INTEGRATION OF ARTIFICIAL NEURAL NETWORK AND GEOGRAPHIC INFORMATION SYSTEM FOR AGRICULTURAL YIELD PREDICTION

Kanchana Thongboonnak[1*] and Sunya Sarapirome[2]

Received: Nov 23, 2010; Revised: Jan 26, 2011; Accepted: Jan 26, 2011

Abstract

The main objective of this study was to develop the Artificial Neural Network (ANN) modules for agricultural yield prediction as an extension of the ArcMap software. The Object-Oriented methodology was used for both design and programming. The application coding was done in VB.NET. The ANN modules developed were tested with longan yield prediction in Chiang Mai and Lamphun provinces. The ANN input data are soil group and climate data for the years 2006 – 2008, which relate to longan yield in 2007 and 2008. All data were normalized in the same range of 0-1 to be suitable as the input of the ANN model. The normalized weekly highest, lowest, and average temperature, average sunlight, and rainfall were interpolated. They were then averaged to spatially represent districts in the study area, which corresponded to the longan yield districts. These data were varied with several input variations. The cross validation process was applied to each variation. The optimal parameters including learning rate, number of nodes in the hidden layer, and number of iterations obtained from testing were 0.4, 6, and 3,000 respectively. These parameters were applied for all training and testing processes. The best accuracy achieved is 99%. The ANN modules developed for the ArcMap environment worked well for longan yield prediction with accurate results despite the limitations of the data set.

Keywords: Artificial Neural Network (ANN), agricultural yield prediction, Longan, Geographic Information System (GIS)

Introduction

The principle of agricultural yield prediction is to search for techniques or models that identify the functional relationship between influencing factors and production. There are both linear models such as linear regression and non-linear models such as Artificial Neural Network (ANN) and Fuzzy Logic. The weakness of the linear models are low flexibility because they cannot be applied to other areas unless the environmental conditions are similar.

[1] *Faculty of Science, Chiang Mai Rajabhat University, 202 Changpueak Rd., Muang District, Chiang Mai 30000, Thailand. E-mail: kanjana@cmru.ac.th*

[2] *School of Remote Sensing, Suranaree University of Technology, 111 University Avenue, Muang District, Nakhon Ratchasima 30000, Thailand. E-mail: sunyas@sut.ac.th*

* *Corresponding author*

Sudduth *et al*. (1996) found that the linear method generally failed to produce good functional approximations of spatial yield variability. Some adaptive non-linear models have been recently introduced for prediction purposes. There are also many reports that claim to give better efficiency (Malik, *et al*., 1999; Liu *et al*., 2001; Ezrin *et al*., 2009).

ANN was developed to forecast the yield of several types of economical orchard plants. The most important economical orchard crop in Northern Thailand is longan. In the past, longan yield forecasting was handled quite simply by collecting information from agricultural organizations and performing surveys to find out the total orchard area then multiplying this number with a yield. The result could change depending on the precision of the surveying. The yield was taken from the previous year's statistics. Using this method a forecasting mistake could be made and reflected in the planning which could cause economic problems in the longan yield. Therefore, instead of using this conventional method, this research concentrated on using modeling to predict the yield of longan. The unconventional ANN method is considered to provide more efficient and accurate results. In addition, the model deals mainly with spatial data prepared under the GIS environment. It will be more convenient if the module developed can work as an extension in the popular GIS software such as ArcMap.

Concept of ANN Modeling

Artificial Neural Network (ANN) is a massive parallel process that is formed through a combination of neurons that emulates the neuro-synaptic structure of the human brain. A neuron is the smallest computation unit of data. Each neuron is linked to its neighbors with varying coefficients of connectivity. They learn the correlation between the input and the target from examples and then solve the corresponding problems. These examples are called training sets which are entered into the network several times so that the network can learn from them. During the training process the connection weight of the network is adjusted. The training process may be supervised by the set of known target outputs corresponding to the inputs provided for the network (Boonprasom, 2003). Figure 1 shows an example of a multiple-input neuron.

The scalar input p is multiplied by the scalar weight w to wp, one of the terms that is sent to the summation. The other input is multiplied by a bias b and then passed to the summation. The summation output n, often referred to as the net input, goes into a transfer function f which produces the scalar neuron output a.

