International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 14 (2015) pp 34346-34352

A Novel Artificial Neural Network Model for Prediction of Soil Total Porosity

Ashok Kumar D*, Kannathasan N#
*Department of Computer Science, Government Arts College, Tiruchirapalli-620 022, Tamil Nadu, India
akudaiyar@yahoo.com
#Department of Computer Science, Kanchi Mamunivar Centre for Post Graduate Studies, Puducherry-605008, India
nkthasan@gmail.com

Abstract:
Background/Objectives: Many mapping of contaminated soil sites and the resulting cleanup are time consuming and expensive tasks, requiring extensive amounts of geology, hydrology, chemistry, GIS in environmental contamination and data mining and pattern recognition techniques. Artificial Neural Networks (ANN) has been inspired by biological neural networks, and are popular tools in the application of classification, prediction and recognition based problems. **Methods/Statistical Analysis:** This work proposed a novel Artificial Neural Network model with a Feed-forward Back-Propagation has been designed to predict soil Total Porosity. Soil samples were collected from International Soil Reference and Information Centre. The network structure used the default Levenberg–Marquardt algorithm (TRAINLM) for training, gradient descent with momentum weight/bias learning function (LEARNGDM) for learning and Mean Square Error (MSE) function for Performance estimation. **Findings:** The ANN model outputs for training, validation and testing in order to determine the relationship between the outputs of the network and the targets. After the network was trained, validated and tested, the generated model can be used to predict the parameter Total Porosity through new input pH. It was found that experimental analysis of Total Porosity was close to the predicted data calculated from the configuration and this confirms the developed ANN model is suitable for predicting the salinity concentrations. **Applications/Improvements:** Development of novel methods for classification in data mining and pattern recognition like genetic based, and hybrid based is required to find out the optimal number of clusters for soil sites based on soil texture with Total Porosity.

Keywords- Artificial Neural Network, Prediction, Total Porosity, Feed-forward Back-Propagation, Levenberg-Marquardt.

1. Introduction

An important factor influencing the productivity of the various ecosystems of the earth is the nature of their soils. Soils are vital for the existence of many forms of life that have evolved on our planet. Few feet of land from the surface is referred to as soil and it acts as a natural filter to sieve many substances that mix with the water.

But water is the medium that transports some contaminants into the groundwater from the soil surface (Fig.1).

