Int. J. Advance. Soft Comput. Appl., Vol. 2, No. 2, July 2010
ISSN 2074-8523; Copyright © ICSRS Publication, 2010
www.i-csrs.org

Pest Clustering With Self Organizing Map for Rice Productivity

Shafaatunnur Hasan and Mohd Noor Md Sap

Soft Computing Research Group,
Faculty of Computer Science and Information Systems
University Teknologi Malaysia, 81300 Skudai, Johor
shafaatunnur@gmail.com

Abstract

Rice, Oryza sativa, also known as paddy rice is produced by at least 95 countries around the globe with China and India are the largest producers of rice in the world; while Thailand, Vietnam and America are the largest world rice exporters. To sustain rice productivity, advance agriculture technologies have always been deployed to increase the productivity of this food grain. This is due to the pressure for high productivity and plant pests' attacks. Geographical Information Systems (GIS) and Global Positioning Systems (GPS) have been used for variable rate application of pesticides, herbicide and fertilizers in Precision Agriculture applications. However, due to the weather uncertainties that affect the rice growth, intelligent solutions have been integrated in current pest management practices. Therefore, this study presents intelligent solutions by implementing spatial analysis and Kohonen Self Organizing Map (SOM) to cluster types of pests for better agricultural rice pest management in Malaysia.

Keywords: *Rice, Clustering, Neural Network, Kohonen Self Organizing Map (SOM)*

1 Introduction

Rice cultivation in Peninsular Malaysia is nearly 383000ha in areas scattered over the eleven (11) states. The rice bowl is the area under the Muda Irrigation Scheme

on the north west coast. Here and in the states of Seberang Perai, Perak and Selangor, rice is grown extensively on lowland plains of marine alluvial clay. On the east coast in Kelantan and Terengganu, rice is grown in the less fertile riverine clay soils. Apart from the coastal plains, paddy is cultivated in the flat narrow inland valleys of Melaka, Negeri Sembilan, Pahang and parts of Perak. Depending on the availability of water, rice in Peninsular Malaysia can be classified into 3 categories; the unirrigated, the partially irrigated and the fully irrigated [1].

Parasites, predators and pathogens play a major role in the regulation of rice pests. Most parasites of rice pests belong to the order Hymenoptera and some few to Diptera. Egg parasites (mostly Hymenoptera) play a major role in limiting the growth of rice pests. A similar role, though to a lesser degree, is also played by larval, pupal and adult parasites. Major group of predators such as frogs, birds, and bats play a minor role [2]. Predation need not be confined to rice pests alone; beneficial species, if abundant, may also be attacked. When pray densities are low, spiders, dragonflies and damselflies become cannibalistic. Spiders have been known to eat their own offspring. The erratic feeding habits of predators make the assessment of their economic value difficult. The factors that play a role include: the capacity of the predator to feed and kill, its selectivity in this, but also to the ability to find the prey. The economic value of parasites is more easily determined, because of their more specific behavior. The pathogens that attack insects include nematodes, viruses, bacteria and protozoa [3]. Their importance such as suppressing agents of rice pests has as yet received little attention. In Peninsular Malaysia, several pathogens have recently been identified.

The crop losses in 1979 an extensive outbreak of Surcifera occurred in the Muda Irrigation Scheme causing damage to an estimated 7163 ha, resulting in a loss of (MY) 1.5 million. However, the worst pests of rice which caused considerable damage in almost all paddy fields in Malaysia are rat. Recently, losses at the national production level have been estimated to be around 7%, representing a monetary value of about MYR6.2 million a year [4]. Otherwise, the estimations of overall crop losses due to the rice pests are complicated matter. The infestations differ from location to location and from season to season. In certain years, hardly to mention certain pests that suddenly causes the populations rise without obvious reasons. The pests are sometimes thinly spread over large areas and in other occasions attacks are severe and localized. In all cases, the losses result from an accumulation of damage inflicted by one or a few major pests and many minor species. The species responsible and its share in the damage seem difficult to assess. However useful information about losses can be obtained by combining data from large-scale enquiries, sample surveys and field trials. In Peninsular Malaysia, the losses from insects, birds and rats are estimated to be between 10% and 15% [1].

