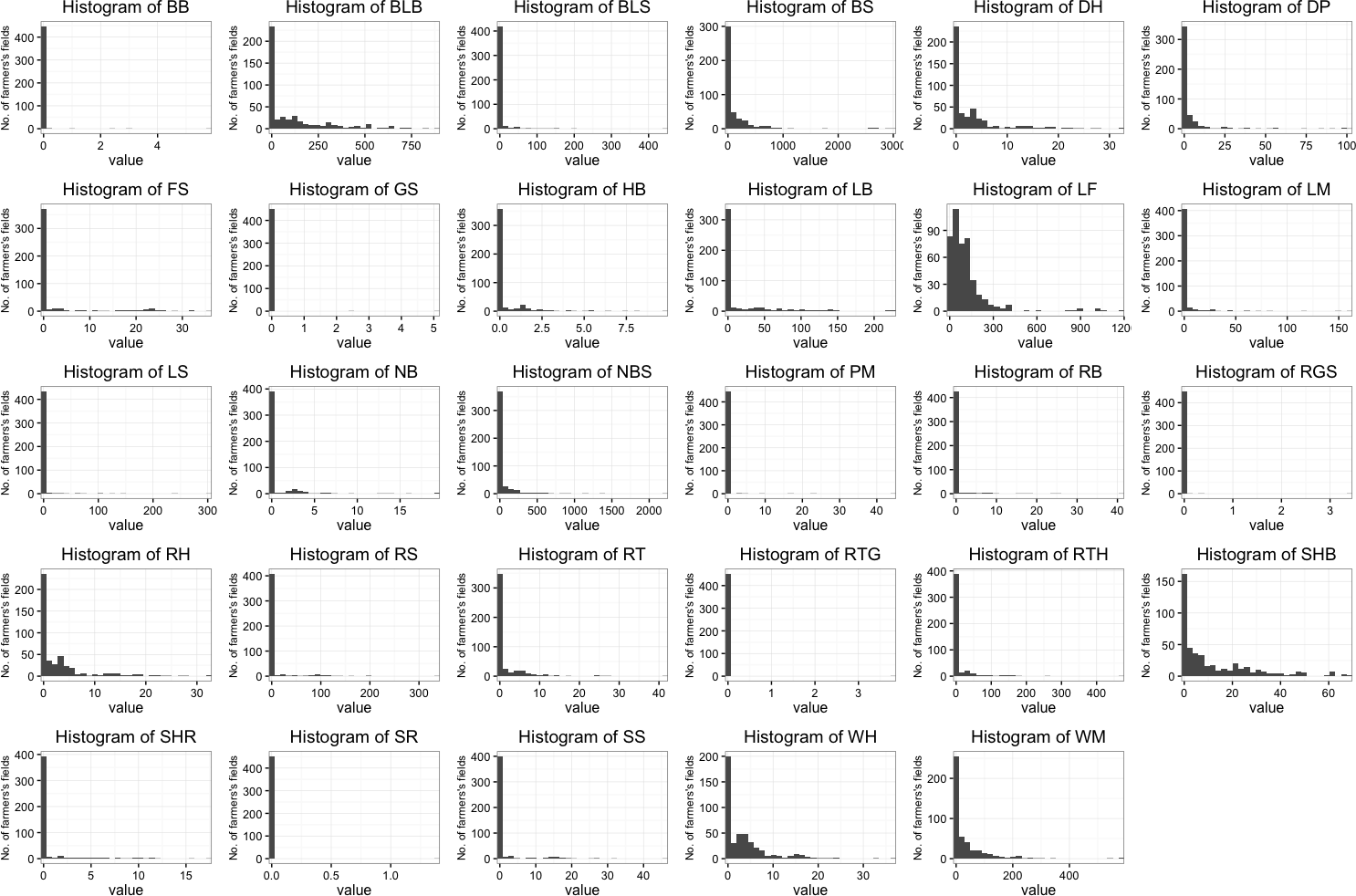


Figure 3-2: Histograms showing the distribution of rice injury values in crop health survey data. BB:Bug burn, BLB:Bacterial leaf blight, BLS: Bacterial leaf streak, BS:Brown spot, DH:Deadheart, DP: Dirty panicle, FS:False smut, GS:Gressy stunt, HB:Hopper burn, LB: Leaf blast, LF: Leaffolder injury, LM: Leaf miner injury, LS: :Leaf scald, NB:Neck blast, NBS:Nerrow brown spot, PM: Panicle mite injury, RB: Rice bug injuries, RGS:Ragged stunt, RH:Rice hispa injury, RS: Red stripe, RT:Rat damage, RTG: Tungro, RTH:Rice thrip injury, SHB:Sheath blight, SHR:Sheath rot, SNL:Snail damage, SR:Stem rot, SS:Silver shoot, WH:White head, WM:Whorl maggot injury.

