JOURNAL OF FOREST SCIENCE, *61*, 2015 (10): 448–455

doi: 10.17221/26/2015-JFS

Detection of the effects of management and physical factors on forest soil carbon stock variability in semiarid conditions using parametric and nonparametric methods

Y. Parvizi, M. Heshmati

Department of Soil Conservation and Watershed Management, Agriculture and Natural Resource Research Center of Kermanshah, Kermanshah, Iran

ABSTRACT: Forest soils in western parts of Iran are being degraded by inappropriate management. The soil organic carbon (SOC) stock was dominantly affected by this type of degradation. On the other hand, SOC is an important sink for atmospheric carbon dioxide and can play a key role in global warming. This study was conducted to evaluate the effects of 15 different physical and 8 different management factors on the SOC content and to determine relative importance of these exploratory variables for SOC estimation in a semiarid forest using multiple least-squares regression, tree-based model, and neural network model. Results showed that the CART model with all physical and management variables and 24-2-1 neural networks had the highest predictive ability that explained 81 and 76% of SOC variability, respectively. Neural network models slightly overestimate SOC content. ANNs have a higher ability to detect the effects of management variables on SOC variability and the advantage of CART was to distinguish the effects of physical variables. In both methods the management system dominantly controlled SOC variability in these semiarid forest conditions.

Keywords: soil organic carbon; CART; modelling, neural networks

Soil organic carbon (SOC) plays a very important role in the global C cycle as it is the largest terrestrial C pool (Lal 2008). It has been speculated that improved land management could result in the sequestration of a substantial amount of carbon in soils within several decades and therefore can be an important option in reducing atmospheric CO_2 concentration (Lal 2010). Forest ecosystem C sequestration is of particular interest because, at global scales, forests account for 80–90% of terrestrial plant C and 30–40% of soil C. Forest soils can act as a sink or source of greenhouse gasses by carbon sequestration to mitigate climate changes depending on their management system. This can be done by management and land use activities (Tan, Lal 2005).

The application of statistical methods including multiple linear least squares regression was limited in SOC estimation because of oversimplification and ignorance of complex nonlinear interactions. Collinearity affects the statistical estimation of the parameters as it inflates the variance of at least one of the estimated regression coefficients, and consequently also inflates the estimation of the confidence interval around the predicted values. To avoid these problems, an alternative approach to dealing with nonlinear systems is to use nonlinear and nonparametric methods such as classification and regression tree (CART) algorithm and artificial intelligence (AI) paradigms such as artificial neural network (ANN) (Zhang 2004; McCullagh 2005). They have been successfully applied in various soil studies (McBratney et al. 2002; Park, Vlek 2002; Spencer et al. 2004; Amini et al. 2005; Sarmadian et al. 2009)

Supported by the Agriculture and Natural Resource Research Center of Kermanshah, Project No. 88005.

This study was conducted to predict forest SOC sequestration and degradation cause and effect in semiarid conditions of Iran. The multiple linear regression (MLR) technique as a parametric method was used in this investigation, in comparison with artificial neural networks (ANNs) and classification and regression tree (CART) methods. Exploratory variables to predict SOC consisted of climatic, soil and topographic factors as physical variables and vegetation status and management operations as management parameters. Since the estimation model was data based, the selection of appropriate input and output variables is important. Thus, sensitivity analysis of exploratory variables was carried out to select the best input combinations for modelling and estimating SOC.

MATERIAL AND METHODS

Site description. The research area was located in Kermanshah province, in northern parts of the Karkheh river basin in the west of Iran. An experimental forest site with semiarid conditions and Mediterranean climatic zone and area of about 2,200 ha was selected for this study. The elevation of this site ranged from 1,610 to 2,000 m with annual average precipitation about 700 mm. Soil temperature and moisture regimes are mesic and xeric, respectively. In this site, the soil with clay or silt clay texture and blocky structure is covered by about 25–60% fine and coarse gravel. Forest types consist of *Quercus brantii* with *Crataegus meyer* and *Amygdalus orientalis* with canopy cover about 10 to over 50%.

