Self-Organized Maps based Spectral Prediction of *Rotylenchulus reniformis* numbers

Rushabh A. Doshi[*] and Roger L. King[**]

GeoResources Institute,
Mississippi State University,
Mississippi State, MS-39762, USA
[*] rad82@msstate.edu and [**] rking@engr.msstate.edu

Gary W. Lawrence

Department of Entomology and Plant Pathology
Mississippi State University,
Mississippi State, MS-39762, USA
glawrence@entomolgoy.msstate.edu

Abstract— Rotylenchulus reniformis **nematodes present in the soil are one of the major nematode parasite species significantly affecting the growth and development of cotton plants. Recent studies have shown that the nematode numbers in the plant's rhizosphere has direct impact on the reflectance of the plants. In this paper, authors utilize this correlation in developing a field worthy methodology for predicting nematode population number extant in the plant's rhizosphere from variable plant's reflectance. To accomplish this task, a supervised Self-organized map (SOM) was trained using the hyperspectral data signatures of cotton plants affected by different known nematode numbers. The hyperspectral signatures used for training were collected from the cotton plants grown in controlled environment. Twelve field samples (uncontrolled environment) with known nematode numbers obtained from lab analysis of the soil were presented to the supervised trained Self-Organized Map. The location of the sample on the labeled supervised-SOM was used to determine the estimated nematode population of the field sample. In addition to the map grid, the locations of the samples were also visualized using U-matrix, to determine whether the samples were not corrupt or located in the junk part of the map. In addition to the primary goal, hyperspectral signatures of both training and testing data were divided into three sub-regions: Visible region, NIR region and Mid-IR region to observe whether any particular region was the most effective in predicting nematode population.**

Keywords- Hyperspectral, Nematode, Rotylenchulus reniformis, Self-Organized Maps

I. INTRODUCTION

The reniform nematode *(Rotylenchulus reniformis)* has become one of the most prevalent species affecting cotton crops throughout the sub-tropical and tropical regions of the world. Females of *R. reniformis* penetrate the root cells of the cotton plants and establish a feeding site, altering the flow of the nutrients from the plants. This reduction of nutrients causes severe damage to the development of the cotton plants, causing yield loss up to 60%. It is theorized that the higher the nematode numbers present in the soil, the more the damage it causes to the cotton plants. Hence early detection of nematode numbers is absolutely vital in preventing further damage to the cotton plants [1-2]. In 1975, Gausman found that there was significant change in the reflectance of the cotton plants affected by *R. reniformis* numbers as compared to the spectral reflectance of non-nematode affected cotton plants [3]. Kelley suggested that there exists a correlation between the spectral reflectance of the cotton plants with nematode numbers present in the soil [1-4]. Current practice is to count nematode numbers directly from the soil. This requires cotton producers to collect numerous soil samples from various parts of the field and send them to a diagnostic lab for detection and enumeration of the nematode numbers. This labor-intensive procedure is not only inefficient but also time-consuming and costly. Additionally, nematode numbers are not uniformly distributed throughout the field. In this paper, the authors make use of the change in spectral reflectance and supervised-SOM in prediction of estimated nematode population numbers extant in plant's rhizosphere. The authors also investigate whether any particular region in EM spectrum was more effective for prediction of estimated nematode numbers.

Section II discusses the data collection procedure. Section III briefly introduces SOM and its supervised version. Section IV explains the methodology of using supervised-SOM in prediction of nematode numbers. Section V summarizes the results and appropriate conclusions. Finally, Section VI briefly discusses the future work for the given method.

II. DATA COLLECTION

A. Traning Data:

For the year 2001, cotton plants were grown in microplots, small fiberglass cylinders, located in the R.R. Foil North Plant Science Research North Farm located at Mississippi State University. There were 25 microplots filled with similar type of soil. During cottonseed planting, each microplot is artificially infested with five initial population levels. These initial population levels are: 0, 500, 1000, 1500, and 2000 nematodes per 100cc of soil. Hence, there will be 5 microplots with each population levels. Each microplot was sprayed with similar insecticides and was provided with similar nutrients. The hyperspectral data (signatures) were collected from the cotton leaves (single leaf) of each microplot using hand-held Analytical Spectroradiometer (ASD). During each spectral reading, corresponding soil samples from the microplot were also taken for the lab analysis. During the lab analysis, the soil samples were analyzed to count the actual nematode numbers. Based on the nematode numbers present in the soil sample, actual nematode population numbers extant in the plant's rhizosphere was estimated. Microplots were inspected

throughout the growing season. Hyperspectral signatures along with their soil samples were collected at regular intervals. These dates include 19th June (33 Days after plantation (DAP)), 25th June (39 DAP), 10th July (54 DAP), 6th August (81 DAP), and 20th August (95 DAP) [2].

