Thailand Statistician
January 2011; 9(1) : 51-64
http://statassoc.or.th
Contributed paper

Principal Component-based Modeling Approaches for Predicting Soil Organic Matter

Kamolchanok Panishkan* [a], Natdhera Sanmanee [b] and Sirikanlaya Pramual [c]

[a] Department of Statistics, Faculty of Science, Silpakorn University, Nakhon Pathom 73000, Thailand.

[b] Department of Environmental Science, Faculty of Science, Silpakorn University, Nakhon Pathom 73000, Thailand.

[c] Mathematics Program, Faculty of Science and Technology, Sisaket Rajabhat University, Sisaket 33000, Thailand.

* Author for correspondence; e-mail: kamolcha@su.ac.th

Received: 22 September 2010
Accepted: 8 December 2010

Abstract

The objective of this research is to study three statistical modeling approaches; namely stepwise multiple linear regression, a feed-forward artificial neural network and a genetic algorithm for predicting quantity of organic matter in soil. Soil samples were selected from three fruit farming agricultural areas in the western region of Thailand; Nakhon Pathom, Samut Sakhon and Samut Songkram. Seventeen soil properties were measured on the soil samples and are used as original variables. To reduce the number of original variables and eliminate data collinearity, a principal component analysis was applied. The models were based on the first five principal components which accounted for 75.81% of total variance. Model performance was measured by performance indexes which are IA, RMSE, MBE and MAE. The results of this study indicated that the genetic algorithm model performs the best among these three models in a validation step and is the most efficient model to predict soil organic matter.

Keywords: feed forward artificial neural networks, genetic algorithm, multiple linear regression, principal components, soil organic matter.

1. Introduction

The development of mathematical models for efficient prediction is becoming more popular. This enables us to foresee potential situations, and thus find ways to prevent unexpected outcomes in the future. Model development is a method that is much employed in scientific applications. In environmental sciences, the method was employed in the prediction of ozone concentration as a warning of possible dangers [1] and also in the prediction of dwelling fire occurrences for management planning and reducing material loss [2]. The method was also employed in agricultural sciences to predict organic carbon [3] and to predict organic matter [4, 5], which are soil quality indications for planning and improving soil quality. Studies have also shown the employment of mathematical model development in medical sciences [6, 7], production engineering [8], and finance and banking [9]. These examples are evidence of the necessity and importance of model development in all work areas.

Thailand is an agricultural country. The agriculture is mostly done in a traditional way. In many cases agricultural land has been cultivated for a continuous length of time without any soil nourishment and sufficient addition of organic matters. This has caused a decrease of organic matters in the soil and is harmful to agriculture in the long run. The development of mathematical models for predicting soil organic matter is, therefore, relevant as a tool for examining the soil quality, which helps determine soil management planning for yielding optimum results and soil quality improvement to help develop agricultural products.

In this study, soil chemical characteristics are studied in relation to soil organic matter in predicting the quantity of soil organic matter (OM), an important indicator of soil quality. The researcher, thus, sees the importance of developing an efficient model which can examine the quantity of soil organic matter.The data collected were measured soil chemical characters in relation to soil organic matter taken from three fruit farming agricultural locations in the western region of Thailand; Nakhon Pathom, Samut Sakhon, and Samut Songkram provinces. In the first step, principal component analysis (PCA) was used to decrease the number of input variables and to eliminate collinearity. The first five principal components will be used as independent variables in predicting soil organic matter. The first method was stepwise multiple linear regression (SMLR). This method is widely used in predicting variables with continuous data that are linearly related to dependent variables, which may result in correlation between independent variables and non-linearity of organic matters. The second method was developing models by artificial neural networks, and this study employs the most commonly-used method; feed-forward artificial neural network (FANN). This method is also widely used in developing models and prediction and can be used to analyze non-linear data, and can solve complicated problems [2]. The third method was a genetic algorithm (GA), which is a relatively new method showing a clear process and analytical strength. In this method, a potential solution to the problem is called a chromosome. The genetic algorithm then creates a population of chromosomes and applies genetic operators such as mutation and crossover to evolve the chromosomes in order to achieve the best solution. The results from the three model methods were compared using performance indexes. This will yield the most efficient and best model in predicting soil organic matter which can lead to a good adaptation of new technologies into agriculture. The objectives of this study, hence, were: (i) to study the relations among independent variables and decrease the data dimensions by analyzing the principal components (PC) and (ii) to compare the model performances by using predictor variables from the principal components employing stepwise multiple linear regression, artificial neural networks, and genetic algorithm methods.

