

Determination of Soil Organic Carbon Variability of Rainfed Crop Land in Semi-arid Region (Neural Network Approach)

Yahya Parvizi (Corresponding author) & Manochehr Gorji

Soil Science engineering Department, University of Tehran, Karaj, Iran

Tel: 98-261-271-3431 E-mail: yparvizi@ut.ac.ir

Mahmoud Omid

Department of Agricultural Machinery, University of Tehran, Karaj, Iran

Mohammad Hossain Mahdian

Agricultural Research, Education and Extension Organization, Iran

Manochehr Amini

Swiss Federal Institute of Aquatic Science and Technology (EAWAG), Duebendorf Switzerland

Abstract

Soil organic carbon (SOC) is a very important component of soil that supports the sustainability and quality in all ecosystems, especially in semi-arid region. This study was conducted to evaluate the effects of 15 different climatic, soil, and geometric factors on the SOC contents in different land use patterns and to determine relative importance of these desired variables for SOC estimation in one of the semi arid watershed zones in the western part of Iran. Feed forward back propagation artificial neural networks (ANN), was used to model and predict SOC. The performance of the model was evaluated using R^2 and MBE values of tested data set. Results showed that 31-2-1 neural networks have highest predictive ability that explains %76 of SOC variability. Neural network models slightly overestimated SOC content, and had higher ability to detect management variables effects on SOC variability. In all ANN structure, management system dominantly controlled SOC variability in rainfed crop land of semi arid condition.

Keywords: CART, Modeling, Neural networks, Soil organic Carbon

1. Introduction

Soil is an important resource that is vital for the sustainability of life in ecosystems, and acts as a buffer to global climatic change (Sparling et al 2006). The soil organic carbon (SOC) is an important component, since it plays a key role in soil fertility and hydrology, contaminants control, and acts as a sink or source of terrestrial C, which affects the concentration of atmospheric CO₂. Atmospheric carbon is increased exponentially at a rate of 1.5% per year. Soil is regarded as an important sink and source for carbon sequestration and modifier of climate change. This can be done by proper land use selection and management practices (Tan and Lal, 2005).

Application of statistical methods include multiple linear regressions and partial least square regression, in SOC estimation, was limited, because of oversimplification, ignorance of complex nonlinear interactions. Another approach in dealing with nonlinear systems is to use non-linear and nonparametric regression methods included regression and nonparametric regression such as artificial intelligence (AI) modeling paradigms such as artificial neural network (ANN). ANN has been successfully used in classification, prediction and recognition problems (McCullagh, 2005, Zhang, 2004). The potential benefits of these methods include greater prediction reliability, cost-efficient estimation and solving complex problems involving nonlinearity and uncertainty. They have been successfully applied in various soil studies (Spencer et al., 2004). These applications include predictions of soil structure, hydraulic properties, pedotransfer functions (Amini et al., 2005, Sarmadian et al., 2009), pedometric use in soil survey (McBratney et al. 2002), environmental correlation of soil spatial variability (Park and Vlek, 2002)

This study was conducted to predict SOC variability at watershed scale across rainfed land use type in a semi-arid condition of Iran by using artificial neural networks (ANNs). Desired variables to predict SOC were consisted of climatic, soil, and topographic factors as physical variables, and tillage, rotation, crop residue

management, ownership, and soil degradation factors, as management parameters. Since the estimation model was data based, selection of appropriate input and output variables was important. Thus, sensitivity analysis on exploratory variables was carried out to select the best input combinations for modeling and estimating SOC.

2. Materials and Methods

2.1 Site description

Merek sub basin, from Karkheh river watershed, with area about 24000 ha, as an experimental site was selected for this study. The elevation of this site ranged from 1450 to 1850 m with annual average precipitation about 500 mm. This area consists of agricultural rainfed lands with area about 14500 ha, dominantly under Wheat and Pea rotation. Climate is semi-arid and cold. Soil temperature and moisture regimes are mesic and xeric. In this site, soil texture is ranged from clay to silt. Soils, in mountains and hill lands were covered by about 25-60% fine and coarse gravel. The pH varied between 7.3 and 7.9, EC ranged from 0.4 to 0.8 (dS/m), and lime content in topsoil was 4-60%.