Sampling and dataset. A sampling design based on a randomized systematic method was performed with considering soil, slope and aspect map. In each sampling site, soil samples were prepared from topsoil (0–0.3 m depth). Fig. 1 shows experimental sites with a general scheme of sampling and sample positions. Soil samples were air-dried and then sieved with a 2 and 0.5 mm sieve. Soil organic carbon of these samples was determined. Some soil physiochemical properties including calcareous percent (through titration with normal NaOH), sand, silt and clay percentages, volumetric gravel percent and saturation percent were determined (Sparks 1996).

In addition to the soil variables described above, climatic variables including mean annual temperature (MAT), mean annual rainfall (AR), potential evapotranspiration (ETP) were recorded and climate types (C_{type}) were determined using the Amberger method. Topographic variables including 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 was prepared based on a digitized contour line map with 20 m vertical lag apart. The transformed aspect (TA), which aligns the index along the SW-NE axis, for the sites was calculated according to Beers et al. (1966) using the Equation 1:

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

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

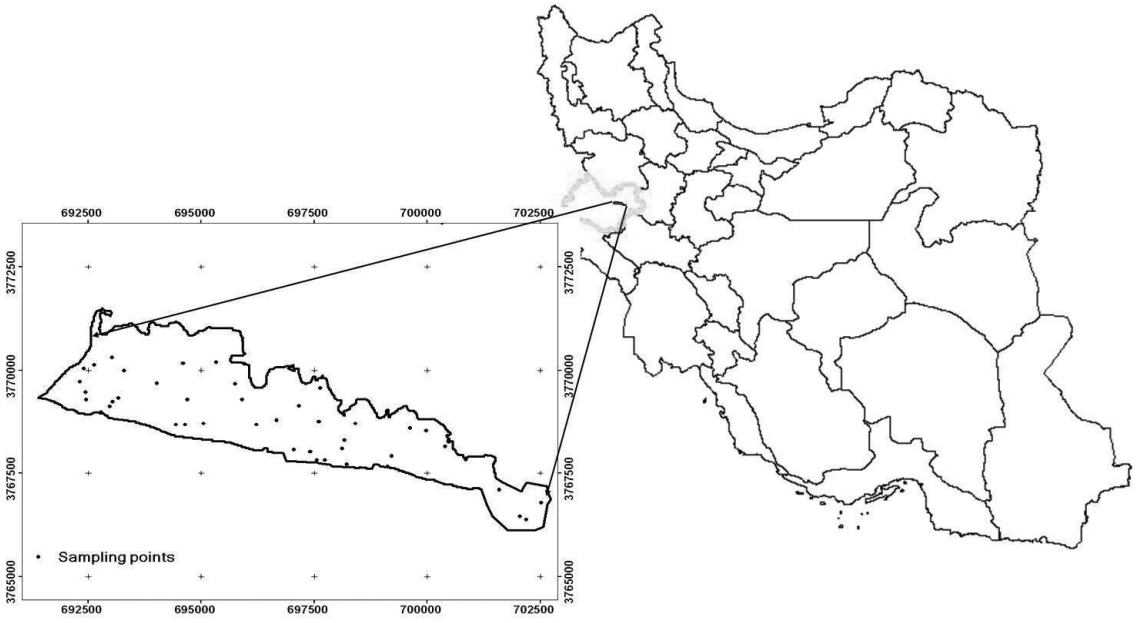

Fig. 1. Forest land of experimental site and sampling point distribution

To investigate land cover and management system effects, by use of information that was collected in a field study, 8 individual quantified indices as management variables were defined. These variables included livestock density (L_d), grazing intensity (G_i) based on a local grading method, invasive plant existence (I_p), deforestation and extensive pruning (D_p), existence of Rosaceae (R_x) [hawthorn (*Crataegus meyer*), *Amygdalus* spp. and *Prunus* spp.], maple (*Acer monspessulanum*) and seedling regeneration stands in vegetation type composition (MS_x), understory density (U_d) in forest layering and accelerated soil degradation status class (Er).