2. Methodology

2.1 Soil samples used in this study

The soil samples were collected from three fruit farming agricultural locations in the western region of Thailand; Nakhon Pathom, Samut Sakhon, and Samut Songkram provinces and 58 sites in all. The samples were examined and soil chemical characteristics were measured that contain17 parameters. Basic sample statistics of soil chemical characteristic data included Aluminum (Al), Manganese (Mn), Iron (Fe), Chromium (Cr), Magnesium (Mg), Zinc (Zn), Copper (Cu), Lead (Pb), Potassium (K), Sodium (Na), Calcium (Ca) , Fulvic Acid (FA), Humic Acid (HA), Cation Exchange Capacity (CEC), Percentage of Clay (%clay), Total Nitrogen (TN), and Organic Carbon (OC) and are shown in Table 1.2.2 Sample sizes were 30, 50 and 100.

2.2 Reduction of the Number of Original Variables

Principal component analysis is a technique to reduce the number of variables and eliminate the relations among input variables by developing a set of new variables that are linear functions of the original variables. This set will retain properties of the original ones, provided that the number of new variables will not exceed the original number. That is, if the original number of variables is p and the number of new variables is m, then $m \leq p$. The number of variables m is chosen components to sufficiently explain the variation of the data.

Table 1. Basic statistics of soil chemical characteristic data from the samples collected.

Variables	Minimum	Maximum	Mean	SD.
1) Al (mg/kg)	7929.24	34947.96	14163.40	4137.62
2) Mn (mg/kg)	190.15	2180.88	836.92	394.33
3) Fe (mg/kg)	11810.52	28602.88	19846.78	3394.37
4) Cr (mg/kg)	13.27	48.71	22.40	4.98
5) Mg (mg/kg)	1270.10	5717.70	3131.85	820.79
6) Zn (mg/kg)	31.36	114.10	60.78	19.91
7) Cu (mg/kg)	11.47	267.44	39.49	42.46
8) Pb (mg/kg)	24.01	61.38	37.83	8.33
9) K (mg/kg)	747.57	4831.45	2167.00	746.72
10) Na (mg/kg)	437.32	5101.70	1697.41	643.47
11) Ca (mg/kg)	3957.08	16816.74	12475.71	2567.58
12) FA (mgC/kg)	10.65	609.80	151.60	113.61
13) HA (mg/g)	6.57	49.47	32.91	12.24
14) CEC(cmol/kg)	14.73	34.29	24.68	4.79
15) % clay(%)	14.27	87.87	44.28	15.21
16) TN (%)	0.07	0.22	0.14	0.04
17) OC(%)	0.70	2.32	1.41	0.42

The variables were standardized due to the difference in the units of measurement. The applicability of the PCA to the data sets used in this study was verified through the application of Bartlett's sphericity test expressed by the following equation:

$$\chi^2 = -\left[(n-1) - \frac{2(p+5)}{6}\right] \ln|\boldsymbol{R}| \tag{1}$$

where χ^2 has $\frac{1}{2}p(p-1)$ degrees of freedom, $\ln|\boldsymbol{R}|$ is the log value of the determinant of the correlation matrix $\boldsymbol{R}$, p is the number of variables, n is the number of data points, and $|\boldsymbol{R}| = \prod_{i=1}^{p} \lambda_i$ where λ_i is the eigenvalue for variable i ; i = 1,2,..., p. The null hypothesis states that the population correlation matrix is an identity matrix. If the obtained chi-square value is significant, then PCA should then be applied. The result from the hypothesis test showed that the chi-square value was equal to 681.672 (p-value = .000). Rejecting the hypothesis means that the strength of the relationships among the variables are strong and appropriate for PCA. To determine the number of components to retain, one of the most commonly-used criterion, called the eigenvalue-one criterion, was applied. With this criterion, the first five principal components with an eigenvalue greater than one were retained. Table 2 indicates the loading values of the first five principal components. These loadings explain the contribution of each variable in a principal component. The bold number means the variable loads on that component. The first five principal components accounted for 75.81% of total variance.