2.2 Sampling and dataset

A sampling design based on randomized systematic method was performed with considering land use, soil, slope, and aspect maps. In each sampling site, soil samples were taken from topsoil (0–0.3 m depth). Figure 1 shows a general scheme of sampling and sample points. Totally, 245 different soil samples were collected. Soil samples were air dried and then sieved by 2 and 0.5 mm sieve. Soil organic carbon of these samples was determined. Some soil physio-chemical properties included; TNV, sand, silt, and clay percentages and SP percentages were determined.

In addition to soil variables that described above, climatic variables included mean annual temperature (MAT), mean annual rainfall (MAR), potential evapotranspiration (ETP) were recorded and climate types (C_{type}) determined using the Amberger method. Topographic variables (included elevation (Elev), slope (P), and aspect of the terrain of the sampling site) were also measured. Geometric factors such as curvature (Curv), and terrain parameter were derived from DEM that prepared based on digitized contour line map with 20 meter vertical lag apart. The transformed aspect (TA), which aligns the index along a SW-NE axis, for the sites was calculated according to Beers et al. (1966) using the following equation:

$$TA = \cos(45 - aspect) \quad (1)$$

TAP parameter was calculated by multiplying TA by sinus value of slope angle. This parameter was used to incorporate the effects of slope on direct-beam radiation.

To investigate management system effects, 13 raw quantified and three combined scenario indices, as management variables were defined by using collected data from field study. Primary 13 variables included: farm area and ownership, in ha (O_h), as ownership index, manuring (Mn), legegum (L.F) and cereal (C.F.) frequency in rotation period, winter fallow existence (W_f), crop residual pasture, straw harvest (S_h), straw burning, domestic density (D_d), machinery energy consumption (E) ($Mjha^{-1}year^{-1}$), tillage index (T_{index}) (Ferrara and Gersa, 2007) tillage direction index (P_{dir}), and accelerated soil degradation class (Er). Crop residual related variables were selected to define related variable as crop residue management scenario (S_{sen}) by classifying them with K-means clustering. Tillage management scenario index (T_{sen}) was defined by combine tillage related variables. Third management scenario variable was made by combining rotation related variables with K-means clustering. This variable, defined as rotation system (R_{sen}) variable.

2.3 Data selection and preprocessing

Selection of input data was based on theoretical contribution of physical and management variable, expert experiences and accessibility. Inputs were selected from climatic factors, soil physical properties, geometric factors and management system factors. The data set was split into a training set and a testing set. Before simulation, all data sets were standardized by using a linear algorithm software. Range normalization to transform the original data was used from $[x_{min}, x_{max}]$ so that each data point falls in the range $[-0.95, 0.95]$. The scaling formula is:

$$x^* = 1.9 \times \frac{(x - x_{min})}{(x_{max} - x_{min})} - 0.95 \quad (2)$$

Where: x_{min} and x_{max} are the minimum and maximum values that make up the series of data. Data is transformed so that 95% of the data points making up x fall between -0.95 and 0.95 . With this technique, data outliers who will fall above 97.5% or below 2.5% of the average value will not affect the normalization process.

2.4 Development of ANNs

A typical ANN consists of interconnected processing elements included an input layer, one or more hidden layers and an output layer (which provides the answer to the presented pattern). Between input and output layers, there could be several other hidden layers (see Figure 2). The input layer contains the input variables for the network, while output layer contains the desired output system, and the hidden layer often consists of a series of neurons associated with transfer functions.

The total error at the output layer is distributed back to the ANN and the connection weights are adjusted. This process of feed-forward mechanism and back propagation (BP) of errors and weight adjustment is repeated iteratively until convergence in terms of an acceptable level of error is achieved. Several methods for speeding up BP have been used including adding a momentum term or using a variable learning rate. In this paper, gradient descent with momentum (GDM) algorithm was used.

