Estimating Soil Salinity Over a Shallow Saline Water Table in Semiarid Tunisia

Fethi Bouksila[1,*], Magnus Persson[2], Ronny Berndtsson[2] and Akissa Bahri[3]

[1] *National Institute for the Research in Rural Engineering, Waters and Forests, 17 rue Hédi Karray, BP 10, 2080 Ariana, Tunisia*

[2] *Department of Water Resources Engineering, Lund University, Box 118, SE-22100 Lund, Sweden*

[3] *International Water Management Institute (Ghana), PMB CT 112, Cantonments Accra, Ghana*

Abstract: Rapid and reliable observations of soil electrical conductivity are essential in order to maintain sustainable irrigated agriculture. Direct measurement of the electrical conductivity of saturated soil paste (*ECe*), however, is tedious and time consuming. Therefore, there are needs to find efficient indirect methods to predict the soil salinity from other readily available observations. In this paper we explore the application of multiple linear regression (MLR) and artificial neural networks (ANN) to predict *ECe* variation from easily measured soil and groundwater properties under highly complex and heterogeneous field conditions in semiarid Tunisia. We compare two methods for dividing the data set into training and validation sub-sets; a statistical (SD) and a random data set division (RD), and their effect on model performance. The input variables were chosen from the plot coordinates, groundwater table properties (depth, electrical conductivity, piezometric level), and soil particle size at 5 depths. The results obtained with ANN and MLR indicate that the statistical properties of data in the training and validation sets need to be taken into account to ensure that optimal model performance is achieved. The SD can be considered as a solution to resolve the problem of over-fitting a model when using ANN. For the SD, the determination coefficient (R^2) when using an ANN model varied from 0.85 to 0.88 and the root mean square error from 1.23 to 1.80 dS m^{-1}. Because of the complexity of the field soil salinity process and the spatial variability of the data, this clearly indicates the potential to use ANN models to predict *ECe*.

Keywords: Neural networks, multiple linear regression, soil salinity, water table, dataset division.

INTRODUCTION

A shallow water table in combination with high soil salinity often leads to permanent soil resource degradation. In arid and semiarid climates, soil salinisation constitutes a major problem for irrigated land sustainability. Throughout the world, about 25% of irrigated areas are affected by salinity and water logging [1]. A shallow water table also constitutes an important soil degradation factor [2-7]. In Tunisia, 36 % of the irrigated areas are strongly sensitive to salinisation [8]. Soil salinisation over a shallow water table depends on climatic conditions, soil properties, vegetation, soil management (irrigation, fertilization, tillage, etc.), and depth to and salinity of the groundwater [9-13]. Evaporation from the soil surface creates a water potential gradient. In response to this gradient, water is transported from deeper levels towards the soil surface where it evaporates and dissolved matter in it increases its concentration in the top soil [14]. To reduce and avoid the risk of salinisation, it is important to control the soil salinity in order to keep it below the plant salinity tolerance.

Measurement of soil salinity in laboratory, especially electrical conductivity of the saturated soil paste (*ECe*), is tedious (sampling, soil preparation, and measurement). In the field, equipment such as time domain reflectometry (TDR) and other salinity sensors are used to give a quick estimate of the soil salinity. These methods give a good assessment of the soil salinity in a limited soil volume. Because of the spatial variability of soil properties, however, it is difficult to apply these methods to larger areas. Because of these constraints, there are needs to infer soil salinity from other more easily observed variables. In the lower valley of Euphrates, Dosso [15] found that the soil salinity in the surface was 10 times higher than in the groundwater. In Tunisia, Bach Hamba [16] found a poor correlation between surface soil salinity and salinity of the shallow water table. The absence of correlation between these parameters was attributed to the importance and complexity of the salinity process in the surface soil (effects of evaporation and precipitation). The key factor controlling the amount of evaporation is the depth to the water table below the soil surface [17]. Another parameter affecting soil salinity is the soil particle size distribution. Generally, capillary rise is larger in a medium-textured (loamy-sandy) soil than in a fine-textured (clay or loam clay) and sandy soil. Servant [18] observed that the surface soil salinity was more important in medium-textured soil as compared to that of fine-textured soil. The soil stratification also has influence on the capillarity rise. Massoumi [19] showed experimentally that the superposition of sand on a silty horizon reduces the capillarity rise as compared to superposition of silt on a sandy horizon. In a field study in Tu-

*Address correspondence to this author at the National Institute for the Research in Rural Engineering, Waters and Forests, 17 rue Hédi Karray, BP 10, 2080 Ariana, Tunisia; Tel: +216-71709033; Fax: +216-71717951; E-mail: bouksila.fethi@iresa.agrinet.tn

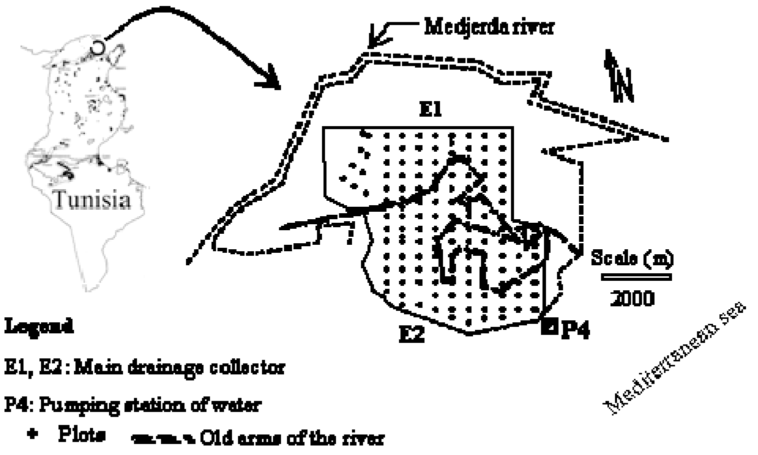

Fig. (1). Experimental area and sampling locations.

nisia, Bouksila [20] observed that the presence of a sandy horizon positioned between two fine-textural horizons constitutes a salt reservoir. To predict field scale spatial salinity (*ECe*) from electromagnetic induction data, Lesch *et al.* [21] showed that the multiple linear regression models (MLR) are theoretically equivalent to and cost-effective relative to cokriging. To increase the prediction accuracy, the MLR incorporate the trend surface coordinates [21].

Many mathematical models have been developed to predict soil salinity [7, 22-24]. Usually these models need a significant number of input parameters (climatic information, soil and water table properties, crop, irrigation water, drainage water, etc.). To apply these models, it is necessary to have an extensive observational data base to provide all input parameters. In many cases, however, it is difficult if not impossible, to supply all these input parameters. Due to this, parallel to the improvement of analytical and mathematical models, statistical techniques with ability to predict salinity levels with a few climatic and soil property input variables have also been developed. One of these techniques utilizes artificial neural networks (ANN). In soil science, ANN has been used to classify soil texture [25], to model nitrate leaching [26], estimating water content and soil solution electrical conductivity from TDR measurements [27,28], prediction of soil hydraulic properties [29]), and to predict soil salinity [30]. However, still limited success has been attained to predict spatial variation of soil salinity using linear and/or nonlinear statistical methods.