Often, 1 neuron with many inputs may not be sufficient. We may need more than 1 parallel neuron operation, in what we call a "layer". A network with several layers is called a multilayer network. Each layer has its own weight matrix W, its own bias vector b, a net input vector n, and an output vector a. Also, different layers may be formed by different transfer functions. The layer whose output results in the output of ANN is named an output layer. All layers which are not the

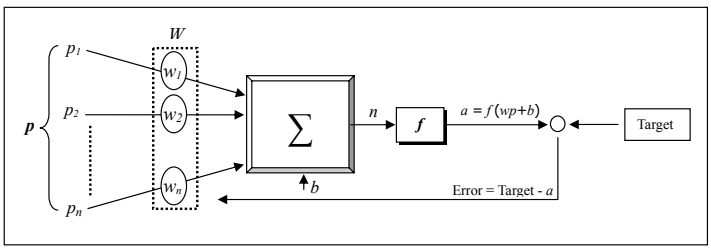

Figure 1. An example of a multiple-input neuron

input layers are called hidden layers.

The best learning method for a multilayer process is the Back-Propagation learning rule which organizes 2 main paths as shown in Figure 2. The first path is called the forward path. In this path, the input vector is applied to the network. The result of the hidden layers is distributed into the output layer. The output vector from the output layer is the actual solution to the network. In this path, the network parameters such as weight and bias are constant. The second path is called the backward path. In this path, all network parameters are changed and regulated. This is accomplished by the error correction rule. An error signal is formed in the output layer. An error vector is equal to the difference between the actual answer and the estimate of the network's answer. The learning rule then adjusts the weights of the network in order to move the network outputs closer to the targets (Hagan *et al.*, 1995). This is the "training process" for the neuron network.

Before using ANN, model training and testing must be done starting from the training process and inputting data into the model until the output is satisfied. After the training process, the weight as well as the knowledge of the neuron network will be discovered. Model testing was done by using data which were not of the same series as the training data set. If the result is satisfactory, the weight can be practically used. Figure 3 shows the steps of neuron network model for longan yield prediction.

Material and Method

Data

Factors involved include temperature, sunlight, rainfall, district boundary, land use, and soil collected from various sources. The characteristics or format of all data collected were examined and further prepared and manipulated to be in the required format and representation that would be effective for the ANN prediction module development.

Temperature, sunlight, and rainfall were collected from 20 Meteorological stations of Hydro and Agro Informatics Institute (HAII) covering 10 districts including Doi Lo, Ban Hong, Doi Tao, Hot, Li, Mae Wang, Chom Thong, Saraphi, San Patong, and Hang Dong. Data from each station were separated into hourly data covering a time span from 2006-2008. Land-use and district data covering the study area were classified by the Land Development Department (LDD) as GIS data layers. Only areas with longan planting were selected and were used together with the soil data layer to obtain the total area extent of the longan planting area of each district and the soil group within each area. Soil data covering the study area were classified by the LDD. Only soil group areas falling into the longan

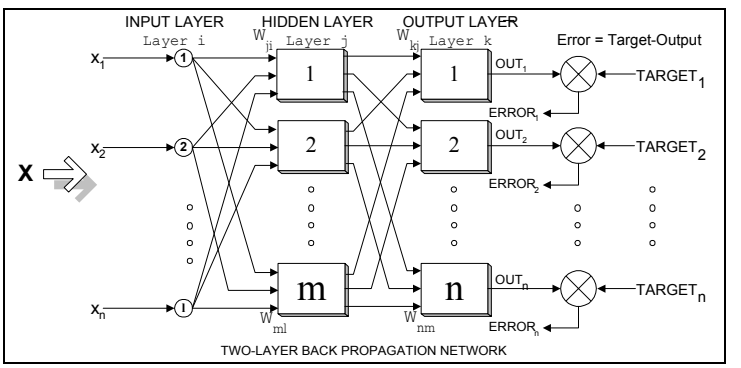

Figure 2. Multilayer Network with Back-Propagation learning rule

planting area were considered for prediction activity. Longan yield data of each district from 2007-2008 within the study area were obtained from Chiang Mai Provincial Agricultural Extension Office and the Office of Agricultural Economics. The unit of yield production is tons/district.

Hourly temperature, sunlight, and rainfall data were formatted into spreadsheet data. Hourly temperature data were prepared into the daily minimum, maximum, and average values and then into weekly minimum, maximum, and average values. Sunlight and rainfall data were prepared into the daily average values and then the weekly values. All weekly data of each HAII station were input through interpolation in order that the average data of each district polygon could be achieved. These data were normalized to be 0 to 1 by using performance of (raw-min)/(max-min). This data was used as input data for ANN training, testing, and prediction analysis.

