Computer and Information Science

www.ccsenet.org/journal.html

Vol. 2, No. 3

August 2009

Estimation of Saturation Percentage of Soil Using Multiple Regression, ANN, and ANFIS Techniques

Khaled Ahmad Aali (Corresponding author)

Department of Irrigation and Reclamation, Faculty of Water and Soil Engineering

University of Tehran, Karaj, Iran

Tel: 98-914-185-3937 E-mail: khaled.ahmadauli@gmail.com

Masoud Parsinejad

Department of Irrigation and Reclamation, Faculty of Water and Soil Engineering

University of Tehran, Karaj, Iran

Bizhan Rahmani

Soil Science Department, College of Agriculture

Azad Islamic University of Karaj, Karaj, Iran

Abstract

The saturation percentage (SP) of soils is an important index in hydrological studies. In this paper, artificial neural networks (ANNs), multiple regression (MR), and adaptive neural-based fuzzy inference system (ANFIS) were used for estimation of saturation percentage of soils collected from Boukan region in the northwestern part of Iran. Percent clay, silt, sand and organic carbon (OC) were used to develop the applied methods. In additions contributions of each input variable were assessed on estimation of SP index. Two performance functions, namely root mean square errors (RMSE) and determination coefficient (R^2), were used to evaluate the adequacy of the models. ANFIS method was found to be superior over the other methods. It is, then, proposed that ANFIS model can be used for reasonable estimation of SP values of soils.

Keywords: Saturation percentage, Soils, MR, ANN, ANFIS, Boukan

1. Introduction

Saturation percentage (SP) is related to the mechanical constituents of soils and can, therefore, be regarded as a quantitative measure of soil texture, water-holding capacity, and cation exchange capacity. Soil profiles may be described in terms of SP, and soil maps may be developed to represent quantitative changes in soil texture within a region. Furthermore, measurement of soil water content is important in simulation of all aspects of hydrological cycle, for estimation of plant water use, and for characterizing most soil physical, chemical, and biological processes. Chemically, water serves as transport agent for dissolved inorganic chemicals and suspended biological components, involved in the processes of soil development and degradation.

The saturation percentage is defined as the ratio of the amount of water added to saturate dry soil samples, to total mass of the fully dried soil. Direct measurement of saturation percentage is time consuming and relatively expensive. In the conventional procedure, initially dried soil samples are saturated with deionized water and then oven dried at 105 C° for a period 24 hrs.

Indirect methods are, then, used as an alternative solution. Numerous attempts have been made to correlate base-saturation percentage with pH in saline suspensions. (Keeney and Corey, 1963; Shaw, 1952) Such efforts have been, at best, only partially successful, particularly if one attempts to estimate precise levels of base saturation from pH measurements.

In view of the importance of accurate estimation of saturation percentage (SP), using basic and readily available soil information, adoption of modern techniques such as artificial neural networks (ANNs) and fuzzy inference system (FIS) can be a viable alternative. Because of the non-linear structure in ANNs models and ambiguity in variables in FIS models, (Piotrowski et al. 1996; Mukhopadhyay 1999), researchers are, recently attracted in using hybrid models such

as Adaptive Neural-based Fuzzy Inference System (ANFIS) to further analyze the variables, which are spatially distributed. (Lee, 2000)

In this study, efficiency of Adaptive Neural-based Fuzzy Inference System (ANFIS), artificial neural networks (ANN) and multiple regression (MR) models were examined in estimation of saturation percentage (SP) using measured data of clay, silt sand and organic carbon (OC), in Boukan plain in the West Azerbaijan Province, Iran.

2. Material and methods

2.1 Description of the study area

The study area is Boukan region which is located in southern part of West Azerbaijan Province, Iran. Boukan covers an area of 47300 hectares (Figure.1), with latitude of 36º 32' and longitude of 46º 13'. Average elevation in this region is 1330 m above sea level. The area falls under the semiarid climate with an average rainfall of 517 mm/year. 600 measured values of clay, silt, sand, organic carbon (OC %) and saturation percentage (SP %) which was previously collected by the Iranian Rojhalat Soil Lab were used in this study. The Walky black method was used to determine OC content in the samples and soil texture (percent sand-silt-clay) were determined using Hydrometer method (Schumacher 2002).