In this paper we explore the ability of ANN to predict the electrical conductivity of the saturated soil paste variation for highly complex and heterogeneous field conditions. In view of the above, the aim of this study is to predict soil salinity from easily measured soil and water table properties for a better water and soil management. We compare different data set divisions and effects on the model target. We also compare advantages of the ANN with multiple linear regression models. We close with a discussion on practical implications.

MATERIALS AND METHODS

Study Area

Field experiments were conducted in the irrigated area of Kalaat Landalous, situated in the northern part of Tunisia (35 km north of the capital Tunis), close to the Mediterranean Sea (Fig. **1**). The irrigated area covers 2900 ha and the main crops are fodder, cereal, and market vegetables. The climate is Mediterranean semiarid with average rainfall of 450 mm/year (rainy period from September to March). The potential evapotranspiration is 1400 mm/year. The soil is an alluvial formation of the Lower Medjerda river (xerofluent), characterized by a fine texture (silty clay to clay). The altitude varies from 2 to 6 m and the average soil surface slope varies from 0.05 to 2%. In 1987, a drainage and irrigation system using Medjerda water was constructed (Fig. **1**). The electrical conductivity of the river water is about 3 $\mathrm{dS\,m^{-1}}$ and the sodium adsorption ratio (SAR) is about 7. Both drip and sprinkler irrigation is used in the system. The drainage system is mainly composed of two primary open ditches (E1 and E2), subsurface PVC pipes, and a pumping station that discharges drainage water to the sea (P4). The subsurface drains have a diameter of 0.08 m, are 150 m long, and a separation distance of 40 m. They follow the average slope, so that the drain depth begins at 1.4 m and ends at 1.7 m before discharging into a secondary open drain. Before the completion of the drainage and irrigation system, the old Medjerda riverbeds (30 to 40 m wide and 1.5 m to 3 m deep) constituted a natural drainage system and the Medjerda water was discharged into these riverbeds during flood periods allowing farmers to irrigate their land.

A 1400 ha area surrounded by two primary open ditches (E1 and E2) was selected within the 2900 ha irrigated area (Fig. **1**) for experimental studies. The experiments were conducted in October 1989, at the end of the summer season.

Data Collection

In total 144 sampling plots, spaced at about 200 by 280 m were investigated (Fig. **1**). In each plot, soil samples were collected at 0.1, 0.5, 1.0, 1.5, and 2.0 m depth. The soil samples were analyzed to determine soil particle size and *ECe*. Soil particle size was measured in the laboratory using the sedimentation method (pipette and hydrometer). In gypsum-rich samples this standard method can not be used [31, 32]. For this reason, only 116 of the 144 plots present complete particle size data (for 0.1 m depth, 115 plots). Five fractions were measured, *clay* (d<2 µm), *fine silt* (2<d<20 µm), *coarse silt* (20<d<50 µm), *fine sand* (50<d<200 µm), and *coarse sand* (200 µm<d<2 mm). Table **1** shows a summary of the

Table 1. Summary Statistics of Soil (Particle Size (%) and Electrical Conductivity of the Saturated Soil Paste (*ECe*; dS m^{-1})) and Groundwater Table Properties (Depth (*Dgw;* m), Piezometric level (*PL;* m), and the Electrical Conductivity (*ECgw;* dS m^{-1}))

Parameter			Minimum	Maximum	Median	Mean	St.Dev.	CV(%)
Soil Depths (m)	0.1	*Clay*	5	57	33	34	10	30
		Silt	37	79	54	55	7	14
		Sand	1	38	9	10	8	75
		ECe	1.1	21.5	5.0	6.1	4.2	69
	0.5	*Clay*	7	62	37	37	13	35
		Silt	13	77	51	49	11	22
		Sand	0	89	7	12	16	129
		ECe	1.7	18.1	5.7	6.1	3.4	55
	1.0	*Clay*	6	62	30	31	13	42
		Silt	2	76	53	50	13	26
		Sand	0	97	11	18	19	106
		ECe	1.6	23.0	6.1	7.1	4.1	57
	1.5	*Clay*	6	67	27	28	11	40
		Silt	4	71	54	52	12	23
		Sand	1	87	15	19	17	88
		ECe	2.1	23.0	7.0	8.2	4.5	55
	2.0	*Clay*	5	60	29	30	12	41
		Silt	4	71	52	50	13	26
		Sand	1	91	14	20	20	101
		ECe	2.1	27.6	6.8	8.4	4.9	58
Ground Water		*Dgw*	1.14	2.90	2.15	2.20	0.31	14
		PL	0.35	4.05	1.92	1.90	0.79	41
		ECgw	3.90	59.6	18.30	15.60	10.10	55

St.dev. (standard deviation), CV (coefficient of variation= 100*St.Dev./Mean).

three particles sizes (clay, silt, and sand) for different depths. The *ECe* measured by the standard method according to USDA [33] was used to estimate the soil salinity.

Beside soil samples, the depth to the groundwater table from the soil surface (*Dgw*) and electrical conductivity of the groundwater (*ECgw*) were measured at each of the 144 plots. The coordinates (*x, y*) and the altitude (*z*) of the plots were measured by GPS (Trimble, model 4600LS, Trimble Ltd. Sunnyvale, CA, USA; accuracy equal to 0.01 m for *x* and *y* and 0.02 m for *z*). The altitude was used to calculate the piezometric level ($PL = z - Dgw$) of the groundwater table.

Modeling Soil Salinity

A suitable regression models is specified that relates the target soil properties (like *ECe*) to a transformed linear combination of the parameters whose influence the *ECe* (such soil and water table properties) and trend surface coordinates. In the statistical literature, this kind of model is commonly called a spatial linear regression model [34]. Two statistical methods were used to predict the soil salinity, the first is a linear model, multiple linear regression (MLR) and the second is non linear model, artificial neural networks (ANN).

Multiple Linear Regression (MLR)

To arrive at a best model depending on an optimal data set division for the MLR, the following steps were adopted:

First step: Choosing input variable. For each plot, there are more than 20 input variables to chose from to predict soil salinity; i.e. 15 particles sizes (*clay*, *silt*, and *sand* for each of the five depths), average of particle sizes for depths above or below the actual depth (e.g., for 1.0 m the average particle sizes of 0.1 and 0.5 and 1.5 and 2.0 m), 3 variables for the groundwater (*Dgw, PL,* and *ECgw*), and coordinates (*x, y*). The two surface coordinate (*x, y*) were included at once as predictor to consider the spatial variation in *ECe* across soil types, landscape types, position of the drainage system, and farming management. The first step was to select the best input variable for the MLR. The software Statgraphics 5 plus (Manugistics Inc., USA) was used to find the best model to estimate soil salinity (*ECe*) for each depth. The software uses combinations of all input variables and calculates the coefficient of determination (R^2) and the root mean square error

(RMSE). The best models will have a minimum RMSE and a maximum R^2. In this step the entire data set was used in the analysis.