The extent of soil groups within the longan planting area of each district was obtained from the union of the soil data layer with the selected district data layer and longan planting area from the land-use data layer. Querying was then performed using a condition of the existing longan area and selected district. The area extent of soil groups within each district was calculated to be a percentage. The area extent as a percentage of the soil group of each district was normalized to be 0 to 1 by dividing by 100. The longan planting area of each district was normalized to be 0 to 1 by using the performance of (raw-min)/(max-min).

The longan yield data were tabulated associated with 16 weekly items of temperature, sunlight, and rainfall data. They were normalized to be 0 to 1 by performing (raw-min)/(max-min) when yields of whole districts in the study area were considered. The example of input data is shown in Table 1.

ANN Module Developments

The developed module covered functions of the ANN model and its related data flow such as the input, training, and testing processes. The steps to develop the ANN module are shown in Figure 4.

The analysis step requires comprehension of the whole functionality of the ANN model and its relationship. The general functions of the model are input, feed forward, back propagation, weight adjustment, and error assessment.

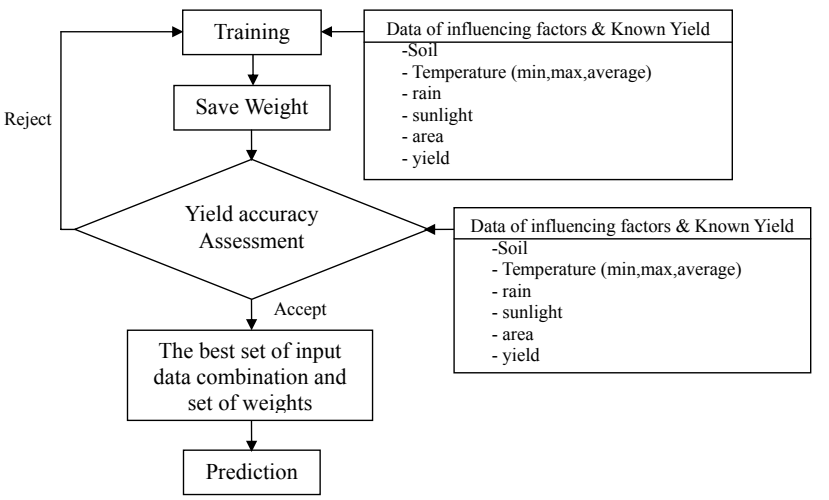

Figure 3. The steps of Neural Network for longan yield prediction

After the analysis and design processes, the module was implemented using the VB.NET language into steps as shown in Figure 5. While coding, the process included configuration to fit the ArcObject, compiling codes to be DLL libraries, and adding the extension module developed as an icon on the ArcMap toolbar (Figure 6).

Training and Testing

Variation of model parameters for cross validations including the number of nodes, learning rate (LR), the number of iterations, and errors acceptance was conducted to obtain the best accuracy for the model. These selected parameters of the model were further used for cross validations with varying input data.

The model parameters were varied for the tests. The results shown in Table 2 reveal that the best LR, hidden nodes, and number of iterations are 0.4, 6, and 3,000 respectively. These were further used in the training and testing of the model.

Based on the studies of Turney (1994), Kohavi (1995), and Lawrence *et al*. (1997),