2.5 Sensitivity analysis

Since the SOC estimation model was data based, selection of appropriate input and output variables was important. A sensitivity analysis was performed on the chosen ANN's so that a better understanding of the relative importance of each input on the output could be examined. Thus, sensitivity analysis was carried out to investigate the dynamic behavior of input variables. This was done by imposing steps changes to various inputs and observing their effects on the network output. These responses were used as guides to select appropriate input and output variables that are suitable for model development.

2.6 Performance criteria

For the purpose of evaluation the accuracy of the prediction models, The performance of the models was evaluated by a set of test data using mean square error (MSE), coefficient of determination (R^2) on testing set, between the predicted values and the target (or experimental) values as follows (Omid et al., 2009). Additionally, two different measures including the mean bias error (MBE) and the correlation coefficient (ρ) were used. The MBE is a measure of bias and reveals the overestimation or underestimation. For ANN's evaluation, the testing sets were used to evaluate the effectiveness of techniques for predicting organic carbon.

2.7 Software

To design various neural networks, a commercial software package, NeuroSolutions (version 5.02), was used. The expression used to calculate the MSE is given by NeuroSolutions for Excel (2005). Statistical analysis was done by SPSS 16 and XLSTAT pro-7.5 package.

3. Results and Discussion

Descriptive statistic and variations of the SOC are given in Table 1. The SOC content varied from 0.34 for dry lands with abundant erosion, to 3.72% in low lands and manure applied soils. The correlation coefficients (ρ) between variables are given in Table 2. The correlation between SOC and T_{sen} was positive and with T_{index} and S_{sen} was negative, significantly ($\alpha = 0.050$). Significant correlation between climatic variables and SOC amounts was predictable, theoretically. But, there is insignificant correlation between geometric variables and SOC. On the other hand, comparing to physical variables, correlation coefficient of majority of management variables and SOC were significantly high. This result indicated that management measures eclipsed physical condition effects on SOC variability in rainfed condition of the semi arid regions in study area. TNV content of soils significantly lowered SOC contents, because of biological activity restriction in lime abundant.

3.1 ANN's structure optimization

Finding the optimum number of hidden neurons in the hidden layer is an important step in developing MLP networks. In neural network design, too many hidden units cause over-fitting, while too few hidden units cause under fitting. A summary of findings and best networks architecture is shown in Table 5. Among the different configurations examined, the N-2-1 configuration, exhibited highest accuracy and least error on cross validation data set (MSE = 0.0768). This optimum feature has N variables as input vector, 2 neurons in its hidden layer and 1 neuron as output vector. The performance of this network is shown in figure 3. After evaluating the optimized configuration with the test set, the MSE were obtained 0.107, 0.113 and 0.120, for all inputs, management, and physical, variables, respectively. The corresponding ρ -value was 0.88, 0.83 and 0.63. The MSE values for the ANN's, with different k = number of nodes in hidden layers values and epoch's, presented that when $k=2$ in calibration and validation data sets, the model was not over trained and optimum epochs in validation set equal to 46, 358, and 376, in all, management, and physical input vector models (Table 3). To find the optimum

number of hidden units, the MSE of the network with different inputs combination was plotted against the number of hidden units. Figure 4 presents this plot for ANN with all variables in input layer.

The scatter plot of the measured against predicted SOC, with different combinations and types of exploratory variables, in the test data set is given, respectively, in figure 5 for the ANN models, which was identified as being the best models for predicting SOC.