Second step: Data set division. Firstly, all available data were randomly divided into two parts (training and validation). In total, 80% of available data were used for training and the remaining 20% were used for validation. Secondly, a trial process was used to divide the data so that the statistical properties of the data in each subset were as close to each other as possible, and thus represented the same population. If the validation data fall outside the range of the data used for training, the results obtained using the validation data can be worse than those obtained using the training data [35]. The statistical data treatment used included the minimum, maximum, range and *t*- and *F*-tests (at a significance level of 0.05), see Shahin *et al.* [35] for details. The data set division that verified all statistical criteria was used to calculate the parameters of the MLR.

Third step: A comparison between the results obtained with statistical data (SD) and random data (RD) set division was used to evaluate the performance of the two data handling types for the MLR model.

Artificial Neural Network (ANN)

Artificial neural networks (ANN) are non-linear models that make use of a parallel programming structure capable of representing arbitrarily complex non-linear processes that relate the inputs and outputs of any system [36]. It provides better solutions than traditional statistical methods when applied to poorly defined and poorly understood complex systems involving pattern recognition [37]. The ANN is structured, similarly to the biological neural network, by interconnected layers composed of neurons. An artificial neuron is the architectural unit of the ANN. It basically consists of a transfer function and two scalar numbers, a weight and a bias. The input is a scalar that is multiplied by the weight and added to the bias. The transfer function is applied to this result. To develop and train a ANN involve (a) choosing a training set that contains input–output pairs, (b) defining a suitable network (number of layers and number of neurons in each layer), (c) training the network to relate the inputs to the corresponding outputs by estimating the ANN weights, and (d) testing the identified ANN. If compared to a conceptual model, (b) is equivalent to the development of the model and (c) is the estimation of the parameters of the designed model. The process of training the ANN consists of a self organizing learning process through a procedure that minimizes the error between the ANN output and the target values. The objective of the training is to find the weights of each neuron that will result in the minimum error. In the present study, a two-layer (one hidden and one output layer) feed-forward ANN trained by a back-propagation algorithm using the Levenberg–Marquardt optimization were used [38]. Back-propagation can be explained as the adjustment of ANN weights and biases by back-propagating the differences between the ANN output and actual target. Prior to ANN application, the original input and target are standardized to ensure that every input receives equal attention during the training [39]. As for the MLR above, the data were split in to two parts, 80% for training and 20% for validation. Each node receives the weighted outputs from the node in the previous layer, which are summed to produce the node input. The node input is then passed through a non-linear sigmoid function to generate the node output, which is passed to the weighted input paths of many other nodes.

Learning and training are fundamental in types of neural networks. Training is the procedure by which the network learns; learning is the end result of that procedure. Learning consists of making systematic changes to the weights to improve the network's response performance to acceptable levels. The network learns by adjusting the weights connecting the layers. The network starts by finding linear relationships between the inputs and the output. Weight values are assigned to the links between the input and output neurons. Once those relationships are found, neurons are added to the hidden layer so that nonlinear relationships can be found. The aim of training is to find a set of weights that will minimize the error. During training, the output predicted by the network is compared with the target and the root mean squared error (RMSE) between the two is calculated. More detailed explanation is available in Changhui and Xuezli [40]. To the output layer, a pure linear transfer function was allocated. As mentioned before, an ANN with one hidden layer and one output layer with a single neuron were used.

Before running the ANN model the following steps were made:

Choice of input. The results of the MLR were not satisfactory for the 0.1 and 0.5 m depths. Therefore, we tried to find other combinations of input variables for these depths. Here, the input for the ANN model was chosen based on (i) the correlation coefficient between the target and the input variable, (ii) the best input for the MLR, and (iii) on an ANN sensitivity analysis for various number of inputs (see Persson and Uvo [28], for details). For other depths (1.0, 1.5, and 2.0 m), the best input found for the MLR was used in the ANN models. To compare the ANN and MLR, the maximum number of input variables in the ANN model will be less or equal to those in the MLR model. For the sensitivity test, we fixed the number of hidden neurons to 7. The best combination of input variables will display the smallest RMSE and highest R^2.

Optimal number of neurons in the hidden layer. We used the principle of constructive algorithms, which essentially start testing a minimum number of hidden neurons and then add neurons until performance ceases to increase [41]. This procedure was used for all soil depths. The optimal number of hidden neuron was then used for the final ANN model.

Data set division. The same methodology as for the MLR was used to choose the data set division (RD and SD). For the SD model, after training each ANN 20 times, the average output was calculated and compared to the target *ECe*. The R^2 and RMSE were then calculated for the training and validation subsets. For the RD model we used 10 different randomly divided data sets when training the ANN. For each data set division the ANN was trained 20 times as described above. The average of 10 times 20 outputs were then calculated and compared to the output of the SD model.

RESULTS AND DISCUSSION

Soil and Groundwater Properties

Table **1** shows a summary of soil and groundwater properties. The average fraction of *clay* varied from 28 to 34 %

Table 2. RMSE and R^2 of the Best Model to Estimate the Electrical Conductivity of the Saturated Soil Paste (*ECe*) (n=116 Except for 0.1 m Depth, n=115)

Depth (m)	Input	R^2	RMSE (d S^{-1})
0.1	***y***, *sand1*.5, *sand*2.0, *Dgw*, *ECgw*	0.253	3.74
0.5	*Dgw*, *ECgw*, *sand*0.1, *sand*1.5	0.524	2.36
1.0	*Dgw*, *ECgw*, *silt*1, *silt* (mean1.5 and 2.0)	0.655	2.46
1.5	***y***, *ECgw*	0.713	2.44
2.0	***x***, *ECgw*	0.628	3.02

S (1.5): sand content at 1.5 m deeps (%), *Silt* (1.5,2.0): mean of the silt content at 1.5 and 2.0 m soil depth (%), *x* and *y* (coordinate of the plots; m); *Dgw* (water table depth, m); *ECgw* (water table electrical conductivity, dSm^{-1})

and *silt* from 49 to 55 %. Contrary to *clay* and *silt*, the average of *sand* fraction increased with depth. The maximum value of *sand* fraction explains some of the textural stratification in the soil profile. The large variation coefficient, especially for *sand*, reflects the alluvial origin of soil and the impact of the change of the Medjerda river bed properties on the particle size distribution. The *soil salinity* varied from 1.1 to 27.6 dS m^{-1}. The average *ECe* for all depths was higher than 6 dS m^{-1}, thus the soil is considered to be saline [33]. The maximum *ECe* at 0.1 m depth was 21.5 dS m^{-1} and at 0.5 m depth 18 dS m^{-1}, lower than at other depths. This is probably a result of natural soil leaching [4]. The variation coefficient for *ECe* is close to 60 % and this variability may be considered as large [42].