On the other hand, spatial analysis can be a useful tool to explore the spatial distribution of pests, and help to formulate and test epidemiological hypothesis of pest establishment and spread [5]. The co-occurrence over space of pests and

different aspects of hosts can help farmers and managers understand pest dynamics. In 2001, geostatistical analyses have been implemented to study the spatial variability of the lettuce downy mildew in coastal California [6]. The relatively short disease influence range, which was estimated by a semivariogram, suggested that the role of inoculum availability in the disease epidemics is less important than environmental variables. Furthermore, spatial analysis together with ANOVA analysis have been conducted to study Pierce's disease (caused by the pathogen Xylella fastidiosa) in Temecula Valley, CA vineyards. The results revealed that the proximity to citrus orchards has influenced the incidence and severity of Pierce's disease [7]. This was an important result, guiding potential management strategies for the vector of the disease, the glassy-wing sharpshooter (Homalodisca coagulata). In another study dealing with the same pathogen, but a different crop used semivariograms to map the differing spatial pattern of almond leaf scorch over several different almond cultivars [5]. Their results reported that both random and aggregate patterns of disease spatial distribution and illustrated how cultivar susceptibility influences the distribution patterns of the disease [8].
In [9], spatial analysis using Self Organizing Map (SOM) has been used to estimate the risk of insect species invasion and [10] used cluster analysis SOM in multi-disease diagnosis. The simulation results show that the proposed model performs well and the proposed multi-disease diagnosis is effective. Furthermore, the study on the effectiveness of gradient-based algorithms has been investigated on Rice Yield in Kedah, Malaysia. The results have shown that the gradient descent algorithm has exhibited promising performance compared to Levenberg-Marquardt, quickprop and BP algorithm [11]. The above promising results have motivated us to propose pest clustering using SOM network that will be discussed in the next section.

2 Clustering

Clustering is a data analysis technique that, when applied to a heterogeneous set of data items, produces homogenous subgroups as defined by a given model or measure of similarity or distance. It is an unsupervised process, where its job is to find any undefined or unknown clusters. In supervised learning method, there are some known clusters (groups), from which the algorithms learn the underlying relationship among the inputs and their corresponding outputs. In this way of learning, the model is developed and used for the prediction of target groups for new data elements whose groups are unknown.

For unsupervised scheme, there is no initial input and output relation. However, groups are only predicted from the input data. So, clustering can be thought of as an exploratory data analysis technique that can be used for the selection of diverse compound subsets and data reduction. Clustering as a methodology for partitioning of various types of datasets has been in use in almost all fields of social and technical sciences. However, the clustering tasks in research includes as

a dimension reduction tool when a data set has hundreds of attributes and for gene expression clustering, where very large quantities of genes may exhibit similar behavior. Clustering is often performed as preliminary step in a data mining process, with the resulting clusters being used as further inputs into a different technique downstream, such as neural networks. Due to the enormous size of many present-day databases, it is often helpful to apply clustering analysis first, to reduce the search space for the downstream algorithms. Cluster analysis encounters many of the same issues in the classification. Researchers need to determine the similarity measure, recode categorical variables, standardize or normalize numerical variables and define the number of clusters [12].

3 Kohonen Self Organizing Map (SOM)

Kohonen networks were introduced in 1982 by Finnish researcher Tuevo Kohonen. Although applied initially to image and sound analysis, Kohonen networks are an effective mechanism for clustering analysis. Kohonen networks represent a type of Self Organizing map (SOM), which itself represents a special class of neural network.

The goal of SOM is to convert a high dimensional input signal into a simpler low dimensional discrete. Thus, SOMs are nicely appropriate for cluster analysis, where underlying hidden patterns among records and fields are sought. SOM's structure the output nodes into cluster of nodes, where nodes in closer proximity are more similar to each other than to other nodes that are farther apart. Ritter had shown that SOMs represent a nonlinear generalization of principal component analysis, another dimension-reduction technique.