B. Testing Data:

For testing purposes, spectral data along with the soil samples were collected for different dates for the year 2004 and 2005. Data for the year 2004 was collected from the field located at Natchez, Mississippi, while the data for the year 2005 was taken from the field located at Belzoni, Mississippi. The dates for the year 2004 include: 14th July (36 DAP) and 17th August (80 DAP), while the dates for the year 2005 include 22nd June (55 DAP), and 16th August (110 DAP).

The key difference between the microplots and field data was that in microplots the nematodes were artificially infested, while nematodes in the field were naturally present. Both microplots and the field plants were subjected to grow in natural environment.

III. SELF-ORGANIZED MAPS (SOM)

In 1990, Kohonen proposed an "unsupervised, competitive and self-learning" neural network called Self-Organized maps (SOM), which adapts itself based on presentation of input data [5]. The foremost benefit of SOM is that it identifies the similarities among the high-dimensional input data, classifies them based on their similarities and finally represents them into a self-ordering, organized meaningful manner on a low-dimensional grid [5-9]. Thus, SOM performs all the three functions of pattern recognition, compression and visualization.

Rectangular or Hexagonal two-dimensional grids are usually used to present the output of the SOM. For each d-dimensional input vector, there exists a d-dimensional prototype vector $m_i = [m_{i1}, m_{i2}, m_{i3} .... m_{id}]$, where i is the neurons in the map [5]. During the training process, Euclidean distances are calculated between the randomly selected single input data vector x and the corresponding prototype vectors of the map. The minimum distance between the input vector and its corresponding prototype vector becomes the winning neuron. The winning neuron is widely known as best matching unit (BMU) and is calculated using the formulae given below [5]:

$$BMU_i = \min_i \{\| x - m_i \|\} \qquad (1)$$

The location of the winning neuron for the given input data vector dictates the location of the input vector in the map [5]. The neighboring neurons surrounding the BMU are then moved closer towards or furthered away from it. The process is repeated for entire data set and each input vector is then mapped based on the location of its winning neurons [5, 8-10]. This process of quantizing d-dimensional input vector into a d-dimensional output prototype vector is called as vector quantization and mapping of the d-dimensional output vector onto a low-dimensional grid is called vector projection [5, 9-11].

For this work, supervised-SOM with batch training algorithm is used. In supervised-SOM classification method, an additional binary class vector is tagged with the d-dimensional input training vector (feature). The size of the class vector depends on the number of classes. The value of class vector is either '1' or '0'. The class vector has only one value with '1', while rest of the value of class vector is '0'. The location of value '1' in the class vector indicates the class of the particular input feature vector [9, 12-14]. During the training, both the feature vector and class vector are considered. In other words, the class vector "influences the representation/ordering of the map" but not during the calculation of the best matching unit [15]. This ordering causes the formation of "class-clustering" in the map [8]. According to SOM Toolbox, "The class of each map unit is determined by taking maximum over these added components, and is labeled accordingly" [12].

Once the ordered map using class information is trained and created, the class vector is removed. The supervised class-clustered map is then provided only with the d-dimensional feature vector of test samples for class estimation/prediction. Based on the similarities of features between the training and testing samples, the test sample is placed on a particular map unit. The location of the d-dimensional test sample on the labeled (class) map unit indicates the predicted class of the test sample [8-9, 12].

IV. METHODOLOGY

The microplot data collected on various dates are grouped together into three classes based solely on nematode numbers. Factors such as growth stage and biomass are not taken into consideration during class grouping. Various classes along with their labels used for the analysis are as shown below in Table I. Bands 350-450 nm are removed from the analysis to take into account the large amount of scattering found at these wavelengths and to compensate for sensor and atmospheric effects. A map is created using supervised-SOM classification method with hyperspectral reflectances of various classes as input (feature). Number of map units assigned to each class depends on the number of input samples (signatures) present in each class. Once the class clustered supervised map is created, twelve field samples (signatures) with known nematode population (from lab) were presented to the SOM. Based on the similarity of the sample signature with unknown class and the training signature of the known class with which the map is created, SOM will place the sample on to the corresponding labeled map unit. The class labels indicate the range of nematode numbers assigned to the hyperspectral signatures during training. The location of the sample with unknown range on the labeled map unit determines the range of nematode population in the soil. The idea is that since the map was created based on some sort of change in the hyperspectral reflectances of the three classes, if the same change is seen in the reflectances of the testing samples, it will place the corresponding sample in map unit which showed same change. The twelve samples with their original nematode population range obtained from lab analysis are shown in Table II. These testing samples are selected from various dates (years) and different days of planting. Six samples are selected from the year 2004 and six samples are selected from the year 2005 with