46

Table III.3: Summary statistics of rice injuries in crop health survey data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Injuries | Mean | SD | Median | Min | Max | Skewness | Kurtosis | SE | Shapiro-Wilk test |
| BB | 0.03 | 0.33 | 0.00 | 0.00 | 5.80 | 14.43 | 228.44 | 0.02 | \*\*\* |
| BLB | 113.64 | 178.34 | 0.00 | 0.00 | 886.67 | 1.86 | 2.93 | 8.38 | \*\*\* |
| BLS | 4.94 | 28.26 | 0.00 | 0.00 | 444.48 | 10.35 | 138.48 | 1.33 | \*\*\* |
| BS | 147.54 | 386.19 | 0.00 | 0.00 | 2999.42 | 5.09 | 30.22 | 18.14 | \*\*\* |
| DH | 3.18 | 5.62 | 0.00 | 0.00 | 32.33 | 2.49 | 6.50 | 0.26 | \*\*\* |
| DP | 3.54 | 12.39 | 0.00 | 0.00 | 101.62 | 5.70 | 36.17 | 0.58 | \*\*\* |
| FS | 2.41 | 6.61 | 0.00 | 0.00 | 35.74 | 2.90 | 7.35 | 0.31 | \*\*\* |
| GS | 0.02 | 0.26 | 0.00 | 0.00 | 5.10 | 17.38 | 317.16 | 0.01 | \*\*\* |
| HB | 0.41 | 1.13 | 0.00 | 0.00 | 9.80 | 4.18 | 22.02 | 0.05 | \*\*\* |
| LB | 17.06 | 38.50 | 0.00 | 0.00 | 226.21 | 2.79 | 8.43 | 1.81 | \*\*\* |
| LF | 114.48 | 156.86 | 76.74 | 0.00 | 1180.29 | 3.94 | 18.95 | 7.37 | \*\*\* |
| LM | 2.93 | 14.33 | 0.00 | 0.00 | 160.47 | 7.63 | 68.27 | 0.67 | \*\*\* |
| LS | 3.13 | 22.06 | 0.00 | 0.00 | 302.08 | 9.82 | 110.91 | 1.04 | \*\*\* |
| NB | 0.60 | 2.14 | 0.00 | 0.00 | 19.32 | 5.69 | 38.72 | 0.10 | \*\*\* |
| NBS | 53.98 | 178.50 | 0.00 | 0.00 | 2213.54 | 6.49 | 58.95 | 8.39 | \*\*\* |
| PM | 0.23 | 2.51 | 0.00 | 0.00 | 44.42 | 14.35 | 229.31 | 0.12 | \*\*\* |
| RB | 0.53 | 3.08 | 0.00 | 0.00 | 40.98 | 8.52 | 87.20 | 0.14 | \*\*\* |
| RGS | 0.01 | 0.16 | 0.00 | 0.00 | 3.40 | 21.02 | 445.24 | 0.01 | \*\*\* |
| RH | 3.18 | 5.62 | 0.00 | 0.00 | 32.33 | 2.49 | 6.50 | 0.26 | \*\*\* |
| RS | 8.14 | 31.13 | 0.00 | 0.00 | 336.71 | 5.32 | 37.10 | 1.46 | \*\*\* |
| RT | 1.47 | 4.02 | 0.00 | 0.00 | 41.56 | 4.90 | 32.92 | 0.19 | \*\*\* |
| RTG | 0.01 | 0.18 | 0.00 | 0.00 | 3.80 | 21.28 | 453.00 | 0.01 | \*\*\* |
| RTH | 9.01 | 34.69 | 0.00 | 0.00 | 470.55 | 7.52 | 79.09 | 1.63 | \*\*\* |
| SHB | 11.67 | 15.49 | 4.72 | 0.00 | 68.65 | 1.60 | 2.08 | 0.73 | \*\*\* |
| SHR | 0.66 | 2.23 | 0.00 | 0.00 | 17.41 | 4.24 | 19.76 | 0.10 | \*\*\* |
| SR | 0.00 | 0.07 | 0.00 | 0.00 | 1.39 | 21.28 | 453.00 | 0.00 | \*\*\* |
| SS | 1.40 | 4.91 | 0.00 | 0.00 | 46.01 | 4.47 | 24.33 | 0.23 | \*\*\* |
| WH | 3.53 | 5.18 | 1.67 | 0.00 | 36.64 | 2.36 | 7.45 | 0.24 | \*\*\* |
| WM | 36.72 | 71.63 | 0.00 | 0.00 | 583.41 | 3.84 | 21.16 | 3.37 | \*\*\* |

The summary statistics, and the result of Shapiro-Wick test of each injury were calculated and summarized in Table III.3. As can be seen, from the previous observa- tions 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. According to Doane and Seward (2011), 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 dis- tributed.