Table 2. Loading Values of the first 5 PC's from soil samples.

Components	PC1	PC2	PC3	PC4	PC5
Al	**0.946**	0.042	0.090	0.022	0.014
Mn	-0.076	-0.071	0.236	**0.788**	-0.114
Fe	**0.775**	0.017	-0.339	0.316	0.012
Cr	**0.937**	-0.058	-0.086	0.007	0.158
Mg	0.109	0.013	-0.166	**0.874**	0.179
Zn	0.535	**0.649**	-0.020	-0.061	0.006
Cu	-0.042	0.268	**0.547**	-0.278	0.271
Pb	0.442	0.109	**-0.669**	0.206	0.208
K	0.377	0.191	-0.299	0.306	**0.611**
Na	0.291	0.009	-0.071	**0.672**	0.293
Ca	-0.240	**0.653**	0.084	0.323	-0.054
FA	0.022	0.310	-0.226	0.016	**0.812**
HA	0.028	0.039	-0.042	0.104	**0.794**
CEC	-0.125	0.022	**0.775**	0.091	-0.302
% clay	0.093	-0.037	**0.858**	0.157	-0.131
TN	0.117	**0.879**	0.014	-0.142	0.324
OC	-0.006	**0.931**	-0.024	-0.092	0.190
Eigenvalue	4.766	2.868	2.135	2.009	1.108
Accumulated variance	0.280	0.449	0.575	0.693	0.758

The first five principal components were then chosen to be the predictor variables in modeling the soil organic matter in stepwise multiple linear regression, feed-forward artificial neural network and genetic algorithm models. The scree plot is shown in Figure 1.

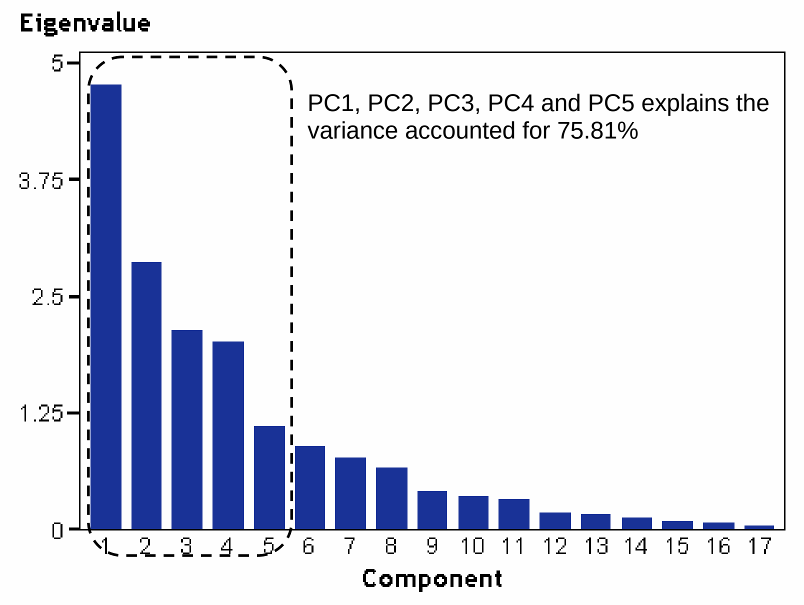

Figure 1. A scree plot for all PC's.

2.3 Performance Indexes

In comparing the precision of models, four performance indexes were used; namely, index of agreement (IA), root mean square error (RMSE), mean bias error (MBE) and mean absolute error (MAE) given by equations (2)-(5), respectively:

$$IA = 1 - \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n}(|\hat{Y}_i - \bar{Y}| + |Y_i - \bar{Y}|)^2} \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{3}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i) \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i| \tag{5}$$

where Y_i is the observed data value of soil organic matter for sample i, $\hat{Y}_i$ is the fitted value of soil organic matter of sample *i*, $\bar{Y}$ is the mean value of quantity of soil organic matter, and n is the sample size. The RMSE and MAE measure residual error. The MBE indicates whether the observed values are overestimated or underestimated.