Selection of input data was based on the theoretical contribution of physical and management variables, expert experiences and accessibility. The data set was split into a training set and a testing set. Before simulation, all data sets were standardized by the software using a linear algorithm. 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 Equation 2 is:

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

where:

x_{min}, x_{max} – minimum and maximum values that make up the series of data.

Data is transformed so that 95% of the data points make up x fall between –0.95 and 0.95. With this technique, data outliers that will fall above 97.5% or below 2.5% of the average value will not affect the normalization process (Omid et al. 2009).

For the purpose of comparison two MLR equations were also constructed. The MLR analysis was performed on the training set that was already used to develop the neural networks. The first model was linearly developed by all physical and management exploratory variables. The second model was performed with stepwise regression with entering all 24 inputs and with minimizing the sum of squares of residuals, significance of the remaining exploratory variables was tested at $\alpha < 0.001$. A validation dataset was used to validate the MLR equations whereas a test dataset was used to test the performances of the MLR equations.

A typical ANN consists of interconnected processing elements including 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 (Fig. 2). The input layer contains the input variables for the network while the output layer contains the desired output system and the hidden layer often consists of a series of neurons associated with transfer functions.

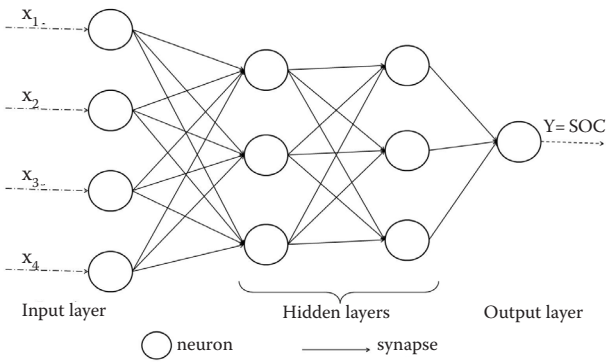

Fig. 2. The structure of the multilayer feed-forward neural network

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 is used.

Since the SOC estimation model was data based, the selection of appropriate input and output variables is important. A sensitivity analysis was performed on the chosen ANNs so that a better understanding of the relative importance of each input on the output could be examined. This was done by imposing step 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 the model development.

The CART methodology represents a unification of all tree-based classification and prediction methods. It transformed the regression tree models in an important nonparametric alternative to the classical methods of regression (Breiman et al. 1984).

The CART algorithm creates a set of questions that consists of all possible questions about the measured variables. Then the splitting criterion was estimated by maximum likelihood and a tree with one node containing all the training data was created. To avoid overtraining, pruning the tree was done by V-fold cross-validation.

The best split is chosen to maximize a splitting criterion. When the impurity measure for a node can be defined, the splitting criterion corresponds to a decrease in impurity. Least-squared deviation (LSD) ($R(t)$) was used as the measure of impurity of a node and was computed using Eq. 3:

Table 1. Descriptive statistics of soil organic carbon (SOC) variables and Pearson's correlation coefficients (ρ) between SOC and physical and management variables