A summary of obtained results and their basic statistics is presented in Table1. The SP values ranged between 28 and 66 (%) with an average value of 48.18%. The respective average values of effective organic carbon, percent of clay, percent of silt and percent of sand were determined as 0.72, 24.43, 51.34 and 24.23 %.

Simple regression analysis was performed to initially establish the predictive relationship between measured parameters. The relations between SP and other measured parameters were analyzed using linear, power, logarithmic, and exponential functions. Models with statistically significant and strong correlations were then selected for further analysis (Table 2). Regression equations were also established among index parameters with SP (Table 3). All obtained relationships were found to be statistically significant according to the Student's t-test at 99% level of confidence.

2.2Artificial Neural Network (ANN)

Artificial neural networks (ANNs) are based on current understanding of biological nervous systems, though much of the biological details are neglected. ANNs are massively parallel systems composed of many processing elements connected by links of variable weights (Lippman, 1987).

In Figure 2 a three-layered neural network consisting of i, j and k layers with the interconnection weights W_{ij} and W_{jk} between layers of neurons is illustrated (Hagan and Menhaj 1994; Kisi and Uncuoglu 2005). The weights are computed through an iterative process based on back propagation algorithm in such a way that the difference between computed and given output (or any error criterion such as mean square error) is sufficiently small. The hidden layer node numbers of each model were determined after trying various network structures, since there is no theory yet available to tell how many hidden units are needed to approximate a given function. Cross validation mode (checking mode) monitors the error to find the optimal termination point for training and also avoid overtraining. Testing mode is used to determine how accurately the network can simulate input-output relationships.

All collected data were divided into three sets, namely; training (3/5 of all data), test (1/5 of all data), and verification (remaining 1/5 of all data). In this study MatLab 7.4 software was used in neural network analysis having a three-layer feed-forward network that consisted of an input layer, one hidden layer, and one output layer. Logsigmoid (transfer) functions for both hidden and output layers were used for analysis network activation.

2.3 Adaptive Neural-based Fuzzy Inference System (ANFIS)

Adaptive Neural-based Fuzzy Inference System (ANFIS) is capable of approximating any real continuous function on a compact set to any degree of accuracy (Jang et al., 1997). Specifically, ANFIS system of interest here is functionally equivalent to the Sugenofirst-order fuzzy model (Jang et al., 1997; Drake, 2000). The hybrid learning algorithm, introduced as follows, combines gradient descent and the least-squares method. As a simple example a fuzzy inference system is assumed with two inputs x and y and one output z. The first-order Sugeno fuzzy model, a typical rule set with two fuzzy If-Then rules, can be expressed as:

$$\text{Rule 1: If x is } A_1 \text{ and y is } B_1, \quad \text{then } f_1 = p_1 x + q_1 y + r_1 \tag{1}$$

$$\text{Rule 2: If x is } A_2 \text{ and y is } B_2, \quad \text{then } f_2 = p_2 x + q_2 y + r_2 \tag{2}$$

The resulting Sugeno fuzzy reasoning system is presented in Figure 3, where the output *z* is the weighted average of the individual rule outputs and is itself a crisp value. The corresponding equivalent ANFIS architecture is presented in Figure 4. Nodes at the same layer have similar functions. The node function is described next. The output of the *i*th node in layer *l* is denoted as O_{li}.