2.4 Models

2.4.1 Multiple Linear Regression

Multiple linear regression is one of the most widely used methods for expressing the relationship between a response variable and several independent variables. Equation (6) is a multiple linear regression model for *p* variables:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{6}$$

where *Y* is the response variable, X_i *(i* = 1,…,*p*) are the independent variables, β_i (*i* = 1,…,*p*) are the parameters usually estimated by least square.

The SMLR model was developed in this study by using stepwise selection. From the total of 58 soil samples, the training data set was formed by randomly selecting 52 of the 58 samples, whereas the test data set consists of the remaining 6 samples (samples 29, 33, 34, 48, 50 and 52). These sets will also be used to develop the FANN and GA models.

2.4.2 Artificial Neural Networks

Artificial neural networks (ANN) are widely used as an alternative way to solve complex and non-linear problems. The most common type of artificial neural network consists of three layers: the input layer, the hidden layer and the output layer. In general, an artificial neural network consists of the processing element or neurons. The input neurons receive data from the outside. The hidden neurons receive signals from the input layer. The output neurons return the output information. Details of the use of ANN can be found in [10]. The ANN model illustrated in Figure 2 can be explained mathematically by equations (7)-(9):

$$u_k = \sum_{j=1}^{p} w_{kj} x_j \tag{7}$$

$$y_k = \varphi(u_k + b_k) \tag{8}$$

$$\upsilon_k = u_k + b_k \tag{9}$$

where X_j (*j = 1,…,p*) are predictor variables or inputs, w_{kj} (*j = 1*,…,*p*) are loadings of neurons at *k*, b_k is a bias value, $\varphi(\cdot)$ is the stimulating function, y_k is the dependent variable or output, and υ_k is the sum of the predictor variables and the bias value.

The most common architecture of ANN is the feed-forward artificial neural network (FANN). FANN allows signals to travel one way from input to output. There is no feedback.

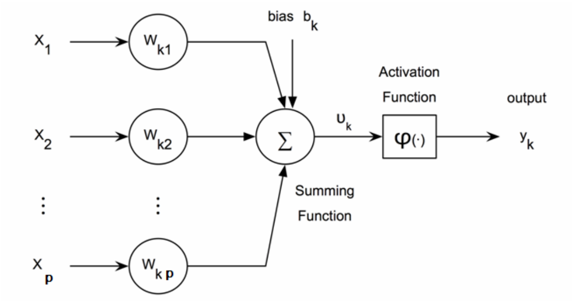

Figure 2. Artificial neural networks model.

The number of nodes in the first level is fixed by the number of predictor variables or inputs in the model, while the number of nodes in the result level or output is equal to the number of results required in the model. One important factor in developing models is choosing the number of nodes and transfer function by FANN. The general format of FANN in this study is as follows:

Organic Matter = FANN (PC1, PC2, PC3, PC4, PC5)

where FANN is the function of the feed-forward artificial neural network in learning and requires a transfer function and an algorithm appropriate for learning. This study will use a hyperbolic tangent transfer function and employ the Levenberg–Marquardt algorithm [11] as the training network method. The data was divided into two data sets: the training data set and the test data set which are the same as for the SMLR model in order to compare the prediction results.

2.4.3 Genetic Algorithm

A genetic algorithm (GA) is a search method that mimics the principle of natural selection. Because a GA requires few assumptions, it can be used to solve a broad range of problems. To use a GA, a solution to a problem is represented as a chromosome. The GA creates a population of potential solutions and applies genetic operators including selection, mutation and crossover to achieve the best one.