Variable	Descriptive Statistics														
SOC	min	max	mean	s^2	SD	Skewness	Kurtosis	CV	SEM						
	0.15	3.67	1.63	0.82	0.91	0.60	0.06	0.56	0.14						
	Pearson's correlation coefficients (ρ)														
Physical	Elev.	P	TA	TAP	Curv.	TNV	SP	gravel	clay	silt	sand	MAT	AR	ET	C_{type}
	0.52*	–0.08	–0.13	0.00	0.16	**–0.44**	**0.69**	0.16	0.31	**0.51**	**–0.49**	-0.27	**0.32**	–0.24	**–0.42**
Management	D_p	Er	G_i	R_x	U_d	I_p	MU_x	L_d							
	–0.03	**–0.63**	**–0.45**	0.18	**0.47**	**–0.49**	**0.44**	–0.23							

s^2 – variance, SD – standard deviation, CV – coefficient of variation, SEM – Standard Error Means, in bold – significant at α = 0.05, Elev. – elevation, P – slope percent, TA – transformed aspect, TAP – transposed aspect multiplied by slope percent, Curv – curvature, TNV – total neutralizing value, SP – saturation percent, MAT – mean annual temperature, AR – mean annual rainfall, ETP – potential evapotranspiration, C_{type} – climate type; D_p – deforestation and extensive pruning, Er – accelerated soil degradation status class, G_i – grazing intensity, R_x – existence of Rosaceae, U_d – understory density, I_p – invasive plant existence, MU_x – seedling regeneration stands in vegetation type composition in forest layering, L_d – livestock density

$$R(t) = \frac{1}{N_w(t)} \sum_{i \in t} w_i f_i (y_i - \bar{y}(t)^2 \quad (3)$$

where:

$N_w(t)$ – weighted number of cases in node t,

w_i – value of the weighting variable for case i,

f_i – value of the frequency variable,

y_i – value of the response variable,

$y(t)$ – weighted mean for node t.

For the evaluation of the accuracy of 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. Additionally, the mean bias error (MBE) as a measure of bias, and revealing the overestimation or underestimation, was used. For ANN evaluation, the testing sets were used to evaluate the effectiveness of techniques for predicting organic carbon.

To design various neural networks, the commercial software package NeuroSolutions (v. 5.02) (NeuroDimension, Inc., Gainesville, USA) was used. The expression used to calculate the MSE is given by NeuroSolutions for Excel (2005). Statistical analysis and CART algorithm were computed by SPSS 16 (SPSS, Tulsa, USA) and XLSTAT 2010 package (SOFTONIC , Barcelona, Spain).

RESULTS AND DISCUSSION

Descriptive statistics for variables are given in Table 1. The SOC content varied from 0.15 for bad land forest with abundant erosion to 3.72% in undisturbed forests soils**.** The correlation coefficients (ρ) between variables are given in Table 1. The correlation between SOC and AR and elevation was positive and it was negative with sand and calcareous percent, significantly (α = 0.05). On the other hand, compared to physical variables, the correlation coefficient between the majority of management variables and SOC was significantly high. This result indicated that the management system eclipsed the effects of physical conditions on SOC variability in semiarid conditions of forest soil.

SOC estimation by CART algorithm

Three different combinations of exploratory variables were applied to the estimation of SOC content by CART algorithm. Fig. 3 indicates that using the CART model with all input variables with exploring about 80% of SOC variability had the highest efficiency in SOC estimation. CART indicated that management variables, compared with physical agents, had a higher influence on SOC variability in forest land use. This method was able to identify 63% of SOC variability by the source of management factors. The application of CART algorithm in all combinations of predictor variables did not show a bias

Table 2. Evaluation indices of MLR models

Regression model	R^2	RMSE	MSE	MBE	MAE	α
OLS	0.824	0.573	0.329	0.000	0.299	0.003
F stepwise	0.680	0.511	0.261	0.000	0.420	< 0.0001

R^2– coefficient of determination, RMSE – root mean square error, MSE – mean square error, MBE – mean bias error, MAE – mean absolute error, α – significance level