**Comparing correlation coefficients of rice injuries from four correlation methods**

I performed pair-wise analysis between each of injuries using all four correla- tion methods (Pearson, Spearman, Kendall correlation, and Biweight mid-correlation). I examined the similarity of correlation coefficients and clustered according to hierar- chical clustering using Euclidean distance. The result was shown in Figure 3-3. Two groups of correlation methods. The first is parametric correlation measures (Pearson correlation and Biweight mid-correlation) and the second group is nonparametric cor- relation 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 (Ta- ble 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.

**Identify the most suitable correlation measure**

According to the previous results, the group of parametric correlation measures

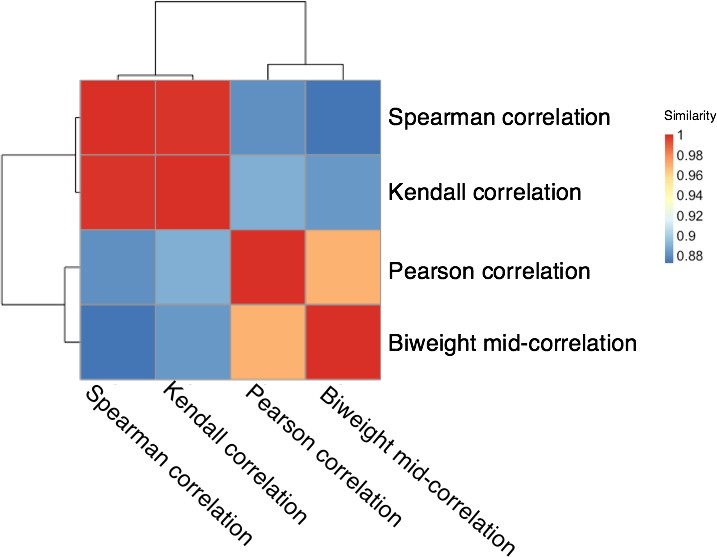


Figure 3-3: Heatmap and dendrogram showing hierarchical unsupervised clustering analysis correlation measures of survey data

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 (Ou, 1985; Barnwal et al., 2013).

**III.1.1 Discussion**

An important criteria to select the suitable correlation measures is to check nor- mality of the variables analysed because an vital assumption in Pearson’s contribution

is the normality of the variables analysed, which could be true only for quantitative vari- ables. Pearson’s correlation coefficient is a measure of the strength of the linear rela- tionship between two such variables. Thus, it is worth to check and test this assumption. Based on visual assessment by histograms, all the variables show left skewed clearly. Skewness and kurtosis values were higher than zero. These indicated that the popu- lations were not non-normal distribution Doane and Seward (2011). Visual inspection of the distribution may be used for assessing normality, although this approach is usu- ally unreliable and does not guarantee that the distribution is normal (Ghasemi et al.,

2012). So normality tests were suggested such as Kolmogorov-Smirnov test, Shapiro- Wilk test. Some researchers recommend the Shapiro-Wilk test as the best choice for testing the normality of data (Peat and Barton, 2005). Shapiro-Wilk test showed that these results were in accord with skewness and kurtosis values.

Evaluation of four correlation measures, including Pearson, Spearman, Kendall correlation, and Biweight mid-correlation by determining similarity of correlation co- efficients from each methods. The results showed two groups clustered according to hierarchical clustering using Euclidean distance. The Spearman and Kendall correla- tion were grouped in rank-base methods, and another group is non-rank-based correla- tions including Pearson correlation and Biweight mid-correlation. From the previous result of testing normality of data, it suggested that the data did not meet the assump- tion of parametric correlation methods, such as Pearson’ method. However, Pearson correlation coefficient is sensitive to outliers. Biweight midcorrelation is considered to be a good alternative to Pearson correlation since it is more robust to outliers(Wilcox,

2012). Among the four correlation methods, Spearman and Kendall are nonparamet- ric rank-based methods. The rank-base methods are nonparametric (distribution-free)

statistics,which uses ranks for correlation and therefore provides a robust measure of a monotonic relationship between two continuous random variables. For this reason, they are particularly suitable for identifying the injuries that increase or decline in monotonic trends in survey data collected during a biological process or developmental stage.

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 (Kumari et al., 2012). The appropriate correlation measures for studied data should closely associate to the prior knowledge of biological correlation. For this study, relationship between dirty panicle and brow spot was detected by Spearman correlation but not by other correlation methods.

**Conclusions**

The analyses I have performed clearly demonstrate the distinct and common performance of four correlation methods. Pearson has been widely used in correlation analyses (Zhang and Horvath, 2005). However, Pearson correlation also limit to be suitable to normally distributed data, and capture the linear relationships. Biweight mid

-correlation is more robust to outlier than Pearson methods, but for this survey data, its output when compared with Pearson methods were not different (in same cluster). The Spearman, Kendall methods performed very well and can capture many relationships in this survey data. Chosen between these methods, Spearman could capture and identify biologically or functionally associated injuries.