Self Organization Map are based on competitive learning, where the output nodes competes among themselves to be winning node (or neuron), the only node to be activated by a particular input observation. As [13] describes it: "The neurons become selectively tuned to various input patterns (stimuli) or classes of input patterns in a course of competitive learning process". A typical SOM architecture is shown in figure 1. The input layer is shown at the bottom of the figure, with one input node for each field. Just as with neural networks, these input nodes do no processing themselves but simply pass the field input values along downstream

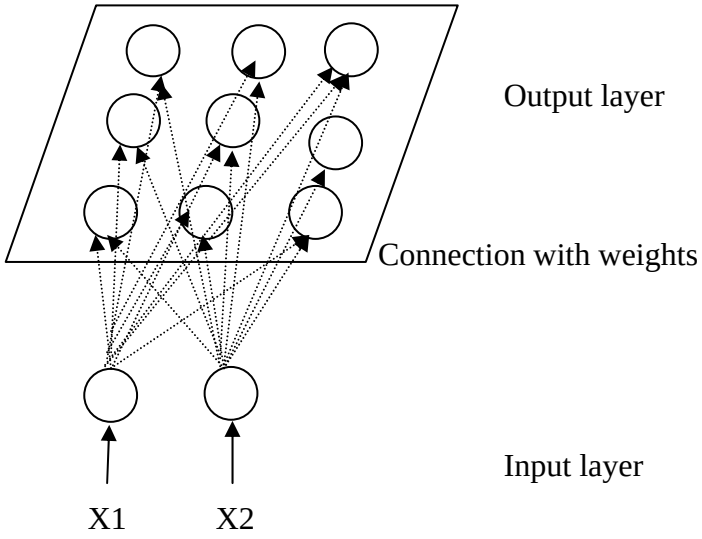

Fig. 1: SOM Architecture

SOM networks are feed forward and completely connected. Feed forward networks do not allow looping or cycling. The network is completely connected to every node in a given layer to every node in the next layer. Similar to neural networks, each connection between nodes has association with it, which at initialization is assigned randomly to a value between zero and one. Adjusting these weights represent the key for the learning mechanism in both neural networks and SOM. Variable values need to be normalized or standardized, just for neural networks, so that certain variables do not overwhelm others in the learning algorithm.

Unlike most neural network models, SOM networks have no hidden layer. The data from the input layer is passed along directly to the output layer. The output layer is represented in the form of a lattice, usually in one or two dimension, and typically in the shape of a rectangle, although other shapes such as hexagons may be used. The output layer shows in figure 1 is a 3x3 square. Finally, SOM exhibit three characteristic processes which is competition, cooperation and adaptation.

3.1 Competition

The output nodes compete with each other to produce the best value for a particular scoring function, most commonly the Euclidean distance. In this case, the output node that has smallest Euclidean distance between the field inputs and the connection weights would be declared the winner.

3.2 Cooperative

The winning node therefore becomes the centre of a neighborhood of excited neurons. This emulates the behavior of human neurons, which are sensitive to the output of other neurons in their immediate neighborhood. In SOMs, all the nodes in the neighborhood share the adaptation given by the winning nodes. They tend to share common features, due to neighborliness parameter, even though the nodes in the output layer are not connected directly.

3.3 Adaptation

In the learning process, the nodes in the neighborhood of the winning node participate in adaptation. The weights of these nodes are adjusted so as to further improve in the score function. For a similar set of field values, these nodes will thereby have an increased chance of winning the competition once again.

SOM Network's Algorithm:

For each input vector x, do:

a) Initialization
Set initial synaptic weights to small random values, say in a interval [0,1], and assign a small positive value to the learning rate parameter α.

b) Competition.
For each output node j, calculate the value $D(w_j, x_n)$ of the scoring function. For example, for Euclidean distance, $D(w_j, x_n) = \sqrt{\sum_i (w_{ij} - x_{ni})^2}$.
Find the winning node j that minimizes $D(w_j, x_n)$ overall output nodes.

c) Cooperation.
Identify all output nodes j within the neighborhood of j defined by the neighborhood size R. For these nodes, do the following for all input records fields.

d) Adaptation
Adjust the weights:

$$W_{ij}new = W_{ij}current + \eta(x_{ni} - w_{ij}current)$$

Standard competitive learning rule (Haykin, 1999) defines ΔW_{ij} applied to synaptic weight w_{ij} as

$$\Delta W_{ij} = \begin{cases} \alpha\left(x_i - w_{ij}\right) & \text{if neuron } j \text{ wins the competition} \\ 0 & \text{if neuron } j \text{ loses the competition} \end{cases},$$

where x_i is the input signal and α is the learning parameter. The learning rate parameter lies in the range between 0 and 1.

e) Iteration
Adjust the learning rate and neighborhood size, as needed until no change occurs in the feature map. Repeat to step (b) and stop when the termination criteria are met.