3.2 Sensitivity analysis

In order to test the hypothesis that not all of the inputs used were required to train the model networks effectively, it was necessary to measure the influence of each input variable on the output. This was done by measuring the mean rate of change of output when a single input was changed by some relatively small amount (0.001). The mean rate of change was determined by testing the model network 594 times using randomly selected input values. Sensitivities, defined as the MSE rate of output changes divided by rate of change of a given input. Sensitivity analysis results are shown in Figure 4. Results indicated that SOC variation was more sensitive to management condition change than physical variables. Changing in range of all management variables, especially tillage related variables, caused significant SOC variation. But among physical variables only aspect related variables (TA and TAP) and climate type affected SOC changes.

3.3 Comparison of ANNs

The test data set was used to evaluate the performance of the different network models for predicting SOC. The results of the research are given in Tables 3, 4, and 5, shows that the ANNs models can predict SOC contents more efficiently by all variables combination. But Spencer et al (2005) showed that ANNs with physical variables had better SOC estimation. On the other hand, ANNs showed higher ability to simulate management effects on SOC variability. This ability is in contract with Somaratne et al (2005) findings. The ANN models could contribute 45% of SOC variability to physical variables. The MBE values indicated that the network models slightly overestimated the SOC. This overestimation was however so small, especially for the network with physical input. The largest RMSE was produced by the neural network with physical predictors.

It seems that there are differences among the ANN models to distinguish kind of predictor variables effects on SOC. The similarity of the different procedures suggests that the relationship between SOC and predictor variables is essentially nonlinear. Significant differences in R^2 and evaluation indices of ANN models executed by different type of exploratory variables indicated that MLP-ANN algorithm can precisely detect interaction effects between physical and management variables.(Wang et al , 2008).

4. Conclusions

The newly developed, artificial neural network, can detect management effects on SOC variability. Management factors especially tillage scenario parameters (included kind of instruments (T_{index}), and tillage frequency), crop residue scenario parameters and also rotation parameters (especially legeum and cereal frequency in rotation period) dominantly determine SOC variability in semi arid conditions of study area. The models tested in this study, using physically based variables included: TAP, TNV, gravel, SP, MAT and AR could account for only up to 40-45% of the variation in SOC in the semi arid conditions of Iran. Analysis of sensitivity accuracy showed that adding more physical variables slightly improved variability prediction and did not significantly improve the modeling results. It seems that for improving prediction power of used methods, management practices, especially tillage, rotation, straw, and grazing management, are essential to be considered.

5. Acknowledgments

The financial support provided by the University of Tehran and Agriculture and Natural Resources Research Centre of Kermanshah Province is gratefully acknowledged.

References

- Amini M., Abbaspour, K.C., Khademi, H., Fathianpour, N., Afyuni, M., & Schulin, R. (2005). Neural network models to predict cation exchange capacity in arid regions of Iran. *European Journal of Soil Science*, 56,551–559.
- Beers, T.W., Dress, P.E., & Wensel, L.C. (1966). Aspect transformation in site productivity research. *Journal of Forestry*, 64,691–692.
- Liu, D., Wang, Z., Zhang, B., Song, K., Li, X., Li, J., Li, F., & Duan, H. (2006). Spatial distribution of soil organic carbon and analysis of related factors in croplands of the black soil region, Northeast China. *Agriculture, Ecosystems and Environment*, 113,73–81.