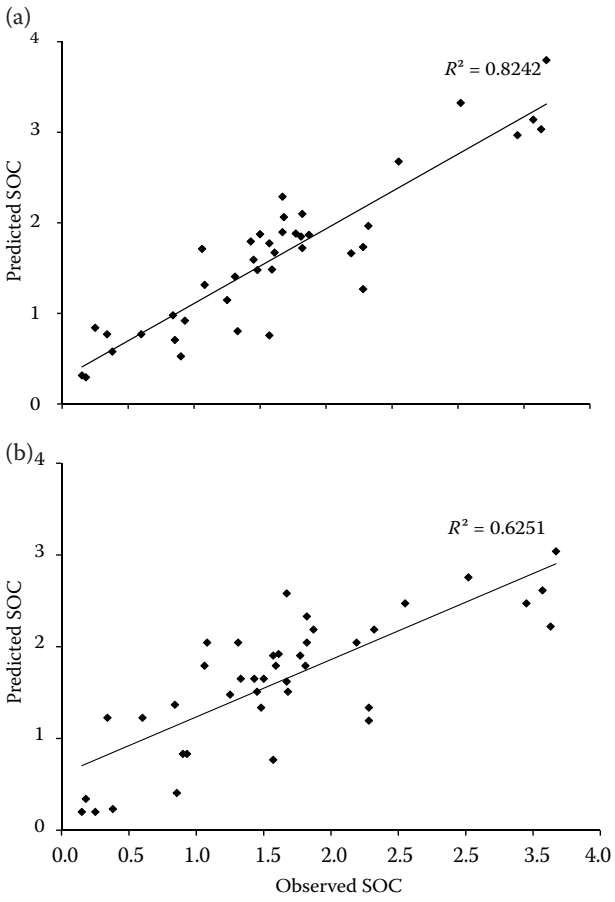

Fig. 3. Scatter plot of observed versus predicted SOC by OLS (a) and Stepwise (b) models

error. But the error estimation was between 0.06 and 0.15% of SOC, and considering the SOC range (0.37–3.72%), this error can be neglected (Table 2). Therefore the CART algorithm could be considered as a good method to estimate SOC content and to determine management effects on it.

ANN structure optimization

Finding the optimum number of hidden neurons in the hidden layer is an important step in developing MLP networks. In a neural network design, too many hidden units cause overfitting, while too few hidden units cause underfitting. A summary of findings and the best network architecture is shown in Table 3. Among the different configurations examined, the N-2-1 configuration exhibited the highest accuracy and the least error on a 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 Fig. 4. After evaluating the optimized configuration with the test set, the MSE values 0.107, 0.113 and 0.120 were obtained, respectively, when inputs included all management and physical variables. The respective ρ was 0.88, 0.83 and 0.63. The MSE values for the ANNs, with different k = number of nodes in hidden layers and epochs, presented that when k = 2 in calibration and validation data sets, the model was not overtrained and optimum epochs in the validation set were equal to 46, 358 and 376 in all management and physical input vector models (Table 4). To find the optimum number of hidden units, the MSE of the network with different input combinations was plotted against the number of hidden units;

Table 3. Mean bias error (MBE) of ANN results with different combinations of input variables

Input variables	All variables	Management variables	Physical variables
	−0.018	−0.009	−0.055