4 SOM Training and Clustering

The SOM consists of a regular, usually two-dimensional (2D), grid of map units. Each unit *i* is represented by a prototype vector $m_i = [m_{i1},\ldots,m_{id}]$, where *d* is input vector dimension. The units are connected to adjacent ones by a neighborhood relation. The number of map units, which typically varies from a few dozen up to several thousand, determines the accuracy and generalization capability of the SOM. During training, the SOM forms elastic net that folds onto the cloud formed by the input data. Data points lying near each other in the input space are mapped onto nearby map units. Thus, the SOM can be interpreted as a topology preserving mapping from input space onto the 2-D grid of map units.

The SOM is trained iteratively. At each training step, a sample vector *x* is randomly chosen from the input dataset. Distances between x and all the prototype vectors are computed. The best matching unit (BMU), which is denoted here by *b*, is the map unit with prototype closest to *x*

$$\|x-m_i\|=\min\{\|x-m_i\|\} \quad (1)$$

Next, the prototype vectors are updated. The BMU and its topological neighbors are moved closer to the input vector in the input space. The update rule for the prototype vector of unit *i* is

$$M_i(t+1)=m_i(t)+\alpha(t)h_{bi}(t)[x-m_i(t)] \quad , \quad (2)$$

where

t equals time

$\alpha(t)$ is an adaptation coefficient

$h_{bi}(t)$ is neighborhood kernel centered on the winner unit

$$h_{bi}(t) = \exp\left(-\|r_b - r_i\|^2\right) \Big/ 2\sigma^2(t) \quad , \quad (3)$$

where r_b and r_i are positions of neuron b and i on the SOM grid. Both $\alpha(t)$ and $\sigma(t)$ decrease monotonically with time. There is also a batch version of the algorithm where the adaptation coefficient is not used.

In the case of a discrete data set and fixed neighborhood kernel, the error function of SOM can be shown to be

$$E = \sum_{i=1}^{N} \sum_{j=1}^{M} h_{bi} \left\| x_i - m_j \right\|^2 \quad , \qquad (4)$$

where N is number of training samples, and M is the number of map units. Neighborhood kernel h_{bj} is centered at unit b, which is the BMU of vector x_i and evaluated for unit j. If neighborhood kernel value is one for the BMU and zero elsewhere, then SOM reduces to adaptive k-means algorithm. If this is not the case, from equation (4), it follows the prototype vectors that are not in the centroid of their Voronoi sets but are local averages of all vectors in the dataset weighted by neighborhood function values.

A SOM was trained using the sequential training algorithm for each data set. All maps were linearly initialized in the subspace spanned by the two eigenvectors with greatest eigenvalues computed from the training data. The maps were trained in two phases: a rough training with large initial neighborhood width and learning rate and fine-tuning phase with small initial neighborhood width and learning rate. The neighborhood width decreased linearly to 1 with Gaussian function. The training length of the two phases was set to 3 and 10 epochs, and the initial learning rate decreased linearly to zero during the training.

5 Data Preparation

In this study, we used similar data as [11]. These data were collected from Muda Agricultural Development Authority (MADA), Kedah, Malaysia from 1996 to 1998 with 4 areas and 27 locations. With two planting season for each year, a total of 6 seasons were generated. There are 35 parameters that affected the rice yield. These parameters are classified to 5 groups. These include: three types of weed which are *rumpai*, *rusiga* and *daun lebar*; Three types of pests: rats, type of worms and *bena perang*; Three types of diseases: bacteria (*blb & bls*), *jalur daun merah* (*jdm*) *and hawar seludang*; one type of lodging and one type of wind paddy. From 35 parameters, only 11 parameters are chosen since these are the most significant features as recommended by the domain expert from MADA.