At the end of the summer, the average *depth to the groundwater table* was 2.2 m (below the PVC drains) and varied from 1.1 to 2.9 m. The variation coefficient was 14 % for all depths. The variation coefficient for water table salinity was considerably higher, 55 % (Table **1**). The *groundwater salinity* varied from 4.1 to 59.6 dS m^{-1}. The similarity between the chemical composition of the highest *ECgw* (59.6 dS m^{-1}) and Mediterranean Sea water indicates that this plot (located at the extreme east of the irrigated area) is situated in a maritime intrusion zone [16]. A previous geostatistical analysis of soil properties (*particle size, saturated hydraulic conductivity, bulk density, and ECe*), and *groundwater salinity* and *depth* showed that the variograms were slightly structured and characterized by a high nugget effect, mainly due to the variability within the sampling distance (grid 200 m x 280 m) [4,16]. Previous analyses of soil hydraulic parameters at different spatial scales did not display a significant reduction of variability below this spatial scale [43]. The farmer practice should take into account the *ECe* variation in order to reduce the risk of soil degradation and to increase the crop production. Indeed, the soil salinity limits water uptake by plants and leads to a decrease in crop production. Therefore, the land use and crop rotation should take into account the crop tolerance to soil salinity. Also, *ECe* was used to estimate the leaching requirement (LR) [20]. An over-estimation of the LR would result in the use of excessive amounts of irrigation water and increased salt loads in drainage systems, which can detrimentally impact the environment and reduce water supplies [1]. The underestimation of LR could increase the *ECe* and the sodium exchangeable percentage (*ESP*) which could result in soil structure degradation. In Kallat Landalous, a negative correlation was observed between the *ESP* and soil saturated hydraulic conductivity [4]. For these reasons, an accurate estimation of *ECe* contributes to sustainable land planning aimed at mitigating soil degradation and increasing crop production.

Prediction of Soil Salinity with MLR

Best MLR Model

According to the Pearson's correlation analysis, the field *ECe* were poor correlated with the soil particle size and the plots coordinate ($-0.39 \leq R \leq 0.26$). The *ECe* were negatively correlated to the *depth to the groundwater table* and to the *piezometric level* ($-0.41 \leq R \leq -0.09$). The best input variable to explain the *ECe* variation was the water table salinity ($0.15 \leq R \leq 0.84$).

For each soil depth, about 22 000 MLR models with different input combinations were tested to obtain the best model based on RMSE and R^2. Table **2** shows these results for each soil depth. For some depths only one spatial coordinate was included in the best MLR model which may seem surprising. Usually two coordinates are necessary to represent the linear trend surface. In a large field study from 10 sets of trend surface variables Lesch *et al.* [34] also found one plot coordinate in some of their best MLR models to predict the *ECe*,. As seen from Table **2** the R^2 increases from soil surface down to the drain depth (1.5 m). Above the PVC drain (0.1 to 1.0 m), the sand and silt variables, characterized by a high variation coefficient (Table **1**) were selected in the best MLR model. For soil depths below the PVC drain (1.5 and 2.0 m depths), the *Dgw* does not appear as input in the best MLR model. The *ECgw* is found as predictor in every model in Table **2**. For the 0.1 m depth, only 25% of the *ECe* variation can be explained by the best MLR model. This poor result reflects the complexity of salt distribution, especially in the surface soil. Probably soil management, irrigation parameters, and climatic conditions not included as input variables have a large impact on the result for the top soil layers. Also, several plots show textural stratification [4]. This stratification causes a discontinuity of the moisture content at the interface of two successive layers which affect the water and salt flow in the soil profile. Unfortunately, the pedologic sampling method can not be used on a large scale. With a fixed soil sampling depth this information is lost. These factors explain the poor MLR results for the soil salinity prediction for the upper soil layers (0.1 m). For other soil layers (0.5 to 2 m), the correlation coefficient from the simple linear regression between the *ECgw* and the *ECe* varied from 0.64 to 0.84. These results reflect the importance of salt

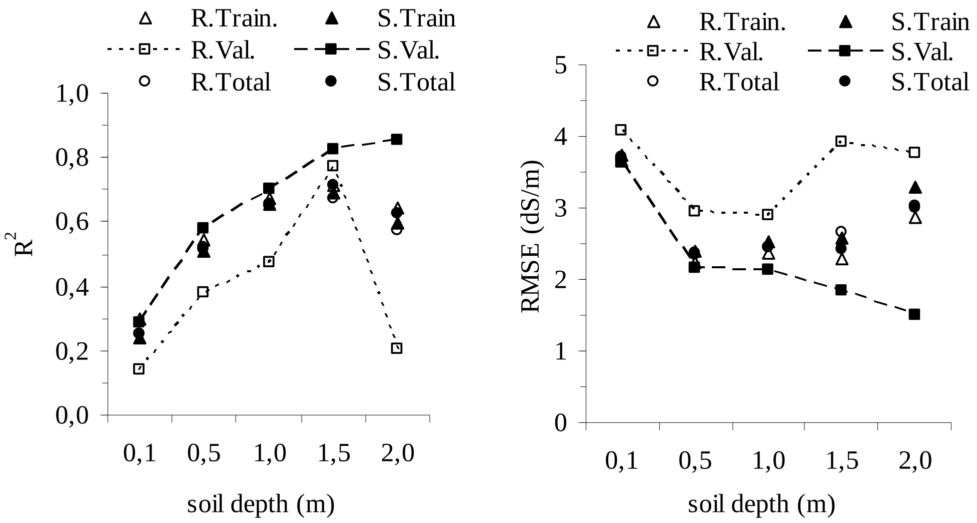

Fig. (2). R^2 and RMSE of the MLR for each depth, various subset (validation (Val.), training (Train.), total) and the dataset division method (Random (R.), statistical (S.)).

build-up in the soil profile from the shallow water table in arid climates.

Effects of Data Set Division

It is difficult to divide the data using SD when many input variables are used. The null hypothesis of no difference between the variance of the validation and training subset was rejected by an *F*-test for the water table input at 0.1 m depth. Fig. **2** shows R^2 and RMSE for the MLR using each subset (validation and training), data set division method (RD and SD), and soils depth. The global R^2 and RMSE (for all plots) is almost identical for both RD and SD divisions. For the validation subset result, however, the difference is large. For all depths, R^2 and RMSE of the SD validation subset are much better than that of RD. The R^2 varied from 0.14 to 0.77 and the RMSE from 2.88 to 4.09 dS m^{-1} for RD. For SD, the R^2 varied from 0.28 to 0.85 and the RMSE from 1.51 to 3.63 dS m^{-1}. With our field data, characterized by considerable variability (Tables **1**), it is evident that SD improves the result of the validation subset (R^2 and RMSE).