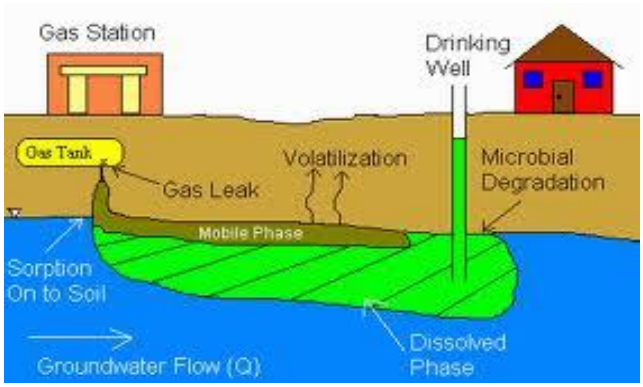


Fig .1.Soil Water Contamination

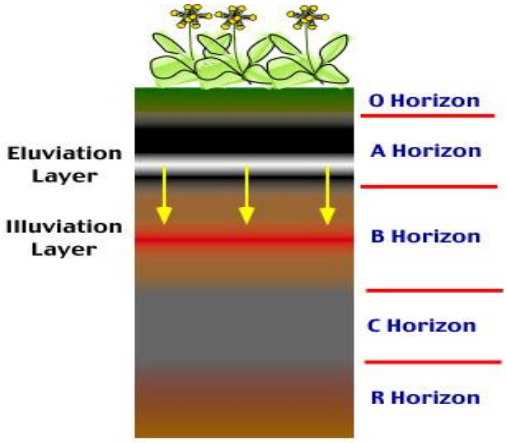


Fig.2.Soil Profile

Soils of various regions have a distinct profile or sequence of horizontal layers. Chemical weathering, eluviations, illuviation, and organic decomposition are the causes for varied nature of layers of soil. A typical soil has five layers viz. O, A, B, C, and R horizons. The top layer of the most of the soils is O horizon and it is composed mainly of plant litter at various stages of decomposition and humus. Below the O layer A horizon is found. This layer is composed principally of mineral particles and has two characters: (i) it is the layer in which humus and other organic materials are mixed with mineral particles and (ii) it is a zone of translocation from which eluviations has removed finer particles and soluble substances, both of which may be deposited at lower layers. Therefore the A horizon is dark in

color and usually light in texture and porous. The A horizon is commonly differentiated into a darker upper horizon or organic accumulation, and a lower horizon showing loss of material by eluviations. The B horizon is a third layer consists of mineral soil layer which is strongly influenced by illuviation. Consequently, this layer receives material eluviated from the A horizon. Due to enrichment of clay particles the B horizon also has a higher bulk density than the A horizon. The B horizon may be colored by iron oxides and aluminum oxides or by calcium carbonate illuviated from the A horizon. The fourth layer, the C horizon is composed of weathered parent material. The texture of this material can be quite variable with particles ranging in size from clay to boulders. The C horizon is also not significantly influenced by the pedogenic processes, translocation, and/or organic modifications. The lower most layers in a typical soil profile is called the R horizon. This soil layer simply consists of unweather bedrock (Fig.2).

The soil properties such as texture, porosity, specific yield depend on the total volume of groundwater recharge, water storage and discharge, also the extent of groundwater contamination[1].

Soil is a mixture of three soil parts: sand, silt, clay. Classification of these parts is based on grain size. The following Table.1 shows the soil part and corresponding diameter of the parts. The relative proportion of soil parts in a particular soil determines its soil texture.

Table.1. USDA Particle Size and Porosity ranges for Sand, Silt and Clay

Name of separate	Diameter range (millimeters)	Porosity (%)
Sand	2.0 – 0.05	25-50
Silt	0.05 – 0.002	35-50
Clay	Less than 0.002	33-60

The ratio of different-sized mineral particles in soil is the basis for soil texture. Clay, silt, or sand particles are considered as mineral particles. Clay particles hold water molecules due to electric force on them and make it suitable for growing crops. Clay rich soil is resistant to erosion. Silt particles too are small enough to hold water in the soil but the silt which is exposed washes away easily, depleting nutrients in the soil. Water passes through sandy soil very quickly and does not hold enough moisture to support plant growth. Exposed sandy soils are much subjected to erosion if the angle of the land, *i.e.* slope, is high. The soil textural triangle gives the textural name, which based on the percentage of sand, silt, and clay within the soil sample (Fig.3). The triangle is divided into 10-percent portions of clay, silt and sand. The summation of the three percentages must total 100 percent[5].

Soil system is composed of air, water, dead organic matter, and various types of living organisms (Fig.4). Various factors such as organisms living in that soil, climate, topography, parent material and time influence the formation of soil. The sum of mineral particles alone does not constitute a true soil but it is influenced, modified, and supplemented by organisms living in that soil.

Fig .3. USDA 12 Basic Soil Textural Triangle

Humus is the biochemical substance that makes the upper layers of the soil usually looks dark, colored dark brown to black. Organic activity is usually profuse in the upper layers of the soil. For instance, one cubic centimeter of soil can be the home of more than a million bacteria.

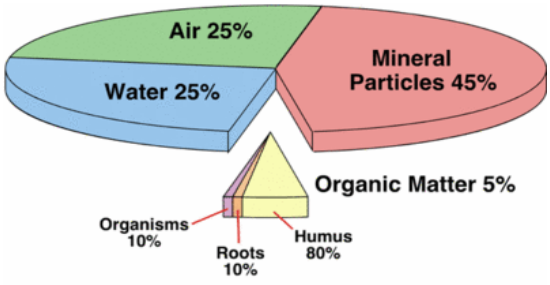


Fig. 4. Components of Soil

The Porosity of a soil is decided by shape and spatial arrangement of soil particles. It is the air space or void space between soil particles. Infiltration, ground water movement, and storage occur in these void spaces. Porosity of soil typically decreases as pH value increases and also particle size increases because of soil contamination or soil pollution (Fig.5).