Layer 1: Every node *i* in this layer is an adaptive node with node function

$$O_{1,i} = 礎_i(x), \qquad \text{for } i = 1,2, \text{ or}$$
$$O_{1,i} = 碟_{i-2}(y), \qquad \text{for } i = 3,4$$

where x (or y) is the input to the *i*th node, and A_i (or B_{i-2}) is a linguistic label (such as ''low'' or ''high'') associated with this node. In other words, O_{1i} is the membership grade of a fuzzy set A (= A_1, A_2, B_1,or B_2) and it specifies the degree to which the given input x (or y) satisfies the quantifier A .

$$\mu A_i(x) = \frac{1}{1+[(x-c_i)/a_i]^{2bi}} \qquad (3)$$

where $\{a_i, b_i, c_i\}$ is the parameter set. As the values of these parameters change, the bell-shaped function varies accordingly, thus exhibiting various forms of membership functions on linguistic label A_i. In fact, any continuous and piecewise differentiable functions, such as commonly used in triangular shaped membership functions, are also qualified as candidates for node functions in this layer (Jang, 1993). Parameters in this layer are referred to as premise parameters. The outputs of this layer are the membership values of the premise part.

Layer 2: This layer consists of nodes labeled Π, which multiply incoming signals and sending the product out. For instance,

$$O_{2,i} = W_i = 礎_i(x)\,碟_i(y), \quad i = 1,2. \qquad (4)$$

Where, each node output represents the firing strength of a rule.

Layer 3: In this layer, the nodes labeled *N* calculates the ratio of the *i*th rule's firing strength to the sum of all rules' firing strengths

$$O_{3,i} = \overline{W}_i = \frac{W_i}{W_1 + W_2} \qquad i = 1,2. \qquad (5)$$

the outputs of this layer are called normalized firing strengths.

Layer 4: This layer's nodes are adaptive with node functions

$$O_{4,i} = \overline{W}_i f_i = \overline{W}_i (p_i x + q_i y + r_i) \qquad (6)$$

where $\overline{W}_i$ is the output of layer 3, and $\{p_i, q_i, r_i\}$ is the parameter set. Parameters of this layer are referred to as consequent parameters.

Layer 5: This layer's single fixed node labeled Σ computes the final output as the summation of all incoming signals

$$O_{5,i} = W_i f_i = \sum_{i=1} \overline{W}_i f_i = \frac{\sum_i W_i f_i}{\sum_i W_i} \qquad (7)$$

In the present study, the triangular and Gaussian membership functions were used. In each application, different numbers of membership functions were tested and the best one, with minimum root mean square error (RMSE) and the maximum R^2, was selected.

A hybrid intelligent system called ANFIS (the adaptive neuro-fuzzy inference system) for predicting SP was also applied. ANFIS was trained with the help of Matlab (version 7.4) and SPSS (15.0 package), and two top models, namely ANFIS11 and ANFIS12 were selected based on RMSE and R^2. ANFIS parameter types for the two models and their values are presented in Table 4.

2.4 Multiple Regression Models

Multiple Regression (MR) is a statistical technique that allows us to predict someone's score on one variable on the basis of their scores on several other variables. MR was used in order to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. The general form of these models is $y = b_1x_1 + b_2x_2 + ... + b_nx_n + c,$ where $\{b_1, b_2, ..., b_n\}$ the regression coefficients are, and y is now written as a function of n independent variables; $x_1, x_2, x_3, ...x_n$. C is the y-intercept (Milton et al., 1997; McClave et al., 1997). Eight MR analysis were carried out to correlate the measured SP to various combinations of measured parameters; namely clay, silt, sand and OC content

3. Application and results

in this study various combinations of measured parameters (clay, silt, sand, organic carbon percentage, and saturation percentage (SP %)) were examined as inputs to ANN models for evaluation of effect of each variable on SP (%). Several NF models were established with different variable added into the input combination at one time. Thus, the input combinations evaluated here were: (i) clay; (ii) silt; (iii) sand; (iv) OC; (v) OC and clay; (vi) silt and sand; (vii) OC and silt; (ix) OC and sand; (x) clay and silt; (xi) clay and sand; (xii) clay, silt and sand (xiii) OC, clay, silt and sand.

In this study, estimated and observed values were statistically compared and analyzed using root mean square error (R-MSE) and determination coefficient (R^2).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \quad (8)$$

Where n is the number of observations, O_i and P_i are the measured and predicted values, respectively. The calculated indices are presented in Table 5. The obtained results shows that ANN models with OC, clay, silt and sand inputs (combination (xiii)) had the smallest RMSE and the highest R^2. This observation emphasizes that all of these parameters have a weight on prediction of SP (%). The best two ANN models among all applied models are presented in the Table 5.