ANN Prediction of Soil Salinity

Choice of Input Variable

The MLR gave poor results in the root zone (0.1 and 0.5 m soil depth). Since these depths are the most important for crops, we tried to improve the performance of the ANN model by choosing input variables using the results from the MLR model together with a sensitivity analysis. Table **3** shows R^2 and RMSE for the ANN models using the different input variables. The best input for the ANN model contained five variables (*x*, *y*, *Dgw*, *PL*, and *ECgw*) for 0.1 m and three variables for 0.5 m soil depth (*x*, *Dgw*, and *ECgw*). The performance of ANN models was improved when the plot coordinates (*x* and/or *y*) were added as input variables (Table 3). At 0.1 m depth, the R^2 varied from 0.40 to 0.77 and the RMSE from 3.38 and 2.05 dS m^{-1} without and with the plot coordinates input (*x*, *y*) respectively. Also, at 0.5 m soil depth, the R^2 varied from 0.74 to 0.87 and the RMSE from 1.74 to 1.25 dS m^{-1}, respectively, without and with the input *x*. These best input variables for 0.1 and 0.5 m soil depths were further used below for the ANN modeling. For other soil depths (1.0, 1.5, and 2.0 m), the best combination of inputs found through the MLR analysis were also used for the ANN modeling.

Effects of Data Set Division

To find the optimum number of hidden neurons in the ANN model, the principle of constructive algorithms was applied. The optimal number was found to be 7 for depths 0.1, 0.5, 1.0 m, 10 for 1.5 m depth, and 11 for the 2.0 m depth (Fig. **3**). In Table **4**, the RMSE and R^2 (average output of 20 different networks) using RD and SD are presented. Using SD, the overall R^2 varied from 0.85 to 0.88 and the RMSE from 1.23 to 1.80 dS m^{-1}. For the validation subset, the R^2 varied from 0.58 to 0.87 and the RMSE from 1.21 to 3.17 dS m^{-1}. The worst result was observed for the upper soil layer. At 0.1 m soil depth, R^2 was 0.85 and the RMSE was 1.8 dS m^{-1}. For the validation subset, R^2 was 0.58 and the RMSE was 3.17 dS m^{-1}. For all depths the performance of the ANN model is better using SD as compared to RD. When RD is used, there is a large difference between the different subset results (the poorest result was observed in the validation subset). Therefore, it can stated that the impact of data set division on the ANN performance is very important, especially for the surface soil layers (0.1 and 0.5 m) and below the PVC drains (2.0 m depth), where the impact of the drainage network is negligible.

At 0.1 m soil depth the result of the ANN model using RD was characterized by over-fitting. This was in spite of that the method applied is used to prevent nonlinear instability and over-fitting, that is, random data order [44] and averaging the output [28]. From the 10 randomly divided inputs, 70 % of the models had an R^2 less than 0.2 for the validation subset and higher than 0.85 for the training subset. Consequently, the model output fitted the data well for the training data, yet produced poor forecasts using validation data. An ANN model is usually capable of learning the signal from the data, but as training progresses, it often starts learning the

Table 3. Results of Sensitivity Test to Predict Soil Salinity with ANN at 0.1 and 0.5 m Soil Depths

Input variables for 0.1 m depth	**R^2**	**RMSE (d S^{-1})**
ECgw, Dgw	0.312	3.55
Dgw, ECgw, PL	0.402	3.38
ECgw, S1.5, S2	0.508	2.99
Dgw, Ecwt, S1.5, S2	0.585	2.83
Dgw, Ecwt, S1.5, S2, y (best input for MLR*)*	**0.708**	**2.30**
ECgw, S1.5, S2, ***x***	0.733	2.26
ECgw, Dgw, ***x, y***	**0.733**	**2.22**
ECgw, Dgw, ***x, y, z***	**0.746**	**2.16**
Dgw, ECgw, PL, x, y	**0.773**	**2.05**

Input variables for 0.5 m depth	**R^2**	**RMSE (d S^{-1})**
ECgw	0.449	2.51
ECgw, Dgw	0.737	1.74
ECgw, Dgw, ***y***	0.767	1.65
ECgw, Dgw, PL	0.781	1.59
ECgw, Dgw, z	0.803	1.52
ECgw, Dgw, S0.1, S1.5 (best input for MLR*)*	**0.827**	**1.42**
ECgw, Dgw, S0.1, S1.5, ***x, y***	0.856	1.31
ECgw, Dgw, ***x****, z*	0.873	1.26
ECgw, Dgw, x	**0.874**	**1.25**
ECgw, Dgw, ***x, y***	0.875	1.24
*ECgw, Dgw****, PL, x, y***	0.890	1.09
ECgw, Dgw, PL, S0.1, ***x, y***	0.898	1.06

S (1.5). Percentage of the soil sand particle size at 1.5 m soil depth, *coordinate (x,y)*, *altitude (z)*, *water table* (depth (*Dgw*, salinity (*ECgw*), piezometric level *(PL)*)

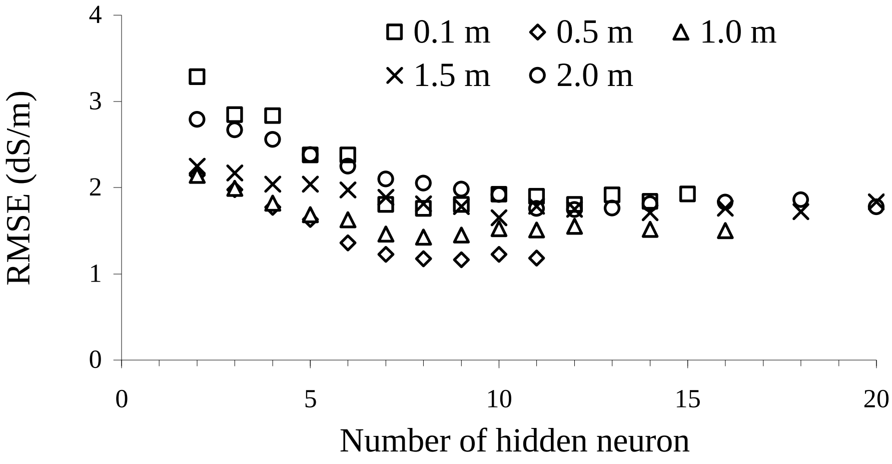

Fig. (3). The average root mean square error (RMSE) of 20 neural network runs plotted against the number of neurons in the hidden layer for the 5 soil depths.

noise in the data (i.e., over-fitting). That is, the forecast error of the model over the validation period first decreases and then increases as the model starts to learn the noise in the training data [45]. To resolve this problem, a technique called early stopping is normally used. In this technique 3 data subsets instead of 2 (training, testing, and validation [28]) are used when training the ANN. Another common technique is to reduce the number of hidden neurons [46]. However, using these approaches did not resolve the problem of over-fitting for the upper soil layer.