- McBratney, A.B., Minasny, B., Cattle, S.R., & Vervoot, R.W. (2002). From pedotransfer function to soil inference system. *Geoderma*, 109,41–73.
- McCullagh, J. (2005). A modular neural network architecture for rainfall estimation, *Artificial Intelligence and Applications*, Innsbruck, Austria, 767–772.
- McVay, K., & Rice, C. (2002). Soil organic carbon and the global carbon cycle, Technical Report MF-2548, Kansas State University.
- Omid M, Baharlooei A., & Ahmadi, H. (2009). Modeling drying kinetics of pistachio nuts with multilayer feed-forward neural network. *Drying Technology*, 27(10),1–9.
- Park, S.J. & Vlek, P.L.G. (2002). Environmental correlation of three dimension soil spatial variability: A comparison of three adaptive. *Geoderma*, 109,117–140..
- Sarmadian F., Taghizadeh Mehrjardi, R., & Akbarzadeh, A. (2009). Modeling of some soil properties using artificial neural network and multivariate regression in Gorgan province, north of Iran. *Australian Journal of Basic and Applied Science*, 3(1), 323-329.
- Somaratne S., Seneviratne, G., & Coomaraswamy, U. (2005). Prediction of Soil Organic Carbon across Different Land-use Patterns:A Neural Network Approach. *Soil Science Society Of American Journal*, 69,1580–1589
- Sparling, G.P., Wheeler, D., Wesely, E.T., & Schipper, L.A. (2006). What is soil organic matter worth?. *Journal of Environment Quality*, 35,548–557.
- Spencer M.J., Whitfort T., & McCullagh, J. (2006). Dynamic ensemble approach for estimating organic carbon using computational intelligence. Proceedings of the 2nd IASTED international conference on Advances in computer science and technology. Puerto Vallarta, Mexico
- Tan, Z. & Lal, R. (2005). Carbon sequestration potential estimates with changes in land use and tillage practice in Ohio, USA. *Agriculture, Ecosystems and Environment*, 126:113-121.
- Tan, Z., Lal, R., Smeck, N., & Calhoun, F. (2004). Relationships between surface soil organic carbon pool and site variables. *Geoderma*,121:187–195.
- Wang, L. & Mao, Y. (2008). A novel approach of multiple submodel integration based on decision forest construction. *Modern Applied Science*. 2(2):9-11.
- Zhang, G. (2004). *Neural networks in business forecasting*, IRM Press, Hershey, PA.

Table 1. Descriptive statistics of SOC variable

Descriptive Statistics	Minimum	Maximum	Mean	Variance	Standard deviation	Skewness	Kurtosis	C.V
SOC	0.34	3.67	1.52	0.51	0.71	1.19	1.15	0.47

Table 2. Pearson correlation coefficients (ρ) between SOC and physical and management variables

Physical variables	ρ	Management variables	ρ
Elev.	0.18	Er.	-0.12
P	0.07	O _h	-0.23
TA	-0.03	Mn.	0.35
TAP	-0.14	L.F	0.28
Curv.	-0.14	C.F	0.09
T.N.V	-0.23	W _f	0.05
S.P	0.24	Rsen.	0.25
Gravel	0.14	Pasture	0.34
clay	0.05	S _h	0.17
silt	-0.12	Burn.	0.45
sand	0.03	D _{den}	0.07
M.A.T	-0.19	Ssen.	-0.46
A.R	0.20	E	-0.51
ET	-0.15	T _{index}	-0.52
C.Type	-0.19*	P _{dir}	-0.09
---	---	Tsen.	0.46

*: In bolds, the correlation is significant at $\alpha = 0.05$

Table 3. ANNs performance with best architecture

Inputs		Train	CV	Test	Best Networks	Train	CV
All variables	MSE	0.001	0.118	0.107	Hidden 1 PEs	9	2
					Epoch #	1000	46
	ρ	0.860	0.862	0.883	Final MSE	0.001	0.043
Management variables	MSE	0.094	0.077	0.113	Hidden 1 PEs	10	2
					Epoch #	5000	358
	ρ	0.808	0.814	0.831	Final MSE	0.0018	0.0737
Physical variables	MSE	0.121	0.083	0.120	Hidden 1 PEs	9	2
					Epoch #	594	376
	ρ	0.651	0.376	0.631	Final MSE	0.0123	0.0413

Table 4. Mean Bias Error (MBE) of ANNs results with different input variables

Input variables	MBE		
	All variables	Management variables	Physical variables
	0.025	0.010	0.003