The coefficient of correlation between measured and predicted values is considered as a valuable indicator of the predictability of models. These relationships, presented in Figure 5, were found to be highly correlated. Cross-correlation between predicted and observed values (Figure 5) indicated that the constructed ANN model is acceptable for prediction of SP.

Multiple regression models were developed to predict SP with different combinations of inputs and the two best models were selected based on R^2 and RMSE indices (Table 6). Cross-correlation between predicted and observed values (Figure 6) indicated that the constructed MR model is acceptable for prediction of SP.

According to the RMSE and R^2 values (Table 7) and cross-correlation between predicted and observed values (Figure 7), ANFIS model constructed has a high prediction performance for prediction of SP.

Analysis of ANFIS models with various combinations of inputs and different type of membership functions with different numbers showed that the model having OC, clay, silt, and sand as inputs with triangular membership function type has the best performance, followed by, the model having clay, silt, and sand combination as inputs with Gaussian membership function.

4. Conclusion

In this study Adaptive Neural-based Fuzzy Inference System (ANFIS), artificial neural network (ANN) and multiple regression (MR) techniques were used for prediction of Saturation Percent (SP) using characteristics of soils; namely, percent sand, silt, clay, and organic carbon (OC). Appropriate models were developed by scrutinizing their performance degrees and the model with minimum RMSE and maximum R^2 was selected as best model. The results showed that constructed ANFIS and ANN models were effectively able to predict SP. The comparison of developed models showed that ANFIS and ANN models are superior as compared with MR functions in estimating SP indices. In this study, option of estimating SP using proposed empirical relationship and models is acknowledged.

References

Drake, J.T. (2000). Communications phase synchronization using the adaptive network fuzzy inference system. Ph.D. Thesis, New Mexico State University, Las Cruces, New Mexico, USA.

Gardner, W.H. (1986). Water content. In: A. Klute Ed. Methods of Soil Analysis, Part 1—Physical and Mineralogical Methods, 2nd ed. American Society of Agronomy, Soil Science Society of America, Madison, WI, pp. 493–544.

Hagan M.T., and M.B. Menhaj. (1994). Training feed forward networks with the Marquard algorithm. IEEE Trans. Neural Networks 6, 861– 867.

Jang, J. S.R. (1993). ANFIS: Adaptive-Network-Based Fuzzy Inference System. IEEE Trans. Syst. Manag. Cyber. 23 (3), 665–685.

Jang, J. S.R. and C. T, Mizutani, E. (1997). Neuro-Fuzzy and soft computing: A computational approach to learning and machine intelligence. Prentice-Hall, Upper Saddle River, NJ.

Keeney, D. R. and R. B. Corey. (1963). Factors affecting the lime requirements of Wisconsin soils. Soil Sci. Soc. Amer. Proc. 27: 277-280.

Kisi, O. and E. Uncuoglu, (2005). Comparison of three backpropagation training algorithms for two case studies, Indian J. of Eng. and Materials Sci., Vol. 12.

Kisi, O. (2007). Development of Stream flow-Suspended Sediment Rating Curve Using a Range Dependent Neural Network. International Journal of Science and Technology Vol. 2, No 1, 49-61.

Lee E.S. (2000). Neuro-Fuzzy Estimation in Spatial Statistics. Journal of Mathematical Analysis and Applications, 249, 221-231.

Lippman, R. (1987). An introduction to computing with neural nets. IEEE ASSP Mag. 4, 4–22.

Matlab 7.4.0. (2007). Software for technical computing and model-based design. The MathWorks Inc.

Mcbratney, A.B. and I.O.A. Odeh, (1997). Application of Fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. Geoderma, 77, p.85-113.

McClave, J. T., Dietrich, F. H., and Sinicich, T. (1997). Statistics. Prentice-Hall. Inc.

Milton, J. S., McTeer, P. M., and corbet, J. J. (1997). Introduction to statistics. McGraw-Hill.662 PP.