Table I Division of Classes based on nematode population

Class	Population range	Labels
Class 1	0-1500	RAN1
Class 2	1501-4000	RAN2
Class 3	Above 4000	RAN3

Table II Testing samples with original class and their labels

Sample	Original Class	Labels
Sample 1	RAN1	SM01
Sample 2	RAN1	SM02
Sample 3	RAN1	SM03
Sample 4	RAN1	SM04
Sample 5	RAN1	SM05
Sample 6	RAN3	SM06
Sample 7	RAN3	SM07
Sample 8	RAN2	SM08
Sample 9	RAN2	SM09
Sample 10	RAN2	SM10
Sample 11	RAN2	SM11
Sample 12	RAN2	SM12

different DAP. There are two disadvantages of using this method:

1. There is the possibility that some samples in the training data might be corrupt and are labeled incorrectly. If some of the test samples are also corrupt, the test sample will be located on the map unit created by corrupt trained samples. Hence the false range is predicted for corrupt signatures.

2. The testing sample might only be corrupt or may not represent any similarity with the trained data set. The method will still place the sample on the map unit that 'most closely matches' the test sample.

In order to solve the first problem a Unified Distance matrix (also known as U-matrix) is also created based on the distances between the neighboring neurons in the map. The main advantage of a U-matrix is that it color-codes the distances between the neighboring neuron with darker color indication large distances and lighter color indicating small distances [12, 16]. This helps in depicting whether or not the training data have corrupt signatures as the location of these samples will have very high distances compared to the other map unit. If the test sample is located in the map created by corrupt training signatures, then one can easily interpret it as corrupt test sample instead of falsely predicting the range. In order to solve the second problem of the possibility of corrupt testing samples but no corrupt training samples, one more class named "junk" class is added to the remaining classes as mentioned in Table I. By adding this extra class of corrupt signatures, any corrupt test sample present in the testing (predicting) analysis will be located in the junk part of the map. In order to verify the solutions, two corrupt test samples were added to the above-mentioned 12 test samples. In order to test the first solution, some corrupt samples (4 in all) were added to class 1. A junk class of corrupt signatures (6 in all) was added in training to check the second solution.

In addition to the above-mentioned goal, hyperspectral signatures for both training data and test samples were divided into three sub-regions: Visible (451-650 nm), NIR region (651-1300 nm), and Mid-IR region (1301- 2500 nm) [17] to investigate the region most effective in predicting correct nematode numbers. The entire methodology was performed on an expanded version of Null's SOMASDGUI and a Graphical User Interface built on a MATLAB based SOM toolbox (available at www.cis.hut.fi) to process and analyze ASD hyperspectral signatures [8, 18]. For our study, map grid of 13 X 13 was used. Default parameters of SOM toolbox was used for both training and prediction purposes.

V. RESULTS AND CONCLUSION

Table III on the next page summarizes various predicted range in each sub-regions along with the entire spectrum. Column 1 of table III indicates the sample number while the column 2 shows the original/actual range estimated from the lab analysis of the soil. Column 3 indicates the range predicted using supervised-SOM. Fig. 1 shows the supervised-SOM map and U-matrix for predicting various nematode ranges for the testing samples. It was seen that more samples were correctly predicted when the supervised map was created using entire spectrum. It was also seen that prediction in the Visible region was least followed by NIR region and Mid-IR region respectively. For our analysis, it was seen that sample 14 was predicted as range 1 when the map was trained using whole spectrum, however, when visualized U-matrix, as shown in fig. 1, it was easily interpreted that the map unit where sample 14 was placed had very high neighborhood distances. This indicates that both the training and testing samples were corrupt or irrelevant to our problem in prediction purposes. It was also seen that sample 13 was located in the junk class indicating that the sample 13 was corrupt and was predicted accordingly.