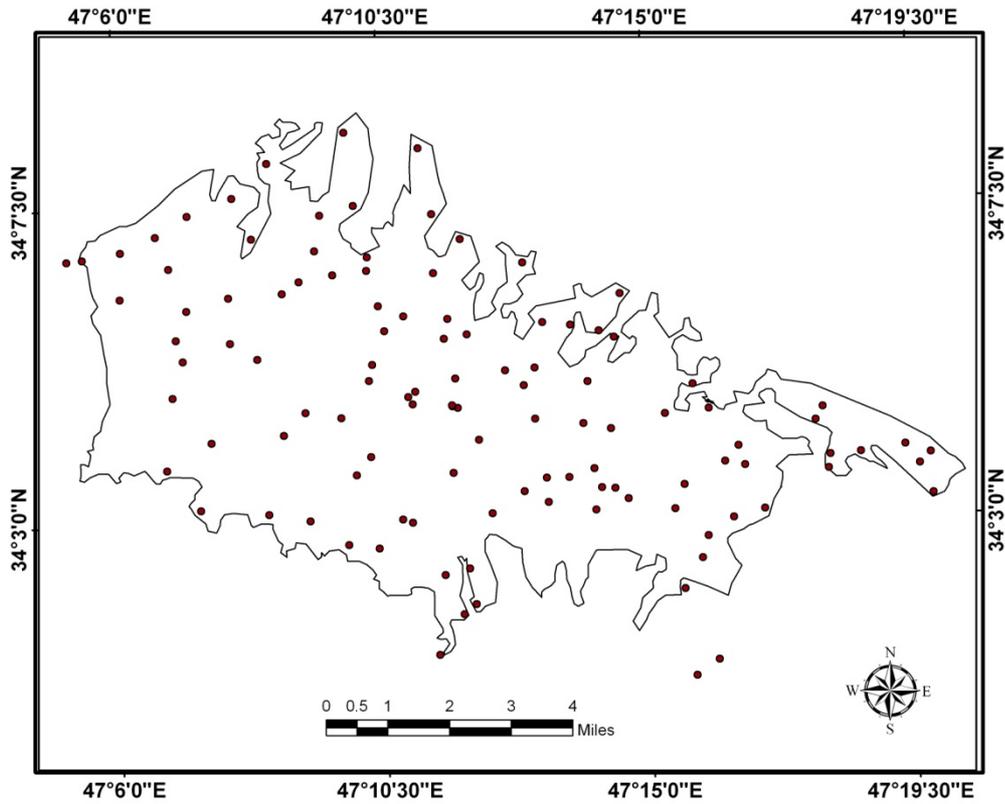


Figure 1. Rainfed land of Merek watershed and sampling point distribution

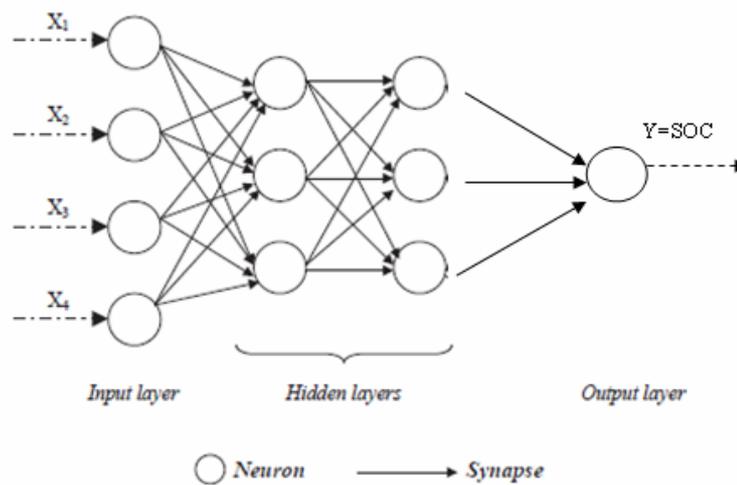


Figure 2. The structure of multi layer feed forward neural network

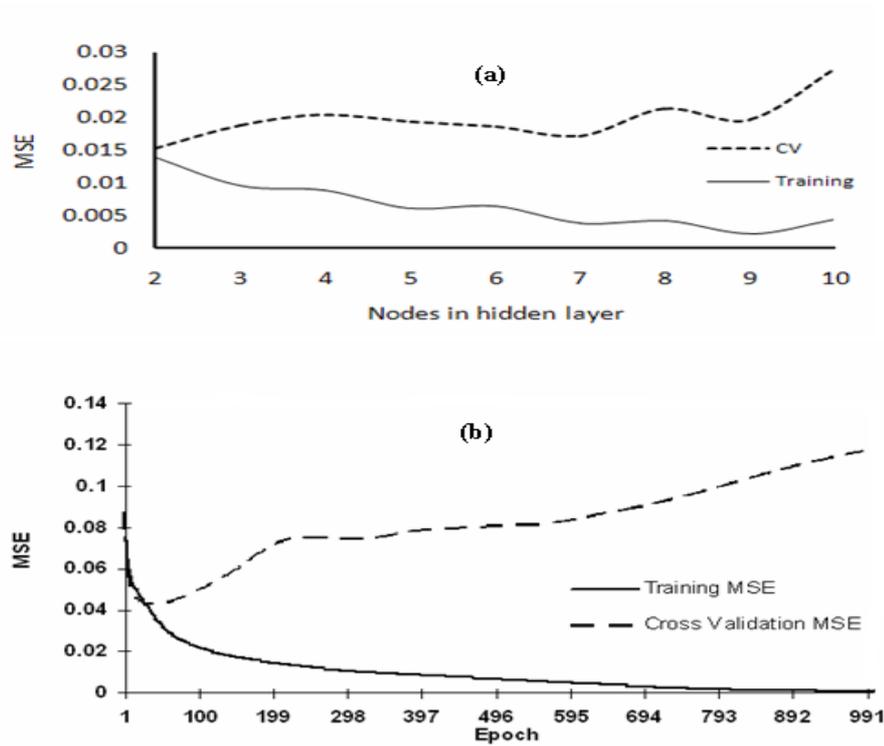