Mukopadhyay A.. (1999). Spatial Estimation of Transmissivity Using Artificial Neural Network. Ground Water, 37(3), 458-464.

Piotrowski J.A., F. Bartels, A. Salski, and Schmidt G. (1996). Geostatistical Regionalization of Glacial Aquitard Thickness in Northwestern Germany, Based on Fuzzy Krigging. Mathematical Geology, 28(4), 437-448.

Shaw W. M. (1952). Report on exchangeable hydrogen in soils. Interrelationships between calcium sorption, exchangeable hydrogen, and pH values of certain soils and subsoils. J. Assoc. Offic. Agr. Chemists 35: 597-621.

SPSS 15.0.0. (2006). Statistical analysis software (Standard Version). SPSS Inc.

Brian A. Schumacher. (2002). Methods for the determination of total organic carbon (TOC) in soils and sediments. NCEA-C- 1282 EMASC-001.

Zimmerman, H. J. (1996). Fuzzy Set Theory-and its applications. Boston, Kluwer Academic Publishers, p.435.

Table 1. Statistical analysis of the measured soil parameters

	SP (%)	OC (%)	Clay (%)	Silt (%)	Sand (%)
Minimum	28.00	0.08	6.00	16.00	3.00
Maximum	66.00	2.16	52.00	71.00	72.00
Range	38.00	2.08	46.00	55.00	69.00
Skew ness	-0.45	0.83	0.54	-1.58	1.33
Kurtosis	0.25	2.18	0.37	4.62	3.11
Mean	48.18	0.72	24.43	51.34	24.23
Std. dev	6.28	0.26	7.01	7.90	10.66

SP: Saturation Percentage OC: Organic Carbon

Table 2. Correlation coefficients (R^2) obtained from simple regression between SP and other measured parameters

model	Dependent	OC	clay	silt	sand
linear	SP	0.089	0.335	0.264	0.586
logarithmic	SP	0.083	0.333	0.267	0.528
power	SP	0.089	0.325	0.313	0.527
exponential	SP	0.09	0.318	0.304	0.616

Table 3. Predictive relationship for assessing SP, using available measured values.

Input combination	Predictive model	R^2
SP- clay	SP=35.517+0.519clay	0.335
SP- clay	SP=9.252+12.344Ln(clay)	0.333
SP- clay	$SP=20.571+ clay^{0.267}$	0.325
SP- clay	$SP=36.430e^{0.011clay}$	0.318
SP- sand	SP=59.103-0.451clay	0.586
SP- sand	SP=80.093-10.312Ln(clay)	0.528
SP- sand	$SP=96.031clay^{-0.226}$	0.527
SP- sand	$SP=61.025e^{-0.010clay}$	0.616

Table 4. Type and values of parameter used for training ANFIS

ANFIS parameter type	Value	
	ANFIS11	ANFIS12
Number of input	3	4
MF type	Trigonal function	Gussian functin
Number of MFs	2 2 2	2 2 2 2
Number of nodes	34	55
Number of linear parameters	8	16
Number of nonlinear parameters	18	16
Total number of parameters	26	32
Number of training data pairs	360	360
Number of checking data pairs	120	120
Number of fuzzy rules	8	16

Table 5. Statistical analysis for the best two ANN models in train and test period

Input combinations	Neuron number	Train		Test	
		R^2	RMSE	R^2	RMSE
clay, silt and sand	23	0.642	3.946	0.5878	3.6458
OC, clay, silt and sand	5	0.654	3.919	0.5886	3.6796

Table 6. Performance indices (R^2 and RMSE) for two preferred MR models for prediction of SP

Input combinations	R^2	RMSE
Clay, silt sand	0.564	3.793
OC, clay, sand and silt	0.577	3.7507

Table 7. Preferred ANFIS's parameter type and their performance indices

Input combinations	Membership functions type	Number of membership functions	train		test	
			R^2	RMSE	R^2	RMSE
clay, silt and sand	Triangular membership function (trimf)	2 2 2	0.649	3.886	0.607	3.585
OC, clay, silt and sand	gussian membership function (gussmf)	2 2 2 2	0.686	3.676	0.593	3.647