Table 1. Shows an example of input data

District	Doi Lo	Ban Hong	Doi Tao	Hot	Li	Mae Wang	Chom Thong	Saraphi	San Patong	Hang Dong
TMAXN1	0.5433	0.6817	0.6125	0.4913	0.4602	0.7474	0.7301	0.5260	0.6609	0.7405
TMINN1	0.8162	0.8529	0.8897	0.9632	0.5662	0.8897	0.9265	0.8971	0.8897	0.9191
TAVGN1	0.5315	0.6076	0.6185	0.6078	0.4776	0.6463	0.6930	0.6356	0.5761	0.6798
SAVGN1	0.1600	0.1400	0.1190	0.1800	0.1840	0.3151	0.0665	0.1133	0.9560	0.1889
RAVGN1	0.0000	0.0004	0.0197	0.7125	0.0004	0.8438	0.0109	0.0101	0.0750	0.0750
…	…	…	…	…	…	…	…	…	…	…
TMAXF4	0.7128	0.9689	0.8997	0.8824	0.7474	0.7993	0.8824	0.6782	0.7647	0.6955
TMINF4	0.5294	0.4926	0.7132	0.5662	0.3897	0.5662	0.5294	0.4559	0.4926	0.4191
TAVGF4	0.5342	0.9378	0.8496	0.8206	0.6814	0.7649	0.6604	0.5087	0.6867	0.5716
SAVGF4	0.1680	0.1800	0.1822	0.2391	0.2000	0.3207	0.1930	0.2317	0.2897	0.1740
RAVGF4	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0003	0.0000	0.0000
S30	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0119	0.1229	0.0000	0.0000
S38	0.0109	0.0678	0.0089	0.0695	0.0000	0.0000	0.0489	0.1781	0.0279	0.0002
S22	0.0850	0.0000	0.0000	0.0000	0.0000	0.0444	0.0000	0.0379	0.3250	0.0755
…	…	…	…	…	…	…	…	…	…	…
S29B	0.0000	0.0645	0.0092	0.0000	0.0949	0.0000	0.0000	0.0000	0.0000	0.0000
S40B	0.0000	0.0647	0.0104	0.0125	0.0238	0.0183	0.0000	0.0000	0.0094	0.0000
AREA	0.1455	0.3762	0.0000	0.0111	1.0000	0.0379	0.5609	0.1396	0.1544	0.0114

where *TMAXN1 is maximum temperature of the first week of November*
TMINN1 is minimum temperature of the first week of November
TAVGN1 is average temperature of the first week of November
SAVGN1 is average sunlight of the first week of November
RAVGN1 is average rainfall of the first week of November
RAVGF4 is average rainfall of the fourth week of February
S30– S40B are soil group
AREA is the longan planting area of each district

the performance of the ANN model of this study was evaluated using an n-fold cross-validation technique. The relevant data set was partitioned randomly into n equally sized sets. Training and testing were carried out *n* times. Each time used *1* distinct set for the testing phase and the remaining *n-1* sets for the training phase. The validation results were averaged over the rounds.

Two-year data collection from 10 districts resulted in 20 records which were used as input data for cross validation of the ANN model. Twenty cross validations were performed one at a time by use of 19 records for training and 1 record for testing. In 1 validation, 19 records were input through the ANN model individually.

After training with the 19 records, a set of weights was achieved and applied to 1 record with a known yield. An error was then estimated when the predicted yield from the model was compared with the known yield of

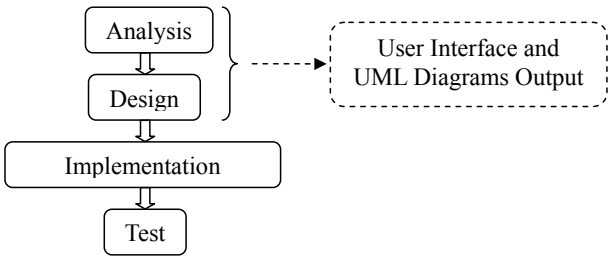

Figure 4. The steps for the module development

Table 2. Results of 12 tests to obtain the best model parameters which are LR, a number of hidden nodes, and a number of iterations

No. of Hidden Nodes	No. of Iterations	LR	% Accuracy of each district										Average % Accuracy
			K.Doi-Lo	Ban-Hong	Doi-Tao	Hot	Saraphi	Li	Mae-Wang	Chom-Thong	Hang-Thong	San-Thong	
4	3000	0.1	91.81	63.88	38.13	84.39	84.92	50.22	97.98	52.29	74.80	84.99	72.34
4	3000	0.2	87.39	90.46	28.84	96.88	71.93	84.96	79.30	50.55	89.75	75.55	75.56
4	3000	0.4	67.79	99.96	73.78	91.30	91.94	86.51	87.95	49.10	79.06	52.76	78.02
4	3000	0.8	68.13	94.19	80.35	63.44	50.93	99.91	67.45	97.09	80.25	41.60	74.33
4	3000	0.4	67.79	99.96	73.78	91.30	91.94	86.51	87.95	49.10	79.06	52.76	78.02
5	3000	0.4	56.86	28.76	51.62	69.18	43.88	98.55	77.69	58.83	60.73	73.68	61.98
6	3000	0.4	78.17	87.14	67.74	75.03	61.85	97.07	90.56	60.32	96.72	86.75	80.13
7	3000	0.4	84.33	83.55	55.17	75.90	83.20	97.23	45.29	61.82	86.98	87.54	76.10
6	3000	0.4	78.17	87.14	67.74	75.03	61.85	97.07	90.56	60.32	96.72	86.75	80.13
6	5000	0.4	84.08	88.37	63.05	68.32	81.35	99.95	74.03	51.98	84.65	63.43	75.92
6	7000	0.4	89.12	29.29	52.71	82.37	64.66	98.05	62.82	80.12	54.25	78.49	69.19
6	1000	0.4	98.61	94.39	67.74	64.12	57.62	97.52	56.11	86.30	69.56	97.92	78.99

the testing record. Twenty errors were reported for 20 cross validations. To this end, the average error indicates the accuracy or efficiency of the model.