Figure 3. MSE of networks with all variables as inputs versus number of nodes (a) and epochs (b)

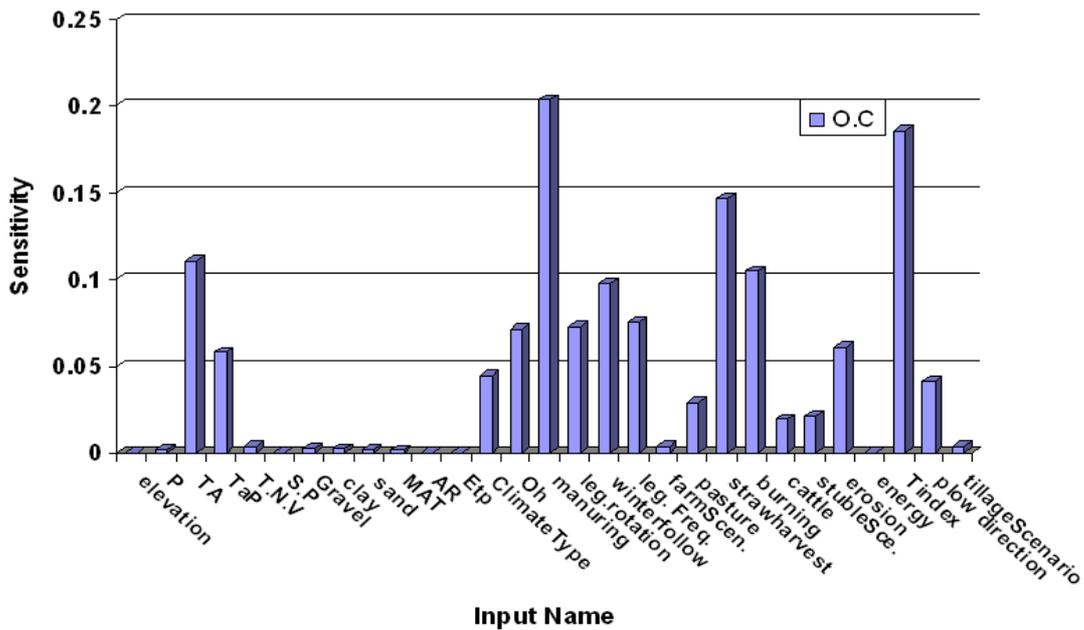


Figure 4. The result of sensitivity analysis about mean of input data for ANN with all variables