Table 4. Influence of the Data Set Division Method on the ANN Model to Predict Soil Salinity (*ECe*)

Depth	Division	Training		Validation		Total	
(m)	method	RMSE	R2	RMSE	R2	RMSE	R2
0.1	Random	1.69	0.881	4.36	0.156	2.41	0.688
	Statistic	**1.32**	**0.933**	**3.17**	**0.580**	**1.80**	**0.851**
0.5	Random	1.20	0.876	2.99	0.442	1.69	0.756
	Statistic	**1.25**	**0.879**	**1.21**	**0.867**	**1.23**	**0.875**
1.0	Random	1.33	0.898	3.18	0.450	1.81	0.810
	Statistic	**1.32**	**0.910**	**2.01**	**0.770**	**1.46**	**0.876**
1.5	Random	1.62	0.864	4.13	0.602	2.31	0.766
	Statistic	**1.60**	**0.881**	**1.91**	**0.849**	**1.65**	**0.867**
2.0	Random	1.71	0.890	4.55	0.288	2.48	0.756
	Statistic	**1.78**	**0.886**	**1.73**	**0.830**	**1.76**	**0.874**

Comparison Between MLR and ANN Models

For all soil depths, the performance of both MLR and ANN models was better with SD as compared to RD (Fig. **2**, Table **4**). Also, with SD the performance of ANN was better than the MLR, especially when the ANN best input was used (0.1 and 0.5 m depths). With SD, the R^2 was 0.58 and 0.28 for the ANN and the MLR model, respectively, using the validation subset at 0.1 m soil depth. For 0.5 m depth, the accuracy to predict *ECe* was better with the ANN (for the validation subset, the RMSE was 1.21 dS m^{-1} for ANN and 2.14 for the MLR model) in spite of using less input variables than in the MLR. For all 5 soils depths, the accuracy of the predicted *ECe* was also better with ANN as compared to the MLR model (Fig. **4**). With SD division, the result of the validation subset obtained by the MLR was usually better than that obtained with ANN model and RD. Also other field studies have shown that statistical methods (principal component analysis, cluster analysis, self organising map etc) can be used to determine the best input variable and to divide the data into relevant subsets for ANN models [47-49].

Based on the above it may be stated that the ANN model can extract more information (related to the *ECe* variation) from the plot coordinates than the MLR. For the best MLR model (Table **2**), the R^2 varied from 0.24 to 0.25 and the RMSE from 3.75 to 3.74 dS m^{-1}, respectively, without and with the input variable *y* for salinity prediction at 0.1 m depth. For the ANN model, the R^2 varied from 0.58 to 0.71 and the RMSE from 2.83 to 2.30 dS m^{-1}, respectively, without and with the input y (Table **4**). This shows that the spatial dependency cannot be represented by a linear model. The nonlinear spatial dependency could, however, be described by the ANN model. In the study area, there are 540 farmers and the farmer's land area varied from 0.15 to 400 ha [50]. The farmer's agricultural management (irrigation, fertilization, crop, agricultural soil practices, etc) is, however, much diversified [50]. This has a considerable effect on the soil salinity distribution, especially at the soil surface. Usually, farms close to each other have similar agricultural practices. Before the completion of the drainage and irrigation systems in 1988, the old Medjerda riverbed constituted a natural drainage system and the Medjerda water was discharged into this riverbed allowing farmers to irrigate their land. These farming plots generally have lower soil salinity [4]). Farmers apply organic and chemical fertilizers, plow, irrigate and cultivate their land during all seasons. Contrary to this, farmers with land in the lower part of the irrigated area, use the land for rainy annual crops and grazing due to salinity and water logging. These management practices significantly affect water and salt transport in the soil [20, 51]. For the large study area, however, it is very difficult to quantify all affecting variables for the soil salinity. In any case, they all add up and contribute to the spatial variability of soil salinity with a specific spatial correlation.

SUMMARY AND CONCLUSIONS

An accuracy estimation of soil salinity is appreciated for both land planners and farmers to make appropriate decisions about crop production and soil and water management In this paper we explored the ability of ANN to predict the spatial electrical conductivity of the saturated soil paste (*ECe*) variation at 5 soil depths (0.1, 0.5, 1.0, 1.5, and 2.0 m) under highly complex and heterogeneous field conditions in semiarid Tunisia.

The input was chosen from more than 20 input variables; plot coordinates (*x*, *y*), *altitude* (*z*), soil particle size at 5 soil depths, and groundwater table properties (*depth, electrical conductivity (ECgw*), and *piezometric level*).

From about 22000 models with different input combinations tested, the plot *coordinate* (*x* and/or *y*) was selected among the best input for the MLR. For the ANN model, at 0.1 m depth, the R^2 varied from 0.40 to 0.77 and the RMSE from 3.38 and 2.05 dS m^{-1} without and with the plot coordinates (*x*, *y*) input, respectively. Consequently, for large fields, the *plot coordinate* indirectly gives input to the statistical models regarding the spatial correlation of parameters that has large effect on the *ECe* (such as farmer's agricultural practices, drainage system efficiency, etc). The final number of input variables used in MLR and ANN are related to the complexity of the soil salinity process, it decreased with the soil depth from 5 to 2.