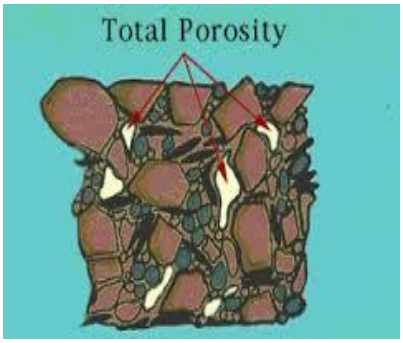


Fig.5. Total Porosity of Soil

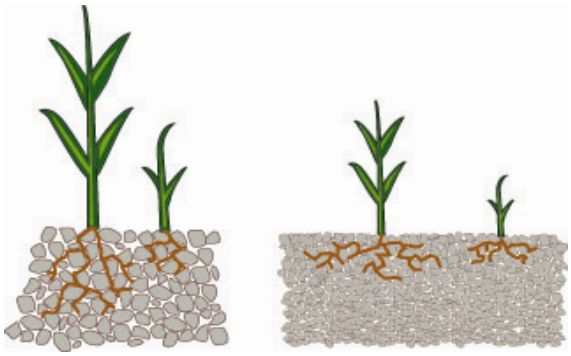
Bulk density is an important indicator of soil compaction and can be calculated as the dry weight of soil divided by its volume and is expressed in g/cm^3. This volume is the sum of volume of soil particles and the volume of pores among soil particles. Structural support, water and solute

movement, and soil aeration are dependent factors of bulk density. Bulk densities above threshold level indicate unsupportive for plant growth (Table.2). This parameter is also used to convert between volume and weight of soil; also it is used to express physical, chemical and biological measurements of soil on a volumetric basis for soil quality assessment. This increases the validity of comparisons by removing error associated with differences in soil density at time of sampling.

The increased porosity of soil is due to mixing and movement of air and water from the soil top to lower layer (Fig.6) where roots of plants occupying (rhizosphere). Increased volume of air and water available to roots has a positive effect on plant productivity. Earthworms and larval and adult insect stages also play a crucial role to produce most of the humus found in a soil through the incomplete digestion of organic matter. Movement of water to lower layers of the soil causes both mechanical and chemical translocations of organic matter.

Table.2. General Relationship of soil bulk density to root growth based on soil textural

Type of Mineral Particle	Ideal bulk densities for Plant Growth (g/cm^3)	Bulk densities that restrict root growth (g/cm^3)
Sand	<1.60	>1.80
Silt	<1.40	>1.65
Clay	<1.10	>1.47



(a) (b)

Fig.6. (a) Soil with good structure (pH <=7) (b) Soil with poor and dense structure (pH >7)

Soil is polluted by various anthropogenic activities such as addition of industrial wastes, pesticides, fertilizers, domestic sewage etc. to soil. Poly nuclear aromatic hydrocarbons and petroleum hydrocarbons and heavy metals are the most common contaminants of soil. Urbanization and industrialization are the major contributors of soil pollution. It leads to numerous health hazards and the chemicals reach human being through contaminated water and plants.

Artificial Neural Network (ANN) is a software system that imitates the neural networks of the brain of man. Neural networks are powerful tools that have the ability to identify underlying highly complex relationships from input-output data only. The study indicates that ANN is efficient in simulating the complicated phenomena during the training process and uses them to simulate the results for new inputs. The technique is capable of dealing with uncertainties in the inputs and can extract information from incomplete or contradictory data sets.