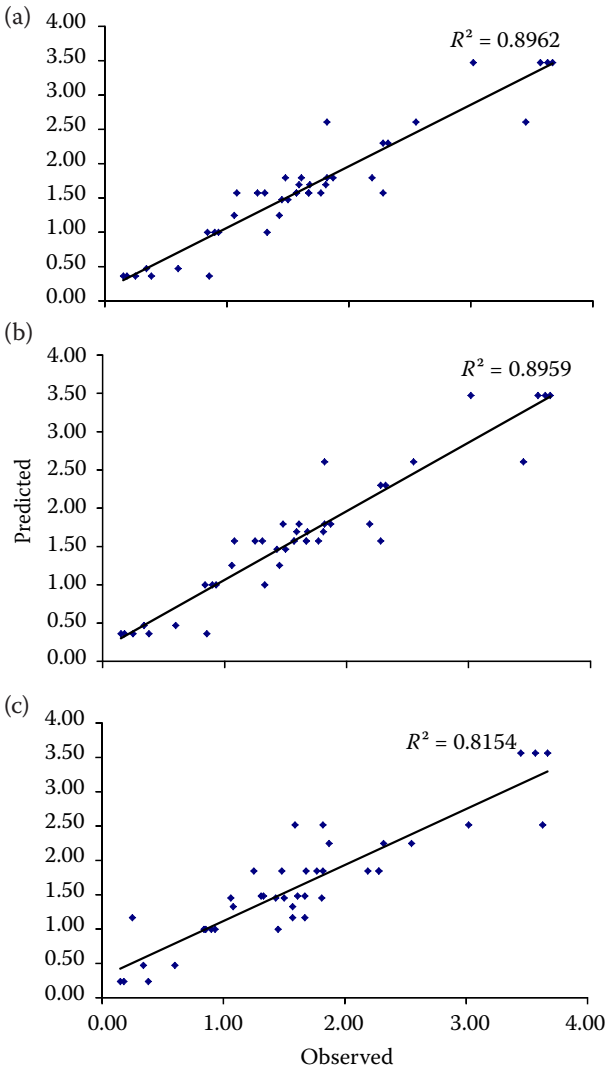

Fig. 4. Plot of predicted vs. measured SOC by CART models with all (a), physical (b) and management (c) variables as predictors

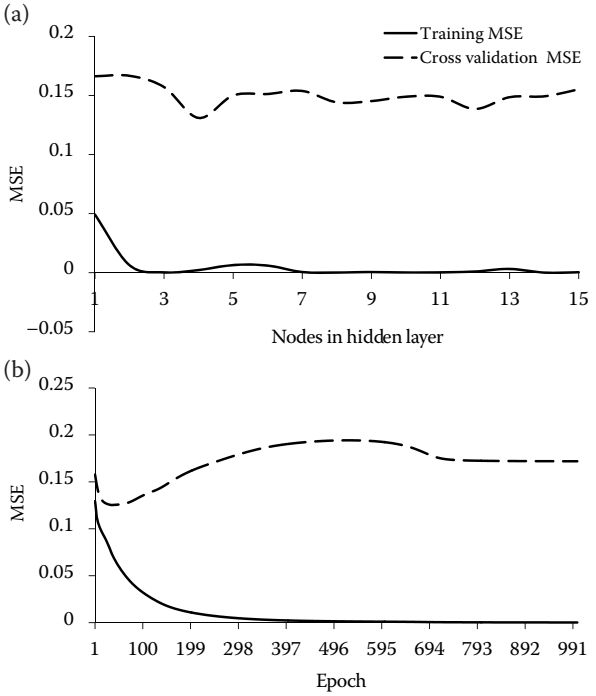

Fig. 5. MSE of networks with all variables as inputs versus number of nodes (a) and epochs (b)

Fig. 4 presents this plot for ANN with all variables in the 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 in Fig. 6 for the ANN model, which was identified as the best model for predicting SOC.

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 each 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 were defined as the MSE rate of change of outputs divided by the rate of change of a given input. Fig. 5 indicates the results of sensitivity analysis. Results indicated that SOC variation was more sensitive to a change of management conditions compared to physical conditions. Changes in the range of all management related variables caused significant SOC variations. But among physical variables only aspect-related variables (TA and TAP) and climate type affected SOC changes.