Figure 1. The study area

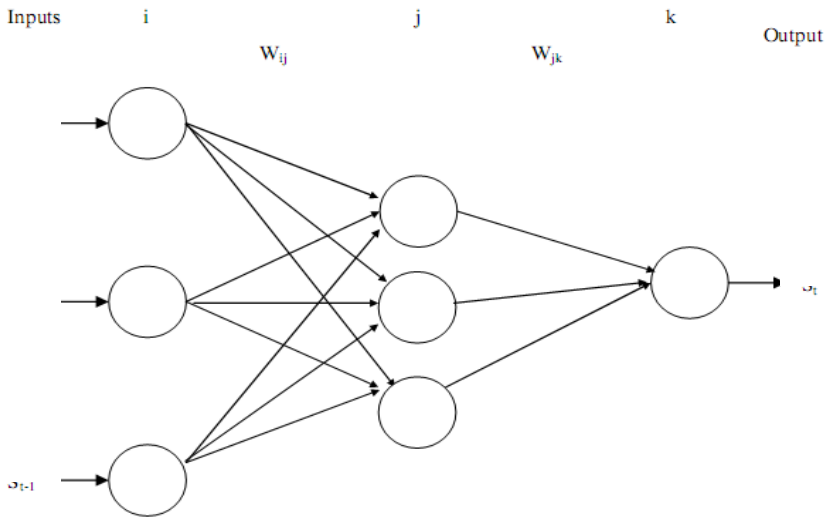

Figure 2. Architecture of three-layer feedforward network (Kisi 2007)