2. Literature Survey

In summary of related 348 papers and reports, ANNs have been applied in solving problems in food quality and safety (35.34%), crop (22.7%), soil and water (14.37%), precision agriculture (6.61%), food processing (2.3%), greenhouse control (2.01%), agriculture vehicle control (1.15%), agricultural pollution (1.15%), agricultural biology (1.15%) and others (2.3%) such as bio energy and agricultural facilities. These ANN applications have been created mainly through classification (45.11%), modeling and prediction (43.97%), control (4.02%) and simulation (2.59%), parameter estimation (2.01%), detection (1.15%), data clustering (0.57%), optimization (0.29%) and data fusion (0.29%) as well.

In summary of 136 related papers and reports, FL has been applied in solving problems in crop management (17%), soil and water (16%), food quality and safety (14%), animal health and behavior (10%), agricultural vehicle control (8%), precision agriculture (7%),greenhouse control(7%), agricultural machinery (4%), food processing (4%), air quality and pollution (3%), agricultural facilities (2%) and others (6%) such as natural resources management and agricultural product design. These applications have been created through FL mainly by control (28%), modeling and prediction (24%), classification (24%), fuzzy clustering (9%), rule-based Inference (7%), multi sensor data fusion (4%), optimization (1%) and others (3%) such as thresholding and pattern inference.

Based on a summary of 83 papers and reports, GAs have been applied in solving problems in crop management (31%), water management (27%), food quality and safety (11%), food processing (6%), precision agriculture (4%), agricultural biology (4%), agricultural machinery (2%), agricultural facilities (2%), animal behavior (2%), and others (11%) such as agricultural vehicle, robotics, and pollution. GAs are basically an optimization and search method. The applications for optimization take the largest portion of the total, 66%. GAs have been also used to assist with modeling and prediction (18%), classification (12%), control (2%), data clustering (1%) and value thresholding (1%).

3. Materials and Methods

Classification and Prediction are two modes of data analysis that can be used to extract the models that describe important data classes or to predict the future data trends. Classification predicts categorical labels. Prediction models continuous valued functions.

Data classification is a process of two-steps:

Model Construction: describing a set of predetermined classes.

- Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute.

- The set of tuples used for model construction is a training set.
- The model is represented as classification rules, decision trees, or mathematical formulae.

Model Usage: for classifying future on unknown objects

- The known label of test sample is compared with the classified result from the model.
- Accuracy rate is the percentage of test set samples that are correctly classified by the model.
- Test set is an independent of a training set.
- If the accuracy is acceptable, use the model to classify data tuples whose class labels are unknown.

Prediction: This data analysis task of numeric prediction, where the model constructed predicts a continuous-valued function, or ordered value, as opposed to categorical label. This model is a predictor. Regression analysis is a statistical methodology which is most often used for numeric prediction[6].

Learning is broadly classified into two types: Supervised Learning and Unsupervised Learning (clustering). Supervised Learning: The training data (observations, measurements, etc.) are accompanied by the labels indicating the class of the observations. New data is classified based on the training set. Unsupervised Learning (Clustering): The class label of training data is unknown. Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data[7, 8,9].

3.1 ISRIC-WISE Soil Profile Data

The test dataset consists of 705 samples of particular region as in Table.3. Using for this research work collected from World Soil Information – ISRIC (International Soil Reference and Information Centre). Version 3.1 of the ISRIC-WISE database (WISE3-World Inventory of Soil Emission Potentials) was compiled from a wide range of soil profile data collected by many soil professionals worldwide. All profiles have been harmonized with respect to the original Legend (1974) and Revised Legend (1988) of FAO-UNESCO[4].