VI. FUTURE WORK

Although the prediction results obtained from using hyperspectral data with supervised SOM gave promising result, it is necessary to validate the results with larger data set. If the results are still promising, it is possible to train the SOM using nematode numbers as classes themselves instead of using population ranges as classes. It was also seen that the use of supervised-SOM in predicting nematode numbers by using hyperspectral reflectance's as features provide unique and completely different perspective in calculating nematode numbers.

REFERENCES

[1] D. Blasingame, W. Gazaway, P. Goodell, R. Kemerait, T. Kirkpatrick, S. Konning, G. W. Lawrence, M. McClure, J. Mueller, M.Newman, C. Overstreet, P. Phipps, J. Rich, S. Thomas, T. Wheeler,and A. Wrather, "Cotton nematodes: your hidden enemies," Education Resources. *National Cotton Council*, Memphis, TN. 2002.

[2] A. T. Kelly, "Estimation of population thresholds of plant-parasitic nematodes on cotton using hyperspectral remotely sensed data," *Master's Thesis*, Mississippi State University, 2003.

[3] H. W. Gausman,, C. M. Heald, Jr., and D. E. Escobar, "Effect of Rotylenchulus reniformis on reflectance of cotton plant leaves," *Journal of Nematology*, Vol. 7, pp.368-373, 1975.

[4] G. W. Lawrence, R. L. King, G. R. Ellis, R. Doshi, K. S. Lawrence, J. Caceres, and S. Samson, "Population estimation of the reniform nematode using hyperspectral reflectance data and applications to variable rate nematicide applications," *Proceedings of the National*

Beltwide Cotton Conference, Vol. 1, National Cotton Council, Memphis TN. 2006. Online: www.cotton.org/beltwide/proceedings.

[5] T. Kohonen, "The Self-Organizing Map" in *Proceedings of Institute of Electrical and Electronics Engineers*, Vol. 78, pp. 1464-1479, 1990.

[6] R. L. King **,** C. Ruffin, F. E. LaMastus, D. Shaw, "Classification of Weed Species Using Self-Organizing Maps," *Proceedings of the 2nd International Conference on Geospatial Information in Agriculture and Forestry*, Lake Buena Vista, FL, Erim International, Vol. 2., pp.151-158, 2000.

[7] R. L. King, A. L. Rosenberger, L. L. Kanda, "Artificial Neural Networks and Three-Dimensional Digital Morphology: A Pilot Study," *International Journal of Primatology,* Vol. 76, pp. 303-324, 2005.

[8] T. C. Null II, "Use of self organized maps for feature extraction of hyperspectral data," *Master's Thesis*, Mississippi State University, 2002.

[9] R. A. Doshi, "Self-organizing maps for classification and prediction of nematode populations in cotton," *Master's Thesis*, Mississippi State University, 2006.

[10] Jiang Bin, "Extraction of Spatial Objects from Laser-Scanning data using a clustering technique," 2004. [Online] Available: http://www.isprs.org/istanbul2004/comm3/papers/270.pdf

[11] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, "Self-Organizing Maps in Matlab: The SOM Toolbox.," *Proceedings of Matlab DSP Conference*, Espoo, Finland, pp. 35-40, 1999.

[12] SOM Toolbox Documentation [Online] Available: http://www.cis.hut.fi/projects/somtoolbox/documentation/

[13] M. Hannula, J. KLaitinen, D. Alassarela, "Classification accuracy of the Frequency Analysis Method: Comparision between SOM-Supervised and k-NN," *Proc of the 4th Annual IEEE Conf on Information Technology Applications in Biomedicine*, pp. 254-257, April 2003.

[14] X. Yun-De, A. Clauset, R. Harris, E. Bayram, P. Santago, II, and J. D. Schmitt, "Supervised Self-Organizing Maps in Drug Discovery. 1. Robust Behavior with Overdetermined Data Sets," *J. Chem. Inf. Model*, Vol. 45(6), pp. 1749–1758, (Article) DOI: 10.1021/ci0500839, 2005. [Online] Available: http://pubs.acs.org/cgi-bin/article.cgi/jcisd8/2005/45/i06/pdf/ci0500839.pdf

[15] X. Yun-De, A. Clauset, R. Harris, E. Bayram, P. Santago, II, and J. D. Schmitt, "Supervised Self-Organizing Maps in Drug Discovery. 2. Improvements in Descriptor Selection and Model Validation," *J. Chem. Inf. Model,* Vol. 46 (6), pp. 137–144, 2006. [Online] Available: http://pubs.acs.org/cgi-bin/article.cgi/jcisd8/2006/46/i01/pdf/ci0500841.pdf

[16] A. Ultsch and H. P. Siemon, "Kohonen's self organizing feature maps for exploratory data analysis", in *Proc. INNC'90, int. neural network conf.*, Dordrecht, Netherlands, Kluwer, pp. 305-308, 1990.