6 Experimental Results and Analysis

In this study, SOM network with 2 Dimensional and 10x10 lattice square neuron is applied with 27 observations, 11 variables, 10 neurons, 1000 times learning cycle with learning parameter from 0.9 to 0.1 and Gaussian Neighborhood as percentage map width start from 50 and reducing to 1. In this experiment, the learning parameter and Gaussian Neighborhood used Exponential Decay to shrink SOM's lattice structure. Table 1 and Figure 2 illustrate the clusters of rice parameters that have affected the rice yield by location. There are 4 locations involved and these include A(A1 to E1), B(A2 to I2), C(A3 to F3) and D(A4 to G4). M196 to M298 are the season starting from 1996 to 1998. In season 1 for the year 1996, the parameter has affected the rice yield for type of pests in most of the location A and type of weeds in location D. In season 2 with the same year, location B mostly infected by type of weeds and location D is bacteria (*blb and bls*). However, season 1 and 2, in 1997 and 1998; show that all locations are mostly infected by types of pests and weeds. These results have proven that pests such as rats, type of worms and *bena perang* are one of the factors that have affected the rice production in MADA. The next experiment provides the specific type of pests for further analysis.

Table 1: Location of each cluster

LOCATION	M196	M296	M197	M297	M198	M298
1(A1)	4	12	3	4	3	4
2(B1)	6	7	7	9	7	1
3(C1)	8	10	7	9	7	7
4(D1)	4	5	3	4	3	4
5(E1)	1	4	8	6	7	7
6(A2)	3	6	4	5	4	5
7(B2)	3	9	6	7	8	10
8(C2)	2	9	6	7	5	6
9(D2)	6	6	4	5	4	5
10(E2)	3	1	6	7	5	6
11(F2)	1	9	6	7	5	6
12(G2)	4	12	3	4	3	4
13(H2)	6	7	3	4	3	4
14(I2)	3	3	1	2	9	9
15(A3)	4	12	3	4	3	4
16(B3)	4	5	5	10	6	3
17(C3)	6	12	7	9	7	7
18(D3)	1	10	3	4	3	4
19(E3)	4	5	2	4	3	4
20(F3)	5	6	4	5	8	8
21(A4)	4	5	3	4	3	4
22(B4)	7	2	8	8	2	3
23(C4)	4	8	2	3	3	2
24(D4)	6	6	4	5	4	5
25(E4)	7	6	4	5	1	5
26(F4)	7	2	4	5	4	5
27(G4)	7	11	1	1	6	3