Table. 3. List of soil variables, their abbreviations and units of measurement

Abbreviation	Description	Units
PHH_2O	Soil reaction in water	pH units
SAND	Sand	%(mass)
SILT	Silt	%(mass)
CLAY	Clay	%(mass)
TOTPOR	Total Porosity	%

3.2 Classification Rules Model

SI. No.	Major Textural Classes	Relative % of Sand, Silt, Clay
1.	Sand	Must contain Sand> 85% and % of silt plus 1.5 times the % of clay shall not exceed 15 and between 25 and 50 of Total Porosity and pH value between 5.8 to 7
2.	Loamy Sand	A) Upper Limit B) Must contain 85 to 90 % of sand and the % of silt plus 1.5 times the % of clay is not less than 15% C) Lower Limit: Must contain 70 to 85% sand and the % of silt plus twice the % of clay does not exceed 30 D) pH value between 5.8 and 7 and Total Porosity between 25 and 50
3.	Sandy Loam	A) Contains 20% or less clay and Total Porosity between 25 to 50 and the pH value between 5.8 and 7 B) The % of silt plus twice the % of clay exceeds 30 and has 52% or more sand , or C) Contains <7% clay, <50% silt, and between 43 and 52% sand
4.	Loam	Contains 7 to 27% clay, 28 to 50% silt, <53% sand and Total Porosity between 37.5 to 46.5 and pH value between 6 to 6.5
5.	Silty Loam	A) Contains 50% or more silt and 12 to 27% clay, or B) Contains 50% to 80% silt and <12% of clay C) Contains 6 to 7 pH value and Total Porosity value between 35 to 50
6.	Silt	Contains 80% or more silt and <12% clay and Total Porosity between 35 to 50 and pH value between 6 to 7
7.	Sandy Clay Loam	Contains 20 to 35% clay, <28% silt, and 45% or more sand and Total Porosity between 37.5 to 46.5 and pH value between 6 to 6.5
8.	Clay Loam	Contains 27 to 40% clay and 20 to 45% sand and Total Porosity between 37.5 to 46.5 and pH value between 6 to 6.5
9.	Silty Clay Loam	Contains 27 to 40% clay and <20% or more sand and Total Porosity value between 37.5 to 46.5 and the pH value between 6 to 7
10.	Sandy Clay	Contains 35% or more clay and 45% or more sand and Total Porosity value between 37.5 to 42.5 and pH value between 6 to 6.5
11.	Silty clay	Contains 40% or more clay and 40% or more silt and Total Porosity between 42.5 to 46.5 and pH value between 6.1 to 6.5
12.	Clay	Contains 40% or more clay, <45% sand, and <40% silt and Total Porosity from 33 to 60 and pH value from 6 to 7

4. ANN for Prediction of Soil Total Porosity

Artificial Neural Networks are generally presented as systems of interconnected nodes which can compute values from inputs. Simple artificial nodes are class of statistical models could be called ‘‘neural’’ if they possess the following characteristics: consist of sets of adaptive weights, i.e. numerical parameters that are tuned by a learning algorithm, and are capable of approximating non-linear functions of their inputs. The adaptive weights are conceptually connection strengths between neurons, which are activated during training and prediction. The network structure is composed of a set of neurons connected by links and organized in number of layers. Each layer is fully interconnected to the preceding layer by weights. Initial suggested weights are progressively adjusted during the training process by comparing predicted outputs with measured data (targets). The computation of network weights and biases is known as ‘‘training step’’. The objective of the back-propagation training algorithm is to find the optimal weights by minimizing the Mean Square Error (MSE) of the output values. Simultaneously, learning functions are used to update the layer’s weight and bias. This procedure is completed in the ‘‘validation step’’, where the network is improved to avoid data over-fitting. After that a set of data is randomly used to examine the network generalization, i.e. the ‘‘test step’’[2].

Each scalar input (p) is multiplied by a scalar weight (W), and then added to a scalar bias (b), resulted in the net input (n= Wp+ b). Finally, the result is passed through the transfer function f, which gets the neuron’s output a, where a= f (Wp+b) (Fig.7). Those functions can be (i) linear transfer functions: It is purely a linear function(y=mx+c) and it is also called as PURELIN transfer function. It is used in the final layer to find a linear approximation to a nonlinear function, or (ii) sigmoid transfer functions: used in the hidden layers to generate the output between 0 and 1 and also known as LOGSIG transfer function and between -1 and 1 for Tan-Sigmoid or TANSIG, even if the input data have infinity values[3].