In this study, a GA is employed together with multiple linear regression. The GA is applied for choosing the best regression coefficients for the multiple linear regression model. To reduce the data set dimension, only the first five principal components are used accounting for 75.81% variance for all the input variables. The objective function is defined as the root mean square error shown in equation (10);

$$RMSE = min\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \qquad (10)$$

where Y_i is the real value of soil organic matter, $\hat{Y}_i$ is the prediction of soil organic matter, and *n* is the sample size. The genetic operations are defined as follows:

Population size	*52*
Probability of Crossover	*0.85*
Probability of Mutation	*0.00*
Replacement	*steady state*
Selection	*tournament (size = 10)*
Maximum Generation	*100.*

The bounds of coefficient randomization are obtained from the 52 learning sets in order to develop a 95% confidence intervals for each of the regression coefficients produced by the GA process. The probability of mutation equals zero due to the differences of intervals of regression coefficients. Table 3 shows the minimum and maximum values of the regression coefficients.

Table 3. Minimum and maximum values of regression coefficients for 95% confidence intervals**.**

Regression Coefficient	Minimum	Maximum
b_0	2.376	2.496
b_1	0.104	0.159
b_2	0.319	0.391
b_3	-0.091	-0.008
b_4	-0.037	0.049
b_5	-0.238	-0.123

3. Results

3.1 Results from SMLR

Stepwise multiple linear regression, using the first five principal components as independent variables, indicated that PC1, PC2 and PC5 explained 89% of the variation in the dependent variable as shown in equation (11). The significance levels for entry into and staying in the model are 0.15. The standard errors of estimates are shown in Table 4.

$$\hat{Y} = 2.423 + 0.135PC_1 + 0.367PC_2 - 0.184PC_5 \qquad (11)$$

Table 4. The parameter estimates and standard error of estimates.

Model	Constant	PC1	PC2	PC5
Parameter estimate	2.423	0.135	0.367	-0.184
Standard error	0.034	0.015	0.020	0.030

3.2 Results from FANN

FANN is used in learning and revising the loading of the network, and found that the best model contained 5 nodes in the input layer, 3 nodes in the hidden layer, and 1 node in the output layer. The loading and bias of each node is shown in Figure 3.

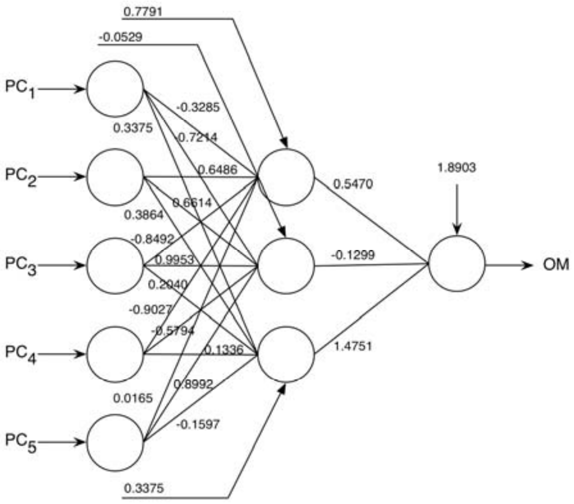

Figure 3. The structure of the best model in learning the network from FANN.

3.3 Results from the GA

By using the genetic algorithm in selecting the best regression coefficients for the MLR equation, the first five principal components explained 98% of the variation of soil organic matter as shown in equation (12).

$$\hat{Y} = 2.440 + 0.117PC_1 + 0.367PC_2 - 0.809PC_3 + 0.047PC_4 - 0.160PC_5 \qquad (12)$$

3.4 Comparisons of the model efficiency

Table 5 and Figure 4 show the predicted values from the 3 models compared to the actual organic matter. The RMSE, MAE, and R calculated from the test sets are illustrated in Table 6.

Table 5. The real values and the predicted values from the 3 models.