Due to the limited amount of research on the physical factors influencing longan production, input data were varied and the results observed. In the present study, a number of input variations were run, as shown in Table 3, and errors from these operations were observed. In each variation, the cross validation was performed and the average

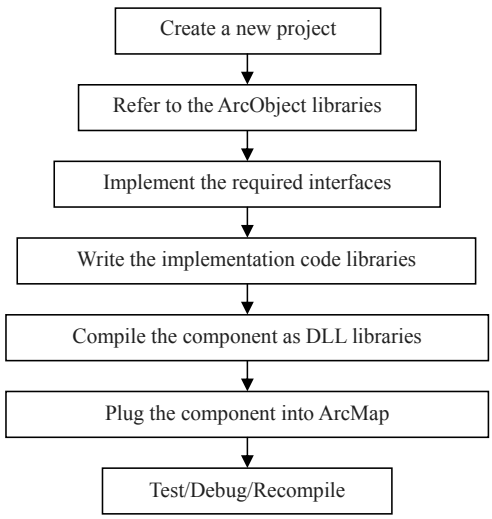

Figure 5. The steps to implement ArcMap extension module using VB.NET

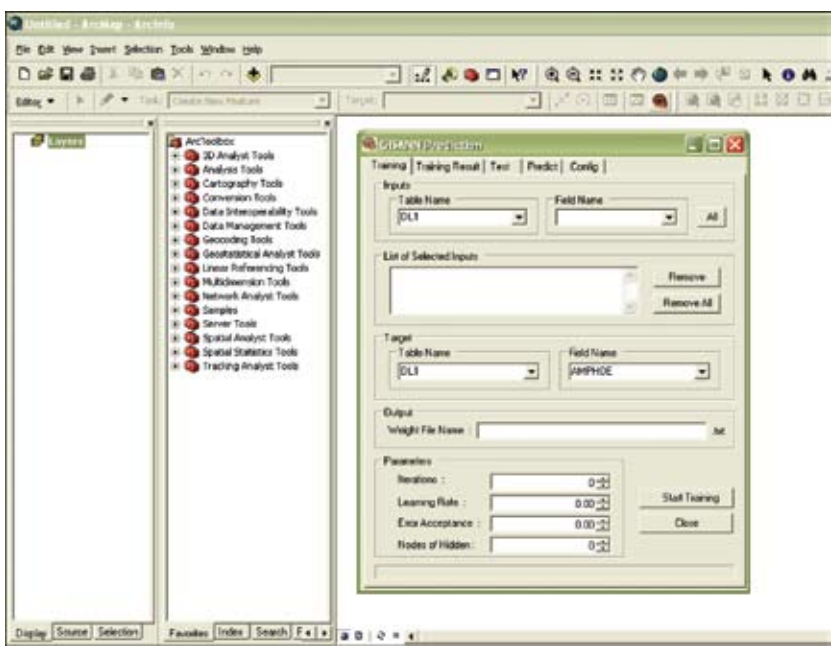

Figure 6. The icon and dialog box of the extension module developed

errors were reported. The variation with the lowest average error or the highest accuracy indicates the best variation for this set of input data and the best set of weights and were used for further predictions.

Result

The ANN prediction model was developed successfully as an extension module for ArcMap software. It can work correctly for model training and testing with data prepared under the software environment. In the study, the results reveal that the temperature and sunlight of 16 weeks from the first week of November to the end of February, soil data, and no rainfall (variation number 3) provided the best accuracy. Therefore, the combination set of input data and set of weights were used for the predictions. Table 4 shows the difference of predicted yield and actual yield for 2 years. Chom Thong is the district showing the best accuracy (99%).

Conclusions

The study results can be concluded as follows:

1) The best accuracy for longan yield prediction using the ANN model developed in the study is 99%. The weight set used was adopted from the best combination set of input data from the tests. The model parameters employed for the tests include LR, hidden nodes, and number of iterations which are respectively 0.4, 6, and 3,000.