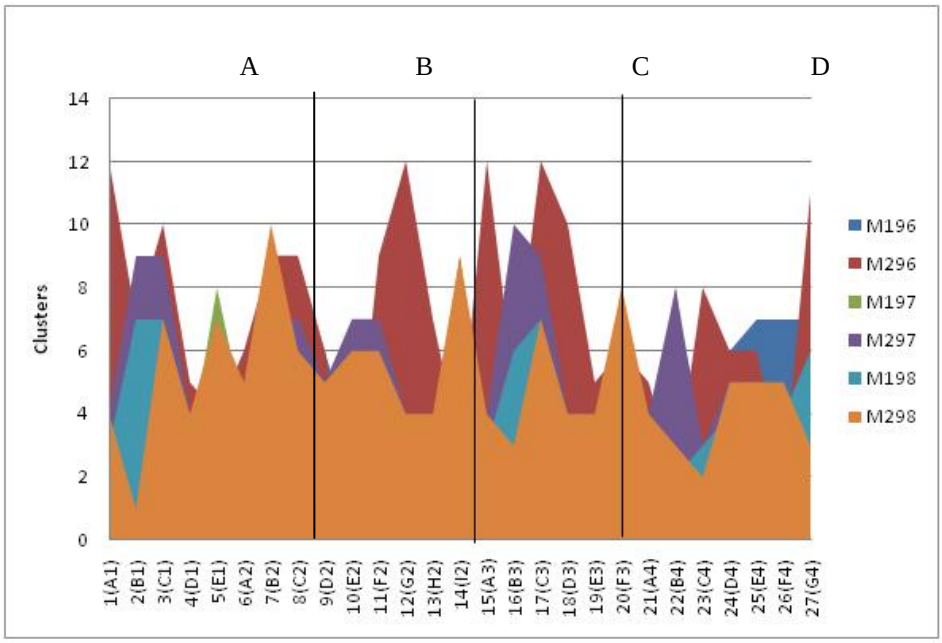

Fig. 2: Location of each cluster

Figure 3, Figure 4 and Figure 5 give the Cluster Means of rice pests. There are 3 types of pests parameter involves in this study: BP for *bena perang*, ULT for type of worms and RAT for mouse. For most of the seasons, ULT yields high range of Cluster Means. For season 2 of each year, BP takes part in the high range of the Cluster Means, while RAT is the lowest rate of Cluster Means with not more than 200.