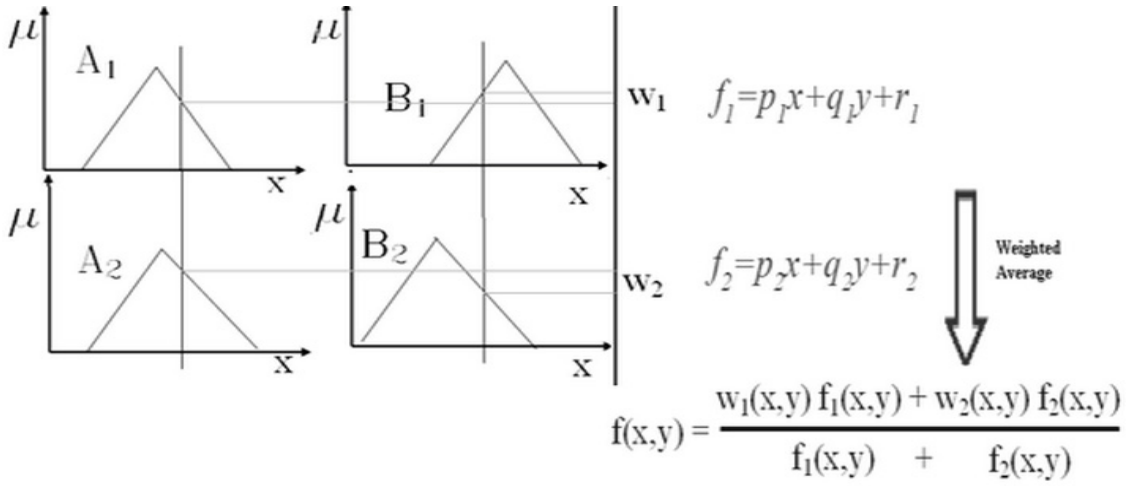

Figure 3. First-order Sugeno fuzzy model with two rules

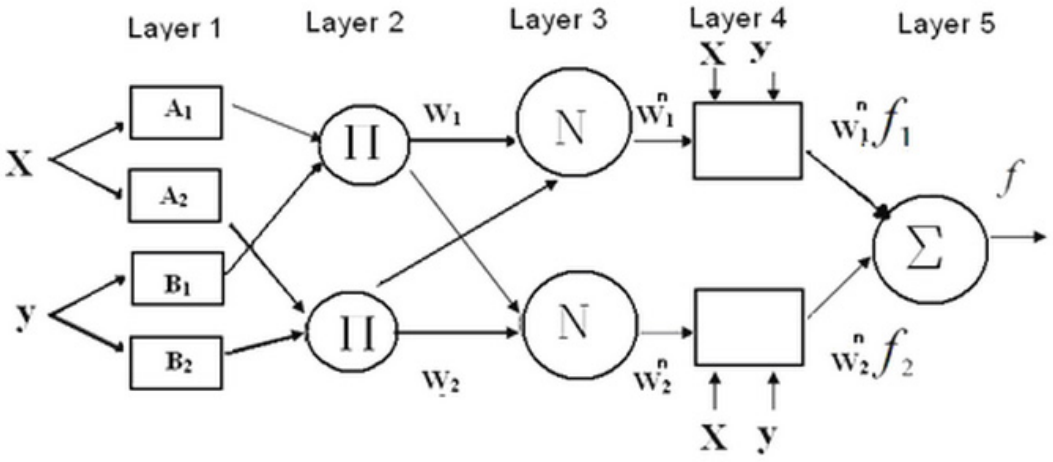

Figure 4. Equivalent ANFIS architecture.

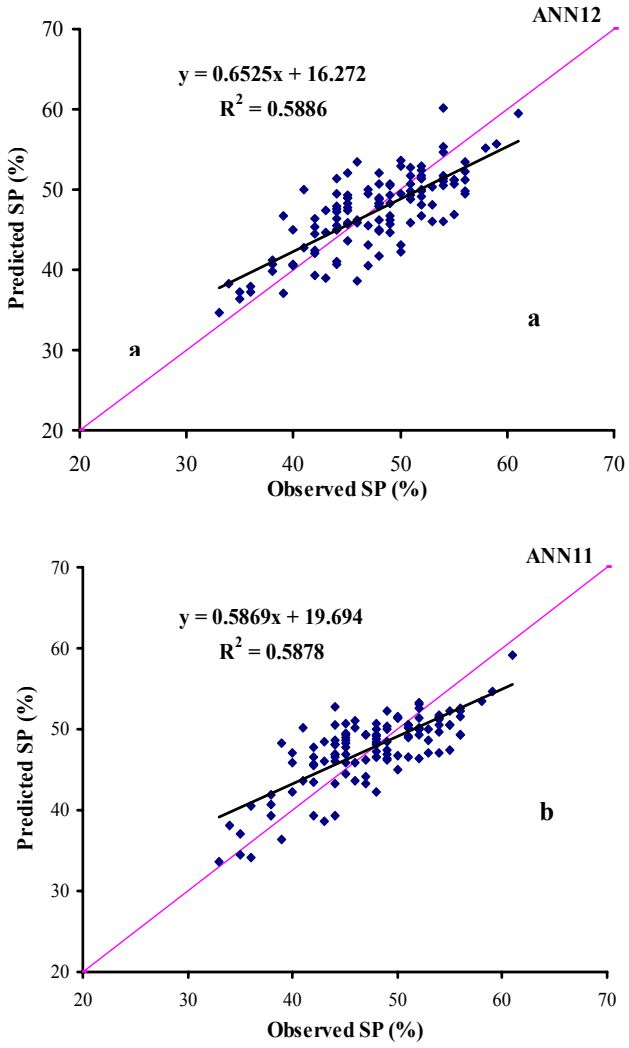

Figure 5. Observed and estimated SP (%) in test period in ANN models with 2 sets of inputs: **a)** OC, clay, silt and sand b**)** clay, silt and sand