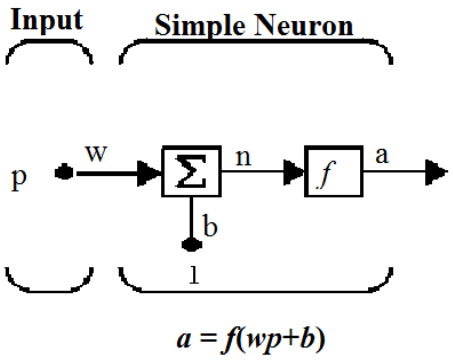


Fig.7. Simple Neuron

A Feed Forward Neural Network (FFNN) with back-propagation training algorithm was applied to correlate the relation between input alkalinity expressed in (pH) and output salinity expressed in Total Porosity. The ANN configuration was identified based on research and through conducting several trials until reaching the best regression results with no over-fitting (Fig. 8). The network properties were as follows:

Input data: pH. Target data: Total Porosity with Fuzzy Logic. Network type: Feed-forward back-propagation. Training function: Levenberg–Marquardt algorithm (TRAINLM). Adaptation learning function: Gradient descent with momentum weight/bias learning function (LEARNGDM). Performance function: Mean Square Error (MSE).

Number of layers: 2 (layer-1: Hidden layer-five neurons and TANSIG transfer function; layer-2: output layer - PURELIN transfer function). Data records were randomly divided into three subsets, i.e. training: 70%, validation: 15% and test: 15%.

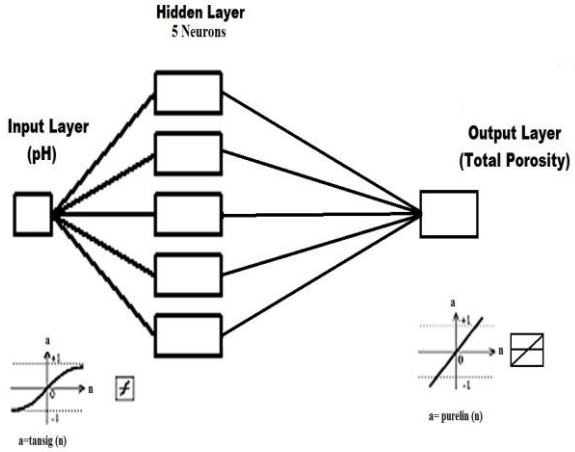


Fig. 8. Artificial Neural Network (ANN) configuration applied to predict Total Porosity value

Analysis of the Soil samples is shown in Table 4.

Table.4. Measured pH ranges and corresponding average values of Total Porosity

Texture	Ranges of pH Values	Average value of Total Porosity (%)
SAND	5.8-7	37.5
SILT	6-7	42.5
CLAY	6-7	46.5

During training, each neuron in the layer adjusts its weight vector toward the closest group of input vectors (Fig.9). The magnitude of the gradient and the number of validation checks were used to terminate the network training. At epoch: 6 iterations, the gradient was equal to 3.758 (i.e. at gradient less than 1e-005, the training will stop). The number of validation checks was equal to 6; which is the appropriate value to stop training. The performance plot (Fig.11) shows the value of the function, in basis of training, validation, and test behaviors, versus the iteration number. The best validation performance, based on the mean square

error, was 0.0035334 at epoch 0. Since the validation and test curves are very similar, therefore no major problems or over-fitting occurred with the training (Fig.10).

The final weights and biases were:

Weight to layer 1 from input 1 (w {1; 1}) = [8.1898; -7.5562; -7.0081; 7.0008; 7.6404] Weight to layer 2 from layer 1 (w {2; 1}) = [1.9045 -1.9357 0.95875 -0.59307 -0.67968] Bias to layer 1(b {1}) = [-9.1555; 3.8046; -0.013523; 3.4992; 6.3596] Bias to layer 2 (b {2}) = [1.6103].