Table 4. Evaluation indices of CART models with different combinations of input variables

Input variables	RMSE	MBE
All	0.292	0.000
Physical	0.292	0.000
Management	0.389	0.000

RMSE – root mean square error, MBE – mean bias error

Comparison of CART and ANN

The test data set was used to evaluate the performance of different network models for predicting SOC. The results of the research given in Tables 2–5 show that the CART model can predict SOC content more efficiently by combinations of all variables. But Spencer et al. (2005) showed that ANNs had better SOC estimation than CART. On the other hand, ANNs showed a higher ability to

Table 5. ANN performance with the best architecture

Inputs		Train	CV	Test	Best networks	Train	CV
All variables	MSE	0.000	0.125	0.221	hidden 1 PEs	8	4
					epoch no.	1,000	36
	ρ	0.803	0.700	0.943	final MSE	0.000	0.172
Management variables	MSE	0.002	0.120	0.214	hidden 1 PEs	12	2
					epoch no.	1,000	102
	ρ	0.694	0.842	0.820	final MSE	0.001	0.140
Physical variables	MSE	0.000	0.170	0.202	hidden 1 PEs	14	5
					epoch no.	1,000	35
	ρ	0.826	0.762	0.946	final MSE	0.000	0.270

CV – coefficient of variation, MSE – mean square error, ρ – correlation coefficient, hidden 1 PEs – number of processing elements in hidden layer 1, epoch no. – number of iterations over the data set in order to train the neural network, final MSE – final mean square of error

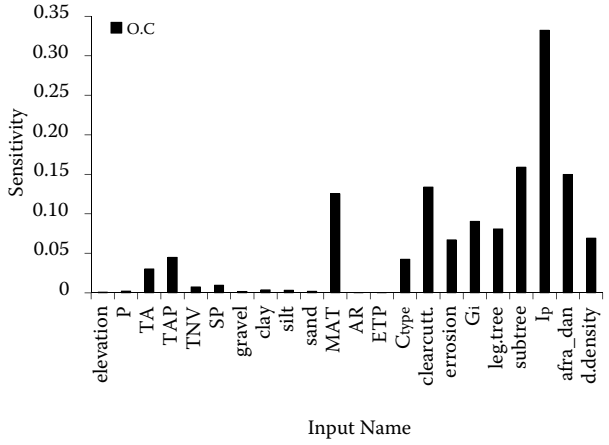

Fig. 6. The result of sensitivity analysis about mean of input data for ANN with all variables (O.C – soil organic carbon)

P – slope percent, TA – transformed aspect, TAP – transposed aspect multiplied by slope percent, TNV – total neutralizing value, SP – saturation percent, MAT – mean annual temperature, AR – mean annual rainfall, ETP – potential evapotranspiration, C_{type} – climate type; clearcutt. – deforestation or forest clear cutting, G_i – grazing intensity, leg. tree – existence of leguminouse trees geniuse in forest floor, subtree – understory density, I_p – invasive plant existence, afra_dan – maple (*Acer monspessulanum*) and seedling regeneration stands in vegetation type composition, d.density – livestock density

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 (Fig. 7). The MBE values indicated that the network models slightly overestimated the SOC but bias errors in CART methods were zero. This overestimation was however so small, especially for the network with physical inputs. The smallest RMSE was produced by the CART with all variables, while the largest RMSE was produced by the network with physical predictors.

It seems that there is a difference between the neural network models and the CART models in distinguishing the kind of predictor variables effects on SOC. The similarity of the two procedures suggests that the relationship between SOC and predictor variables is essentially nonlinear. Significant differences in R^2 and evaluation indices of CART models executed by different type of exploratory variables indicated that the CART algorithm can precisely detect interaction effects between physical and management variables. The only superiority of ANN structures in comparison with CART models mostly contributed to the ability of ANNs to derive management factor impacts