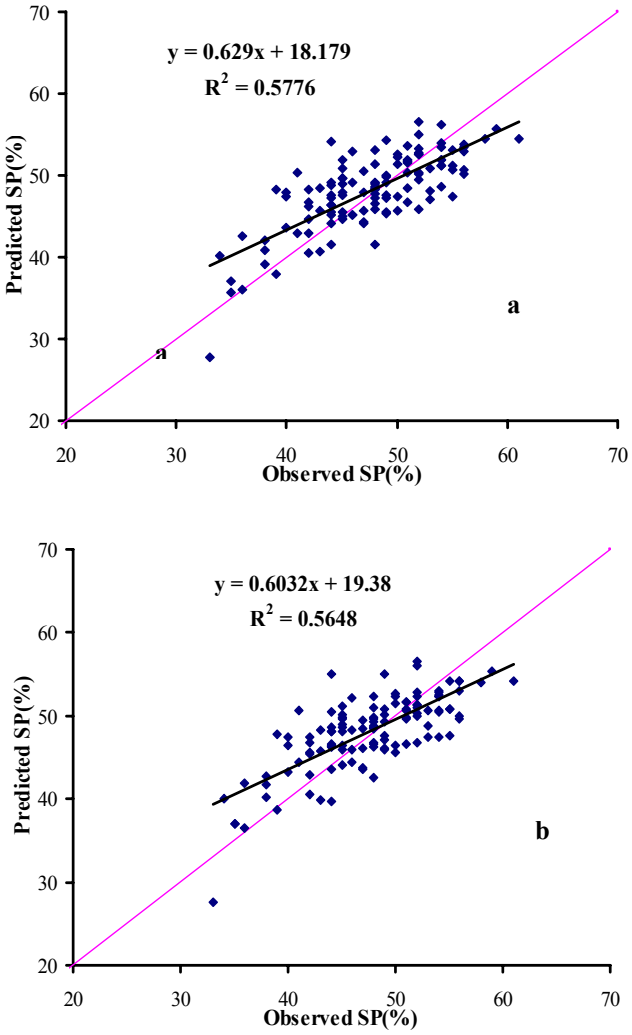

Figure 6. Observed and estimated SP (%) in test period in multiple regression models with 2 sets of inputs: **a)** OC, clay, silt and sand **b)** clay, silt and sand

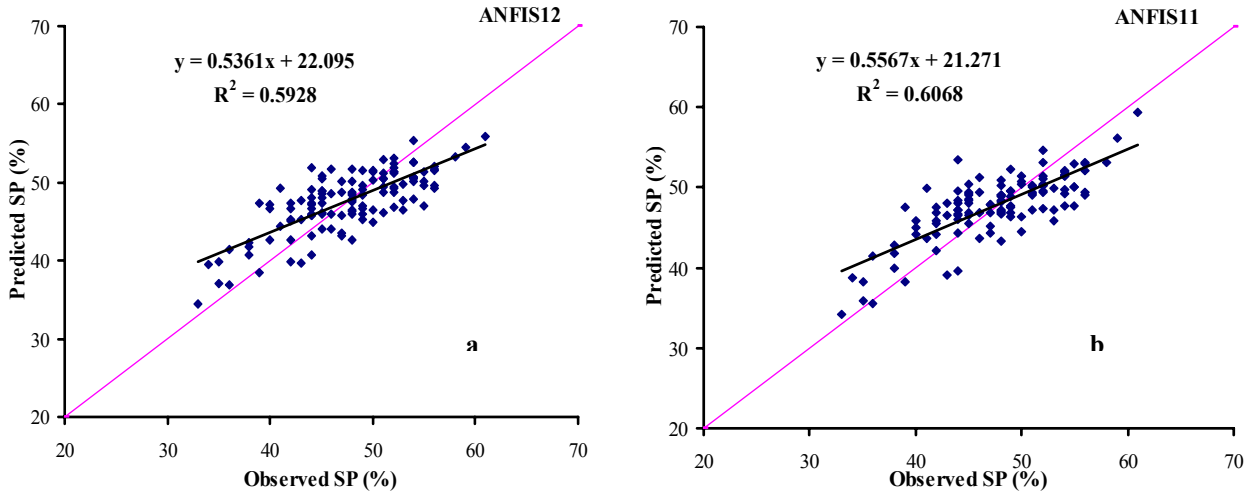

Figure 7. Observed and estimated SP (%) in test period in ANFIS models with 2 sets of inputs: **a)** OC, clay, silt and sand **b)** clay, silt and sand