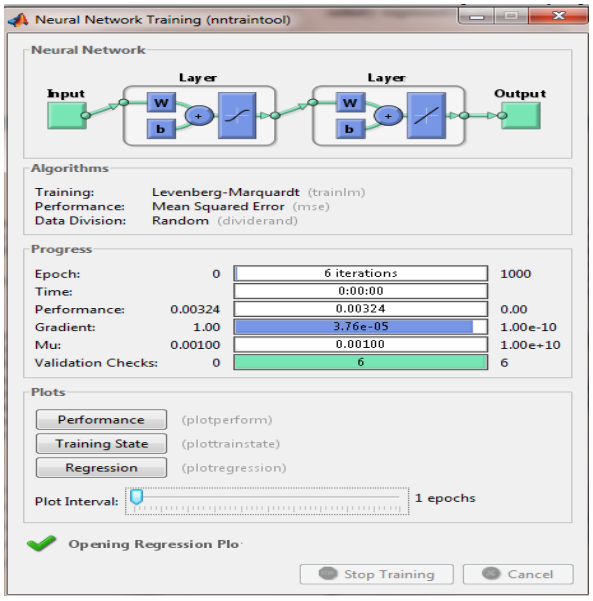


Fig 9. Neural Network Training

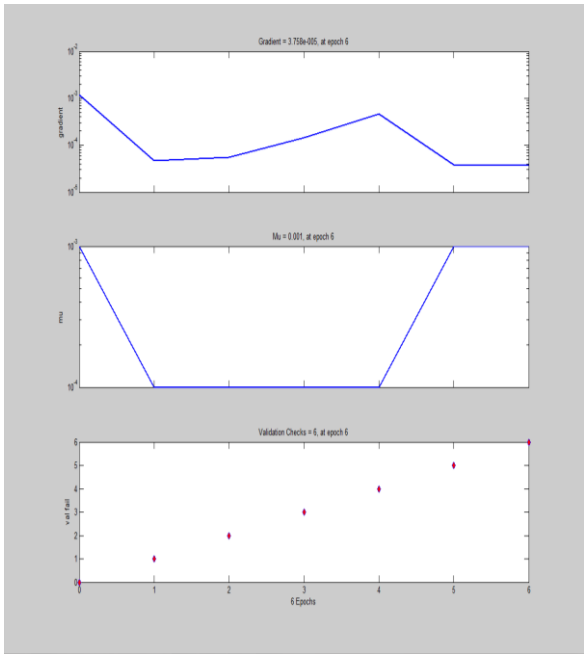


Fig 10. Training state of the generated ANN model

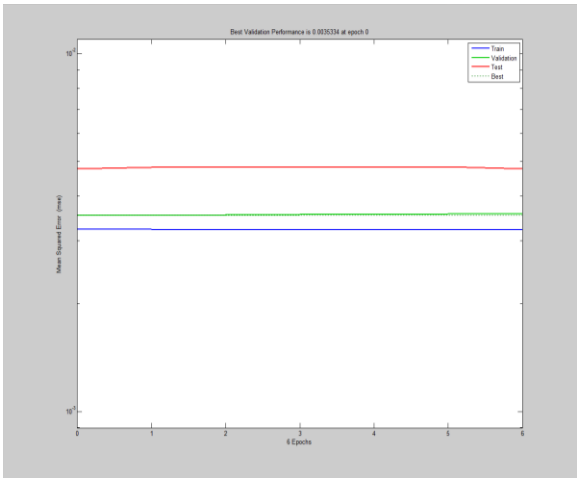


Fig .11. Performance of the generated ANN model

Linear regression plots the network outputs for training, validation and testing in order to determine the relationship between the outputs of the network and the targets (Fig.12). In each plot, the dashed line represents the perfect result, i.e. outputs =targets, whereas the solid line corresponds to the best fit linear regression. As the R-value approaches to one, then there is an exact linear relationship. The regression results (R-value) were 0.36427, 0.31759 and 0.26087 for training, validation and test, respectively. Those results were corresponding to a total response of 0.33523. The lower regression results can be attributed to fewer training data (accounting for 78 points) and/or the ANN configuration, in terms of number of hidden layers and neurons, might not being optimal. After the network was trained, validated and tested, the generated model can be used to predict the parameter Total Porosity through new input pH data.