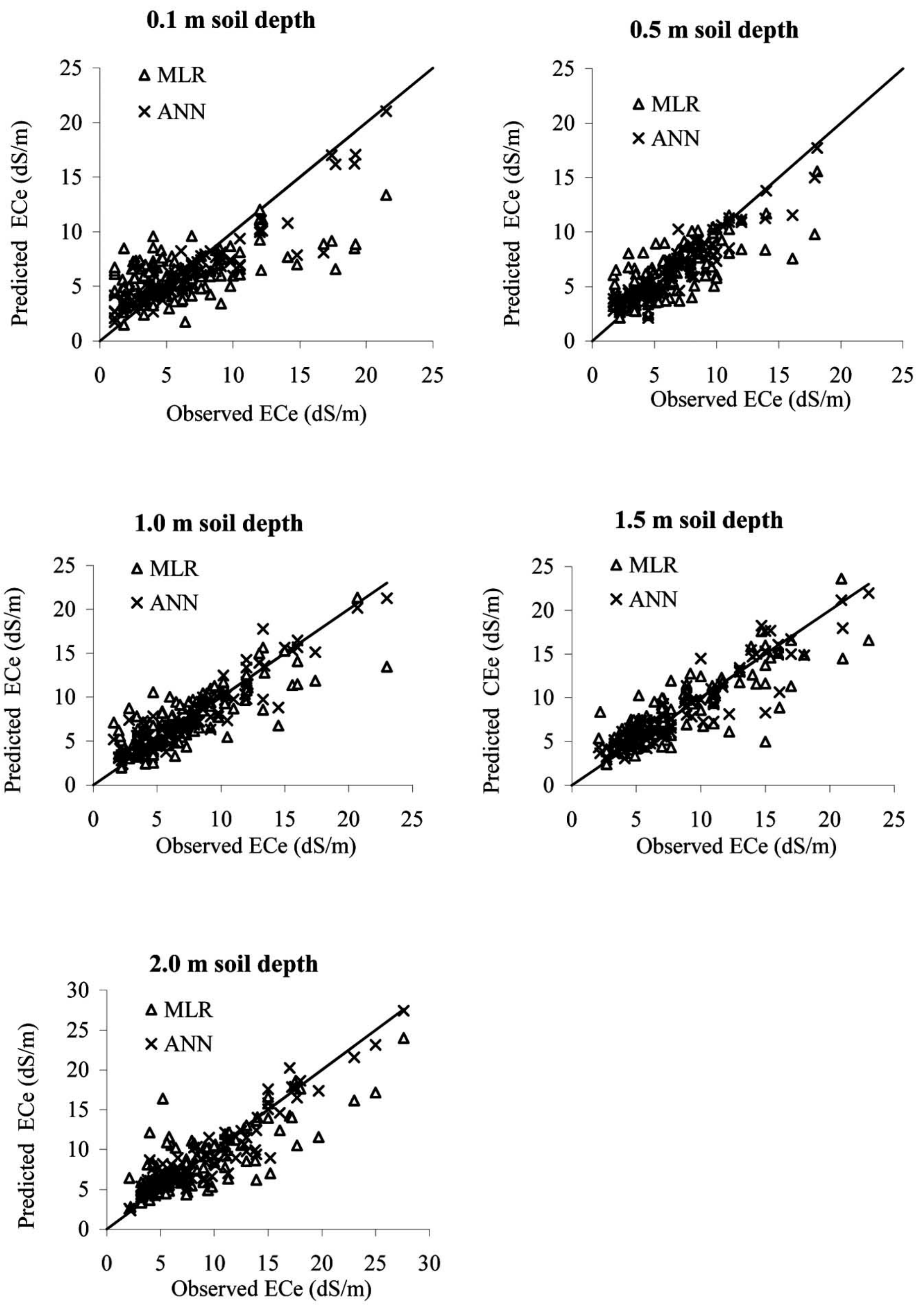

Fig. (4). Comparison between observed and predicted soil salinity (*ECe)* with the MLR and ANN models, using the statistical data set division.

The performance of MLR and ANN models are better with SD as compared to RD division, especially for the validation subset. The statistical properties of the various data subsets (training and validation) need to be considered to ensure that each subset represents the same population. Also, for 0.1 m soil depth, in spite of applied methods to prevent nonlinear instability and over-fitting, the result of the ANN model using RD was characterized by over-fitting. However, with SD, the performance of the validation subset was improved. Consequently, SD can be considered as a technique against the problem of over-fitting. However, when the number of inputs becomes large, it may be difficult to divide the data in a way so as to take statistical properties of the various input variables into account. In general, however, for all 5 soil depths and for the various subsets, the performance of ANN to predict the *ECe* was better than the MLR model. For the ANN model, R^2 varied from 0.85 to 0.88 and the RMSE from 1.23 to 1.80 dS m^{-1}. For the MLR, the R^2 varied from 0.25 to 0.71 and the RMSE from 2.33 to 3.68 dS m^{-1}. Because of the complexity of the field soil salinity process and the resulting spatial variability of the data (1< *ECe* < 28 dS m^{-1} and 1< *ECgw*< 60 dS m^{-1}), results clearly indicated the potential use of ANN models to predict the *ECe*.

ACKNOWLEDGEMENTS

This work was supported by the Swedish International Development Cooperation Agency (SIDA) and the Swedish Research Council. We would like to thank Imed Bach Hamba for providing soil salinity data and Cintia Uvo for her help with the neural network analysis.

REFERENCES

[1] Rhoades JD, Kandiah A, Mashali AM. The use of saline waters for crop production. FAO, irrigation and drainage 1992; p. 48.

[2] Dieleman PJ. La salinité. Seminar of Bagdad. FAO, Irrigation and Drainage 1970; vol. 7: pp. 107-27.

[3] Belhacene H, Chayat M. Evaluation des problèmes d'engorgement des sols, de drainage et de la qualité des eaux dans le périmètre de Tadla, Volume I et II. M-Sc. diss, IAV Hassan II, option Rural Engineering, Rabat (Morocco) 1992.

[4] Bouksila F. Bonification des sols : Cas du périmètre de Kalâat Landalous. Caractérisation physique des sols et étude de la variabilité spatiale de leurs propriétés en vue de la détermination des facteurs et des zones à risques de salinisation. M-Sc. diss, Tunisian National Agronomic Institute (INAT), Tunisia 1992.

[5] Bahri A. Utilisation des eaux et des sols salés de la plaine de Kairouan (Tunisie).Ph.D. diss, Toulouse (France) 1982.

[6] Mustapha ATA, Seliem MH, Bakahati HK. Effect of subsurface drainage on salt movement and distribution salt-affected soil. Isotope and radiation technique in soil physics and irrigation studies. IAEA and FAO 1983; pp. 265-81.

[7] Jorenush MH, Sepaskhah AR. Modelling capillary rise and soil salinity for shallow saline water table under irrigated and non-irrigated conditions. Agric Water Manage 2003; 61: 125-41.

[8] DG/ACTA. Examen et évaluation de la situation actuelle de la salinisation des sols et préparation d'un plan d'action de lutte contre ce fléau dans les périmètres irrigués en Tunisie. Phase 2 : Ebauche du plan d'action. Technical Report, Tunisian Farmland Conservation and Management Department (DG/ACTA), Tunisia, Sept 2006; Ref A.42.

[9] Gardner WR. Some steady state solutions of the unsaturated moisture flow equation with application to evaporation from a water table. Soil Sci 1958; 85: 228-32.

[10] CRUESI. Research and training on irrigation with saline water (1962-1969). Technical Report. CRUESI- Tunis/UNESCO-Paris 1970.

[11] M'Hiri A. Effet de l'irrigation sur la stabilité structurale des sols de texture fine. Proceeding of the First National Congress of Earth Sciences, Tunis (Tunisia) 1981: pp. 295-301.

[12] Rieu M. Elément d'un modèle mathématique de prédiction de la salure dans les sols irrigués : Application au Polders du Tchad. Ph.D. diss., Univ. Toulouse (France) 1983.

[13] Mermoud A, Musy A. Salinisation du sol depuis une nappe peu profonde : Simulation de l'effet d'un abaissement de la nappe sur les remontées d'eau vers la surface. In Special technical session of the 42nd Executive council meeting of the International commission of Irrigation and Drainage (ICID), Beijing (China) 1991: pp. 1-9.

[14] Rudraju TR. Dynamics of salt transport and ion chemistry in the unsaturated zone in the presence of a shallow aquifer. Doct Thesis. Department of Civil Engineering (Suiko). Faculty of Engineering, Kyushu Univ. 1995.

[15] Dosso M. Géochimie des sols salés et des eaux d'irrigation. Aménagement de la basse vallée de l'Euphrate en Syrie. Ph.D. diss.,Toulouse, France 1980.

[16] Bach Hamba I. Bonification des sols : Cas du périmètre de Kalâat Landalous. Caractérisation de la salinité initiale du sol en vue de la détermination des facteurs et des zones à risques de salinisation. M-Sc. diss, INAT, Tunisia 1992.