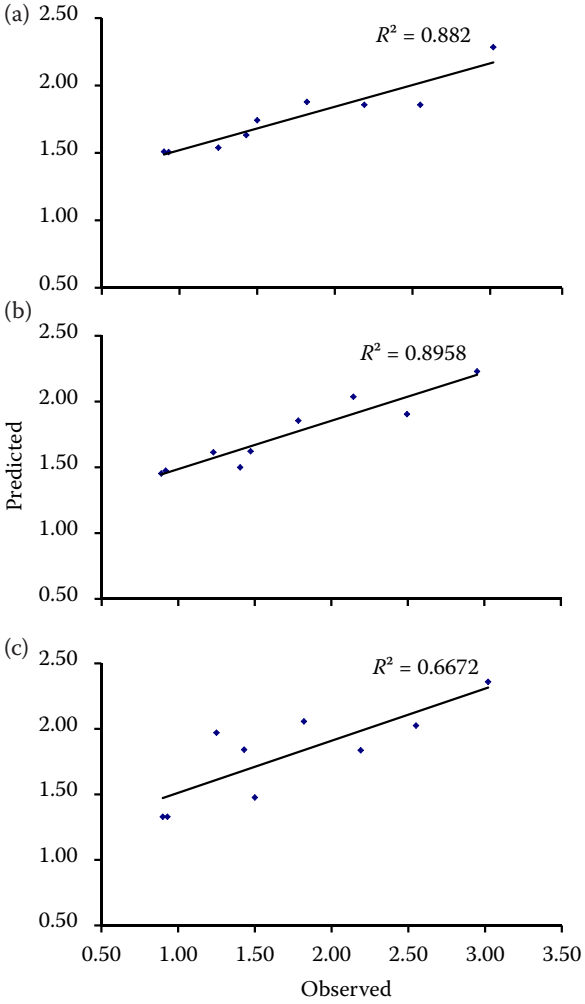

Fig. 7. The scatter plot of the measured vs. predicted SOC using the ANNs with with all (a), physical (b) and management (c) variables as predictors

on SOC content. But CART can detect the effect of physical variables and interaction effects of variables precisely (Wang et al. 2008).

CONCLUSIONS

The newly developed neural network can detect management effects on SOC variability better than the CART models. But CART models can explore nonlinearity and interaction between variables more accurately than ANNs. Especially management factors dominantly determine SOC variability in these semiarid conditions. The models (CART and ANNs) tested in this study, using physically based variables, including TAP, TNV, gravel, S.P, MAT and AR could account for only up to 40–45% of the variation in SOC in semiarid conditions of Iran. The analysis of sensitivity accuracy showed that adding more physical variables can slightly improve variability prediction and did not significantly improve the modelling

results. It seemed we must bring to our attention that to improve the predictability power of our methods considering management factors specially grazing management should be attempted. We hope to include these in our future works.

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.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984): Classification and Regression Trees. Belmont, Wadsworth International Group: 368.

Lal R. (2008): The role of soil organic matter in the global carbon cycle. Soil and Environmental Pollution, 116: 353–36.

Lal R. (2010): Managing soils and ecosystems for mitigating anthropogenic carbon emissions and advancing global food security. BioScience, 60: 708–721.

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): 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. Manhattan, Kansas State University: 216–289.

Omid M, Baharlooei A., Ahmadi H. (2009): Modeling drying kinetics of pistachio nuts with multilayer feed-forward neural network. Drying Technology, 27: 1–9.

Park S.J., Vlek P.L.G. (2002): Environmental correlation of three dimensional soil spatial variability: A comparison of three adaptive. Geoderma, 109: 117–140.

Sarmadian F., Taghizadeh R., 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: 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

Sparks D.L. (1996): Methods of Soil Analysis. Soil Science Society of America Book Series, Vol. 5. Madison, Soil Science Society of America: 1264.

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. In: Proceedings of the 2nd IASTED International Conference on Advances in Computer Science and Technology. Puerto Vallarta, Jan 23–25, 2006, 186–192.

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: 9–11.

Zhang G. (2004): Neural Networks in Business Forecasting. Hershey, IRM Press: 58–116.

Received for publication March 7, 2015
Accepted after corrections October 8, 2015

Corresponding author:

Dr. Yahya Parvizi, Agriculture and Natural Resource Research Center of Kermanshah, Jam-e-jam Street, Kermanshah, P. O. Box: 6715848333, Iran; e-mail: yparvizi1360@gmail.com