It was found that experimental analysis of Total Porosity was close to the predicted data calculated from the configuration and this confirms the validity of this model.

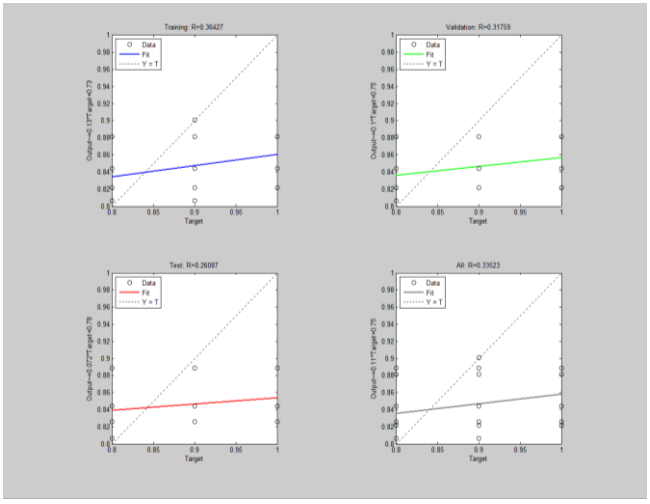


Fig.12. Regression plot of Training, Validation and Test for Total Porosity using ANN

5. Conclusion

Many mapping of contaminated soil sites and the resulting cleanup are time consuming and expensive tasks, requiring extensive amounts of geology, hydrology, chemistry, GIS in environmental contamination and data mining and pattern recognition techniques. The soil salinity based on alkalinity was predicted and ANN with a structure of 1–5–1 was proposed. The network showed an acceptable ability to capture the interrelationship between input: pH and output: Total Porosity concentrations. Values of correlation coefficient (R) of training, validation and test were 0.36427, 0.31759 and 0.26087, respectively. It is concluded that the developed ANN model is suitable for predicting the salinity concentrations. Development of novel methods for classification in data mining and pattern recognition like genetic based, and hybrid based is required to find out the optimal number of clusters for soil sites based on soil texture with Total Porosity.

References

[1] Kumar D, Kannathasan N, "A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining", IJCSI International Journal of Computer Science Issues,2011, 8 (3), pp. 422-428.

[2] Mahmoud Nasr, Hoda Farouk Zahran, “Using of pH as a tool to predict salinity of groundwater for irrigation purpose using artificial neural network”, June 2014.

[3] Demuth, H., Beale, M., Hagan, M., Neural Network Toolbox 5: Users Guide. The Math Works Inc., Natick, MA. 2007.

[4] Batjes NH, ISRIC-WISE global data set of derived soil properties on a 0.5 by 0.5 degree grid (version 3.0). Report 2005/08. ISRIC-World Soil Information: Washington.

[5] Study Guide: Soil Mechanics Level 1, Module 3, Unified Soil Classification System, National Employee Development Staff, Soil Conservation Services. United States Department of Agriculture. 1987.

[6] Ashok Kumar D, Kannathasan N, “Analysis of Linear and Segmented Linear Regression of Fruit Yield on Soil Salinity”, Proceedings of International Conference on Mathematical Modeling and Applied Soft Computing, vol.2, 2012, pp.275-282.

[7] Kumar DA, Kannathasan N, “A Study and Characterization of Chemical Properties of Soil Surface Data Using k-Means Algorithm”, Proceedings of the International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013, pp.264-270.

[8] Kumar DA, Annie MCLC, Begum TUS, “Computational Time Factor Analysis of k-Means Algorithm on Actual and Transformed Data Clustering”, 2012 International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012, pp. 49-54.

[9] Thangavel K, Ashok Kumar D,”Optimization of Code Book in Vector Quantization”, Annals of Operations Research, 2006, 143 (1), pp. 317-325.