Set	Real Organic Matter	SMLR	FANN	GA
1	1.9477	1.7296	1.7872	1.9264
2	3.6756	3.5493	3.4354	3.6401
3	2.5483	2.0695	2.1539	2.4262
4	3.1388	3.1084	3.0072	3.0152
5	2.4351	2.7770	2.5659	2.6354
6	2.2166	2.0418	1.9595	2.1645

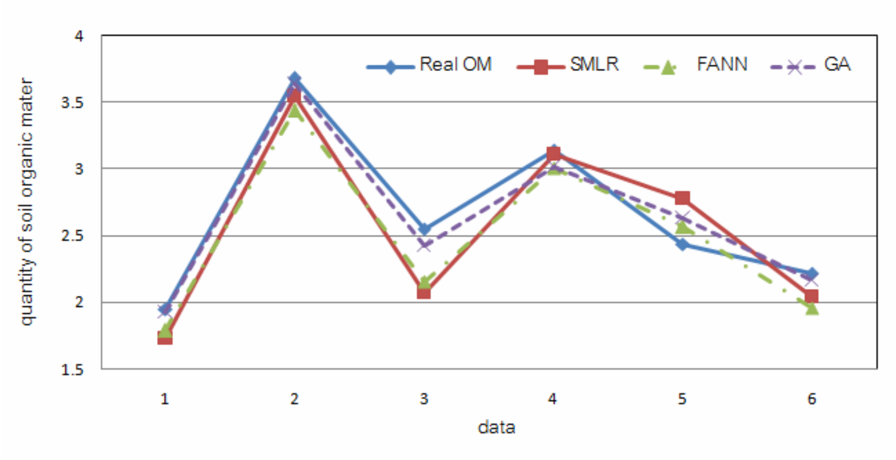

Figure 4. Comparing the quantity of real OM and OM predicted by using the SMLR, FANN and GA models on the test data set.

The results from predictions using SMLR, FANN, and GA models showed that the GA model indicated most accurate results among the three models. According to the IA index, the FANN and GA models gave the same value of 0.99, which was more than that of the SMLR model yielding the value of 0.97. The RMSE and MAE results were also in accordance with IA. The GA model yielded values of 0.11 and 0.09, respectively, while the FANN model yielded values of 0.24 and 0.22. The SMLR model, on the other hand, yielded less satisfactory results of 0.27 and 0.23. MBE indicated that the best prediction method was GA, which yielded the value of -0.03. The second best method was SMLR at the value of -0.11, and FANN came in third at the value of -0.17. Table 6 summarizes the results.

Table 6. Performance indexes for all 3 models.

Performance indexes	SMLR	FANN	GA
IA	0.97	0.99	0.99
RMSE	0.27	0.24	0.11
MBE	-0.11	-0.17	-0.03
MAE	0.23	0.22	0.09

In conclusion, similar performances were obtained with MLR and FANN models but the GA model gives the best values. The comparisons of performance indexes for all 3 models are shown in Figure 5.

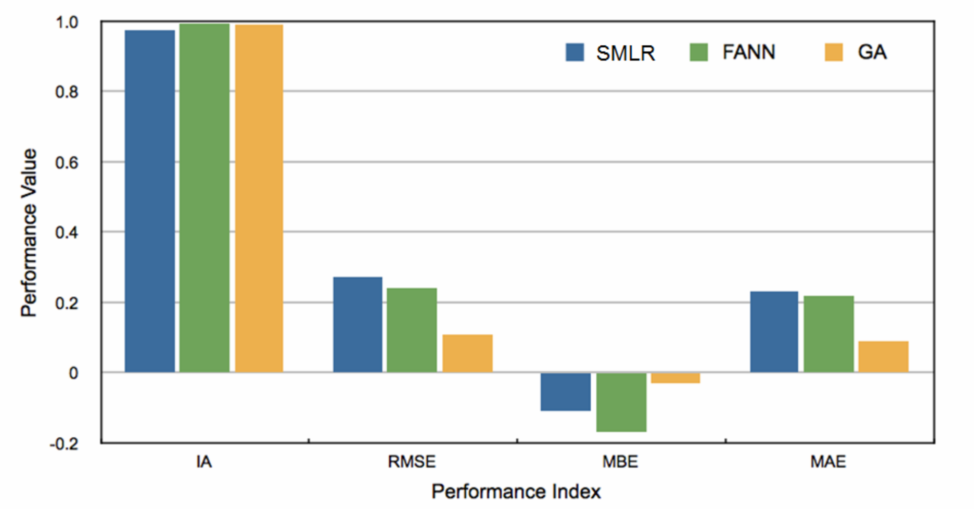

Figure 5. Comparing performance values for all 3 models.