2) Considering the physical data used, the average prediction accuracies of the variation sets are between 80-83%. The best accuracy appeared when all input factors were used. Without either sunlight or rain data, results are lower than the others. This might indicate that these data effects are related to one another. However, with both of them, or without soil, there is not much effect on accuracy.

3) Considering the varying number of weekly data, the weekly data of January and February show better accuracy than others in a range of 2-6%. It could be concluded that

Table 3. A number of variations of input data

Variation	Weekly data	Temperature (min, max, ave.)	Sunlight	Rain fall	Soil Group
1	N1 to F4	☑	☑	☑	☑
2	N1 to F4	☑	☑	☑	☑
3	N1 to F4	☑	☑	☑	☑
4	N1 to F4	☑	☑	☑	☑
5	N1 to F4	☑	☑	☑	☑
6	N1 to D4	☑	☑	☑	☑
7	N1 to D4	☑	☑	☑	☑

when *N1 is the first week of November*
N2 is the second week of November
J4 is the fourth week of January
☑ *is selected factor*
☑ *is non-selected factor*

Table 4. Difference of predicted yield and actual yield for two years

District	2007				2008			
	Predicted Yield (tons)	Actual Yield (tons)	Difference	Accuracy (%)	Predicted Yield (tons)	Actual Yield (tons)	Difference	Accuracy (%)
Doi Lo	16076	16997	-921	94.58	17850	16285	1565	90.39
Ban Hong	31339	26831	4508	83.20	21849	18270	3579	80.41
Doi Tao	6573	5416	1157	78.64	6295	6643	-348	94.76
Li	31967	32863	-896	97.27	28671	24074	4597	80.90
Chom Thong	32166	31849	317	99.00	22022	25332	-3310	86.93
Saraphi	15371	16051	-680	95.76	15122	15826	-704	95.55
San Patong	19764	16005	3759	76.51	10444	10061	383	96.19
Hot	7497	8648	-1151	86.69	14027	13006	1021	92.15
Mae Wang	10748	12749	-2001	84.30	7556	9683	-2127	78.03
Hang Dong	11231	9986	1245	87.53	6757	7840	-1083	86.19
Total	182732	177395	5337		150593	147020	3573	

the varying number of weekly data during November–February does not show a high significance on influencing longan production. A higher accuracy can be obtained using more weekly data.

4) The modules developed are sufficiently flexible to be applicable to yield prediction of other crops. Different types of crops could have their own factors influencing yield productivity; if their data can be prepared to meet the requirements of these modules' input format, their yield could be predicted using these modules.

References

Boonprasom, P. (2003). Unpublished data. Yield Prediction of Tangerine Using Artificial Neural Network (ANN). Chiang Mai University, Thailand.

Ezrin, M.H., Amin, M.S.M., Anuar, A.R., and Aimrun, W. (2009). Rice yield prediction using apparent electrical conductivity of paddy soils. European Journal of Scientific Research (EJSR), 37(4):575-590.

Hagan, M.T., Demuth, H.B., and Beale, M. (1995). Neural Networks Design. 1st Edition. PWS Publishing, Boston, MA, USA, 684p.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 95); Aug 20-25, 1995; Montreal, QC, Canada, p. 1137–1143.

Lawrence, S.C., Giles, L., and Tsoi, A.C. (1997). Lessons in neural network training: overfitting may be harder than expected. Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97); July 27-31, 1997; Providence, RI, USA, p. 540–545.

Liu, J., Goering, C.E., and Tian, L. (2001). A neural network for setting target corn yields. American Society of Agricultural Engineers (ASAE), 44(3):705-713.

Malik, R., Hua, G.B., and Barathithasan, T. (1999). A comparative study of artificial neural networks and multiple regression analysis in estimating willingness to pay for urban water supply. Available from: www.buildnet.co.za/cdcproc/docs/1st/ranasinghe_m.pdf. Accessed date: Oct 1, 2010.

Sudduth, K.A., Drummond, S.T., Birrell, S.J., and Kitchen, N.R. (1996). Analysis of spatial factors influencing crop yield. Proceedings of the 3rd International Conference on Precision Agriculture. June 23-26, 1996; Minneapolis, MN, USA, p. 129–140.

Turney, P. (1994). A theory of cross-validation error. J. Exp. Theor. Artif. In., 6:361-391.