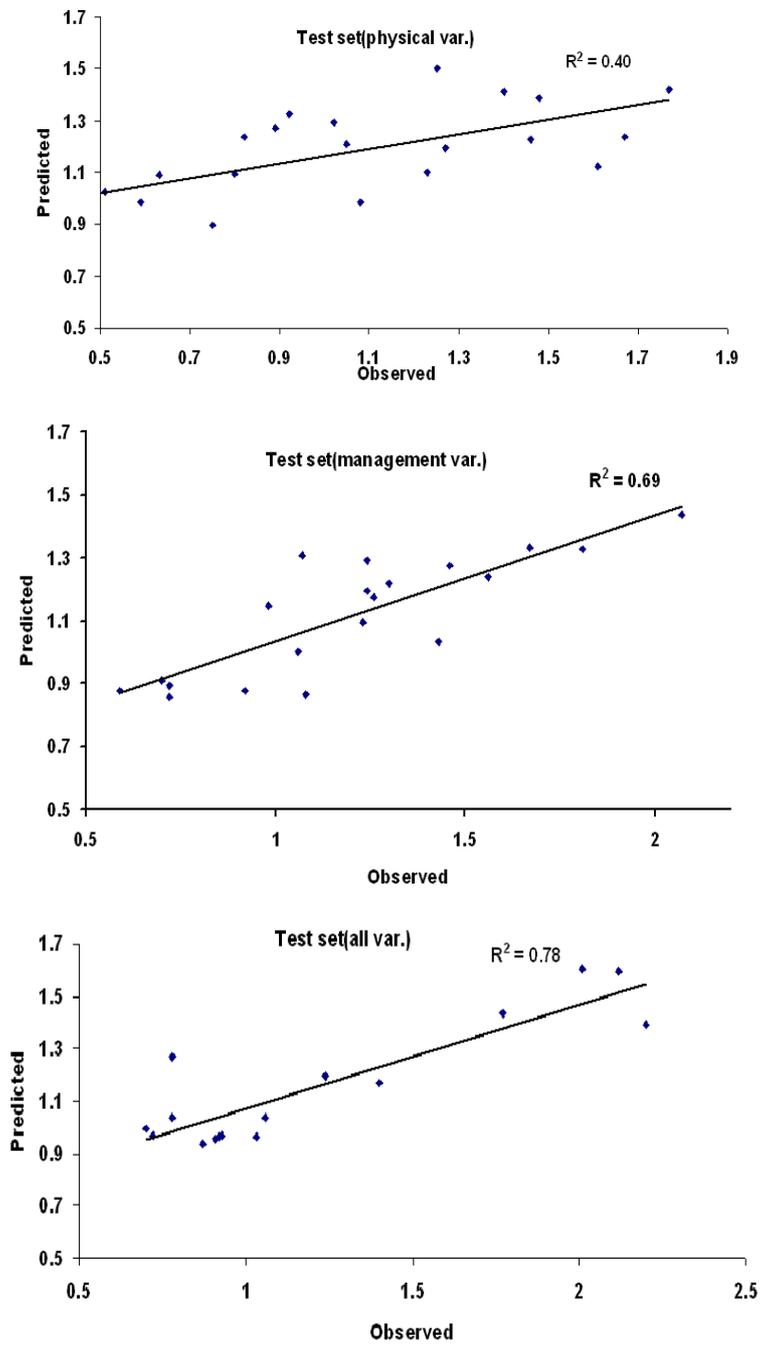


Figure 5. The scatter plot of the measured vs. predicted SOC using the ANNs with different inputs