4. Discussion

In developing models for prediction of soil OM, the most important point is the data preprocessing. By removing unnecessary data when selecting the model, the more accurate the result will be. This study used the analysis of principal component in eliminating original variables or unnecessary data from the models, yet the information contained in the data remains. This enabled us to achieve the appropriate number of the predictor variables in the models. This can also eliminate the relations among the independent variables in model fitting, which may lead to inaccurate predictions.

The prediction of soil organic matter by using SMLR, FANN and GA methods are all different approaches depending on the problem of interest. SMLR is easier to understand, not very complicated, and easy to use. The disadvantage of this method is that the data used in the process must be linear. As the data in connection with this study are environmental data that constantly change and possess non-linear properties, this proved to be a disadvantage using SMLR method. The second method is FANN, which can be used with non-linear data, but how accurate the FANN is in prediction depends on the networks specified, number of nodes in hidden layer, selection of a transformation function and a learning algorithm appropriate for the models. The last method is GA, which is more complicated, but can be used with non-linear data and problems with specified conditions or limited bounds, which is the advantage of this method. This study used GA together with SMLR in selecting the coefficients for SMLR model, thus this also need to take into account the appropriate specification of variable bounds and parameters to make them suitable for the problems that need to be solved.

The prediction of soil organic matter by using SMLR, FANN, and GA methods yield almost the same results, but overall, the GA method yields the most accurate results. This could be because the GA does not require linearity in the data. Although the GA is a complicated method, it is a new and appropriate technique for analyzing environmental data or continuously changing data. This can be adapted in developing more accurate prediction from the models. According to the performance indexes, the results of this study indicated that genetic algorithm model performs better than stepwise multiple linear regression and feed-forward artificial neural network models in the validation step. The GA model is the most efficient model to predict soil organic matter.

References

[1] Sousa S.I.V., Fernando G. Matins, Maria C. Alvim-Ferra, and Maria C. Pereira. "Multiple linear regression and artificial neural networks based on principal component to predict ozone concentrations." Environmental Modelling & Software, 2007:22; 97-103.

[2] Yang, L., Christian W. Dawson, Martin R. Brown, and Michael Gell. "Neural network and GA approaches for dwelling fire occurrence prediction." Knowledge-Based Systems, 2006;19: 213-219.

[3] Ingleby, H.R., and Crowe T.G., "Reflectance models for predicting organic carbon in Saskatchewan soils." Canadian Biosystems Engineering, 2000;42: 57-63.

[4] Ingleby, H.R., and Crowe T.G., "Neural network models for predicting organic matter content in Saskatchewan soils." Canadian Biosystems Engineering Vol.43 (2001) : 7.1 - 7.5.

[5] Panishkan K., Areekijseree M., Sanmanee N. and Swangjang K. “Soil Classification based on their Chemical Composition using Principal Component Analysis.” Environment Asia, 2010;3:47-52.

[6] Jung Yi Kim and others. "Comparative study on artificial neural network with multiple regressions for continuous estimation of blood pressure." In Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, 6942-6945. Edited by IEEE. China : IEEE,1992.

[7] Jae, Song H. and others. "Comparative Analysis of Logistic Regression and Artificial neural network for Computer-Aided Diagnosis of Breast Masses." Academic Radiology, 2005;12:487-495.

[8] Erenturk, S., and Koksal Erenturk. "Comparison of genetic algorithm and neural network approaches for the drying process of carrot." Journal of Food Engineering, 2007;78:905-912.

[9] Rurkhamet, B., Parames Chutima, and Manop Reodech, "Comparative study of artificial neural network and regression analysis for forecasting new issued banknotes." Thammasat Int. Sc. Tech 3; 1998:21-28.

[10] Haykin, S., Neural Networks: A Comprehensive Foundation. 2nd ed. New York: Prentice Hall, 1998.

[11] Matignon, R., Data Mining Using SAS Enterprise Miner. 2nd ed. Hoboken, NJ : John Wiley & Sons, Inc., 2007.