[17] T. M. Lillesand and R. W. Kiefer, *Remote Sensing and Image Interpretation*. John Wiley & Sons., Inc, New York.,2000.

[18] SOM Toolbox. [Online] Available: www.cis.hut.fi/somtoolbox.

Table III Prediction chart of various testing samples in various regions of EM spectrum with green-color indicating correctly predicted and red-color indicating falsely predicted

| Sample	Labels	Original Class	Predicted Range in different regions			
			Visible region	NIR region	Mid-IR region	Whole region
Sample 1	SM01	RAN1	RAN2	RAN1	RAN3	RAN1
Sample 2	SM02	RAN1	RAN1	RAN3	RAN1	RAN1
Sample 3	SM03	RAN1	RAN2	RAN2	RAN3	RAN1
Sample 4	SM04	RAN1	RAN1	RAN1	RAN1	RAN1
Sample 5	SM05	RAN1	RAN1	RAN2	RAN1	RAN2
Sample 6	SM06	RAN3	RAN3	RAN2	RAN2	RAN2
Sample 7	SM07	RAN3	RAN1	RAN3	RAN3	RAN3
Sample 8	SM08	RAN2	RAN3	RAN2	RAN2	RAN1
Sample 9	SM09	RAN2	RAN1	RAN2	RAN3	RAN2
Sample 10	SM10	RAN2	RAN3	RAN3	RAN3	RAN1
Sample 11	SM11	RAN2	RAN3	RAN2	RAN1	RAN2
Sample 12	SM12	RAN2	RAN1	RAN2	RAN2	RAN2
Sample 13	SM13	Junk	Junk	Junk	Junk	Junk
Sample 14	SM14	Junk (RAN1)	RAN1	RAN3	RAN3	RAN1
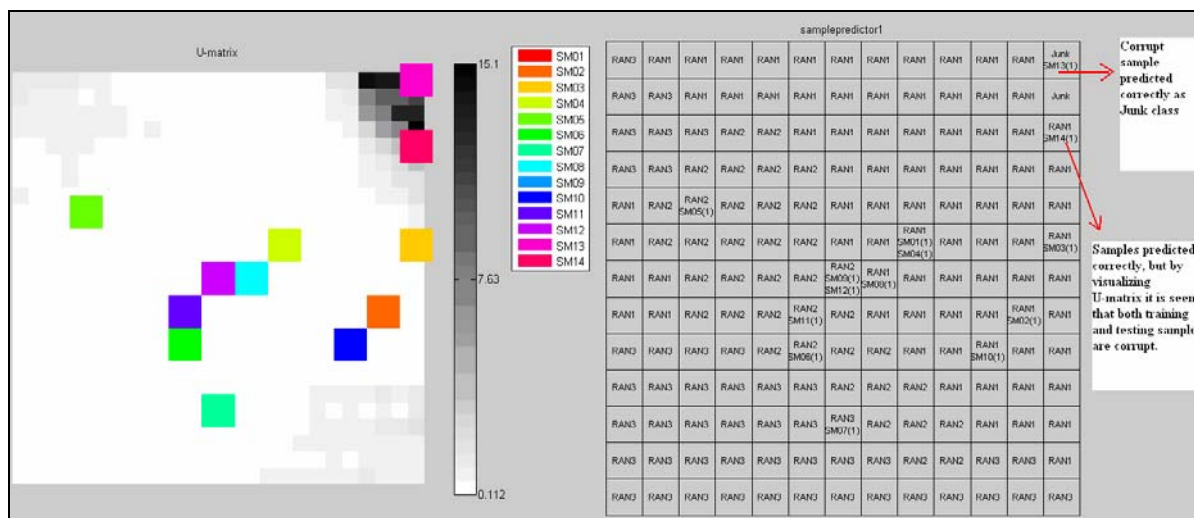
Figure 1 Illustration of the predictor concept (in entire spectrum)