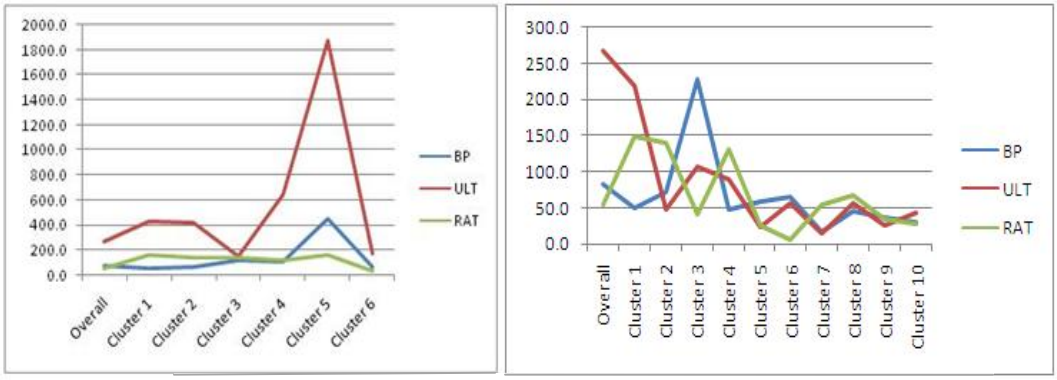

Fig. 3: Cluster Means for Season1 and Season 2 in 1996

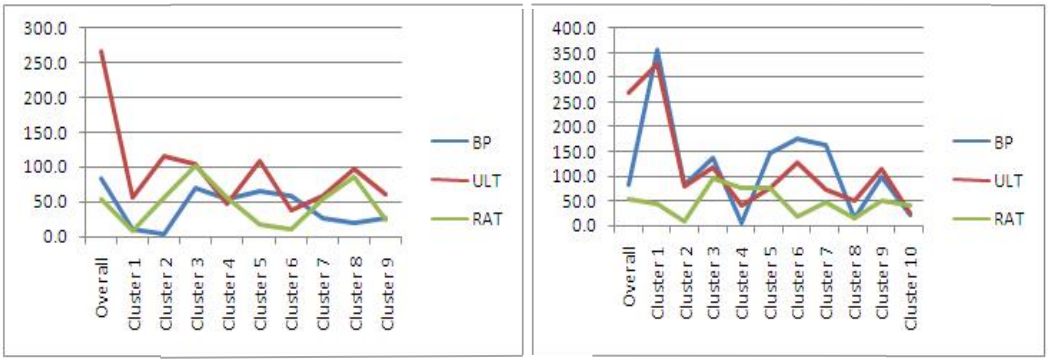

Fig. 4: Cluster Means for Season1 and Season 2 in 1997

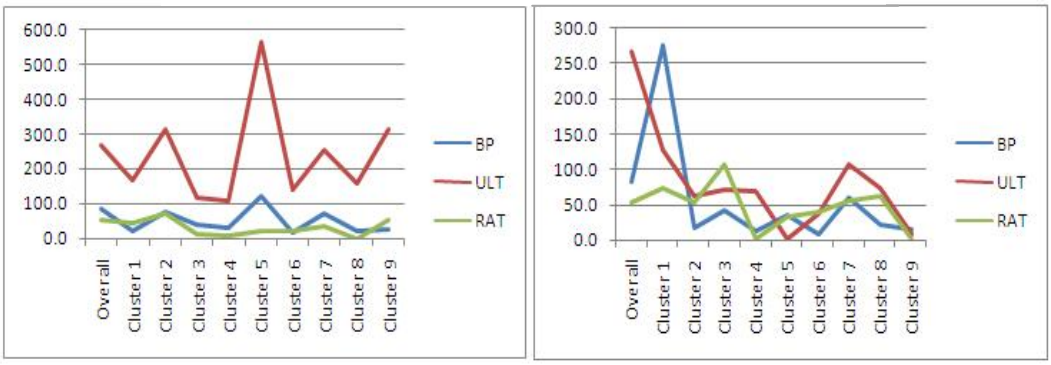

Fig. 5: Cluster Means for Season1 and Season 2 in 1998

6 Conclusion

Pests and weeds are the major factor of the rice yield losses in Malaysia. Hence, intelligent solutions are needed to mitigate the issues of rice productivity. As such, intelligent clustering that is based on SOM network has been successfully applied in spatial analysis for Integrated Pest Management (IPM). Future work will focus on complex spatial data and other machine learning tools for rice yield prediction.

ACKNOWLEDGEMENT

This research was supported by the Research Management Center, University Technology Malaysia (UTM) and the Malaysian Ministry of Science, Technology and Innovation (MOSTI) under vote number 79094. Authors would like to thank Muda Agricultural Development Authority (MADA), Kedah, Malaysia for their support in making this study a success.

References

[1] MARDI, "Manual Penanaman Padi Berhasil Tinggi Edisi 1/2001", Malaysian Agriculture research and Development Institute. Malaysian Ministry of Agriculture (2002).

[2] G.V Vreden,Abdul Latif Ahmad Zabidi. "Pest of rice and their natural enemies in Peninsular Malaysia". Pudoc Wagenigen (1986).

[3] Bongiovanni, R., & Lowenberg-DeBoer, J. "Precision agriculture and sustainability". *Precision Agriculture*, 5 , (2004), 359-387.

[4] Beeri, O., & Peled, A. "Spectral indices for precise agriculture monitoring". *International Journal of Remote Sensing*, 27, (2006), 2039-2047.

[5] Groves, R. L., Chen, J., Civerolo, E. L., Freeman, M. W., & Viveros, M. A. "Spatial analysis of almond leaf scorch disease in the San Joaquin Valley of California: factors affecting pathogen distribution and spread". *Plant Disease*, 89, (2005), 581-589.

[6] Wu, B. M., Bruggen, A. H. C. V., Subbarao, K. V., & Pennings, G. G. H. "Spatial analysis of lettuce downy mildew using geostatistics and geographic information systems". *Phytopathology*, 91, (2001),134-142.

[7] Perring, T. M., Farrar, C. A., & Blua, M. J. "Proximity to citrus influences Pierce's disease in the Temecula valley". *California Agriculture*, 55, (2001), 13-18.

[8] Kelly, M. and Guo,Q. "Integrated Agricultural Pest Management Through Remote Sensing and Spatial Analysis",*Remote Sensing and Integrated Pest Management*, (2007), 191-207.

[9] Watts, M.J and Worner, S.P. "Estimating the risk of insect species invasion: Kohonen self-organising maps versus k-means clustering", Ecological Modelling, 220 (6), (2009), 821-829

[10] Zhang,K., Chai, Y. and Yang, S.X. "Self-organizing feature map for cluster analysis in multi-disease diagnosis"**,** *Expert Systems with Applications,* 37(9), (2010),6359-6367.

[11] Saad,P., Jamaludin, N.K, Rusli,N, Bakri, A. and Kamarudin, S.S. "Rice Yield Prediction- A Comparison between Enhanced Back Propagation Learning Algorithm", *Journal of Information Technology*,(2004).

[12] Jain, A. K., Murty, M.N. and Flynn, P. J. "Data Clustering : A Review". *ACM Computing Surveys,* Vol. 31 (3), (1999), 264-362.

[13] Haykin, S., *Neural Networks : A Comprehensive Foundation*. (2nd ed). Upper Saddle River, New Jersey, (1999).