[17] Salama RB, Otto CJ, Fitzpatrick RW. Contributions of groundwater conditions to soil and water salinisation. Hydrogeol J 1999; 7:46-64

[18] Servant J. Contribution a l'étude pédologique des terrains halomorphes. L'exemple des sols salés du sud et du sud ouest de la France. Ph.D. diss., ENSA, France 1975.

[19] Massoumi AM . Etude expérimentale sur le mécanisme du mouvement capillaire de l'eau et des sels solubles dans le sol : Influence des rapports cationiques des solutions et de la texture sur le processus d'alcalinisation. Ph.D. diss., Univ. of Paris, France 1968.

[20] Bouksila F, Hachicha M, M'Hiri A. Variabilité des propriétés des sols du périmètre irrigué de Kalâat Landalous et risques de salinisation. Proceeding of the First scientific days of the Research Center in Rural Engineering, Nabeul (Tunisia) 1995: pp. 78-89.

[21] Lesch SM, Strauss D, Rhoades J. Spatial Prediction of Soil Salinity Using Electromagnetic Induction Techniques 1. Statistical Prediction Models: A Comparison of Multiple Linear Regression and Cokriging. Water Resour Res 1995; 31(2): 373-86.

[22] Raes D, Denyse E, Deproost P. UPFLOW, a model to assess water and salt movement from a shallow water table to the topsoil. Proceeding of the PCSI Workshop, 28-29 mai 2002, Montpellier (France) 2002. Available from : http://hal.cirad.fr/docs/00/18/03/42/PDF/Raes.pdf

[23] Srinivasulu A, Sujani Rao CH, Lakshimi GV, Satyanarayana TV, Boonstra J. Model studies on salt and water balances at Konanki pilot area, Andhra Pradesh, India. Irrigation Drainage Syst 2004; 18: 1-17.

[24] Wahba MAS, El Ganinym M, Abdel Daym M S, Kandil H, Gobran A. Evaluation of drainmod-S for simulating water table management under semi-arid conditions. Irrigation Drainage 2002; 51: 213-26.

[25] Chang DH, Kothari R, Senior M, IEEE, Islam S. Remotely sensed brightness temperature over the Southern Great Plains. IEEE Trans Geosci Remote Sensing 2003; 41(3): 66-674.

[26] Kaluli JW, Madramootoo CA, Djebbar Y. Modeling nitrate leaching using neural networks. Water Sci Tech 1998; 38: 127-34.

[27] Persson M, Sivakumar B, Berndtsson R, Jacobsen OH, Schjonning P. Predicting the dielectric constant-water content relationship using artificial neural networks. Soil Sci Soc Am J 2002; 66: 1424-9.

[28] Persson M, Uvo CB. Estimating soil solution electrical conductivity from time domain reflectometry measurements using neural networks. J Hydrol 2003; 273: 249-56.

[29] Minasny B, Hopmans JW, Harter T, Eching SO, Tuli A, Denton MA. Neural networks prediction of soil hydraulic functions for alluvial soils using multistep outflow data. Soil Sci Soc Am J 2004; 68: 417-29.

[30] Patel RM, Prasher SO, Goel PK, Bassi R. Soil salinity prediction using artificial networks. J Am Water Res Assoc 2002; 38: 91-100.

[31] Vieillefon J. Contribution à l'amélioration de l'étude des sols gypseux. Cahier ORSTOM, série Pédologie 1979 ; 17(3): 195-223.

[32] Porta J. Methodologies for the analysis and characterization of gypsum in soils: A review. Geoderma 1998; 87: 31-46.

[33] USDA. Diagnostic and improvement of saline and alkali soil. Agriculture Handbook N° 60, U.S. Dept. of Agriculture 1954.

[34] Lesch SM, Corwin DL, Robinson DA. Apparent soil electrical conductivity mapping as an agricultural management tool in arid zone soils. Comput Electron Agric 2005; 46: 351-78.

[35] Shahin MA, Maier HR, Jaksa MB. Evolutionary data division methods for developing artificial neural network models in geotechnical engineering. Department of Civil & Environmental Engineering. The University of Adelaide 2000, research report N° R 171.

[36] Hsu K, Gupta HV, Sorooshian S. Artificial neural network modelling of the rainfall-runoff process. Water Resour Res 1995; 31: 2517-30.

[37] Poff NL, Tokar S, Johnson P. Stream hydrological and ecological responses to climate change assessed with an artificial neural network. Limnol Oceanogr 1996; 41: 857-63.

[38] Hagan MT, Menhaj M. Training feed forward networks with the Marquardt algorithm. IEEE Trans Neural Netw 1994; 5: 989-993.

[39] Maier HR, Dandy GC. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Env Model Software 2000; 15: 101-23.

[40] Changhui P, Xuezhi W. Recent Applications of Artificial Neural Networks in Forest Resource Management: An Overview. In Environmental Decision Support Systems and Artificial Intelligence, Cochairs (eds.) 1999; 15-22. Technical Report WS-99-07, AAAI Press, Menlo Park, CA 1999.

[41] Kwok TY, Yeung DY. Constructive algorithms for structure learning in feed forward neural networks for regression problems. IEEE Trans Neural Netw 1997; 8: 630-45.

[42] Vauclin M. Méthode d'étude de la variabilité spatiale des propriétés d'un sol. Les colloques de l'INRA 1982; 15: 9-43.

[43] Hachicha M, M'Hiri A, Bouksila F, Bach Hamba I. Variabilité et répartition de l'argile et de la salinité dans le périmètre de Kalaât Landalous (Tunisie). Étude et Gestion des Sols 1997; 4(1): 53-66.

[44] Tangang FT, Hsieh WW, Tang B. Forecasting the regional sea surface temperatures of the tropical Pacific by neural network models, with wind stress and see level pressure as predictors. J Geophys Res Oceans 1998; 103: 7511-22.

[45] Hsieh W, Tang B. Applying Neural Network Models to prediction and data analysis in meteorology and oceanography. Bull Am Meteorol Soc 1998; 79 (9): 1855-70.

[46] Tokar SA, Johnson PA. Rainfall-Runoff modelling using artificial neural networks. J Hydrol Eng 1999; 4(3): 232-9.

[47] Brosse S, Giraudel JL, Lek S. Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. Ecol Model 2001; 146: 159-66.

[48] Shahin MA, Maier HR, Jaksa MB. Data division for developing neural networks applied to geotechnical engineering. J Comp Civil Eng 2004; 18(2): 105-14.

[49] Bowden GJ, Dandy GC, Maier HR. Input determination for neural network models in water resources applications. Part 1—background and methodology. J Hydrol 2005; 301: 75-92.

[50] Mekki I, Bouksila F. Vulnérabilité du milieu physique, pratiques des agriculteurs et performance du périmètre irrigué de Kalâat El Andalous, basse vallée de la Medjerda, nord de la Tunisie.. Acte du Séminaire International « Exploitation des Ressources en Eaux pour une Agriculture Durable ». Annales de l'INRGREF, 2008; numéro spécial 11: 74-88.

[51] Hachicha M, Bouksila F, Zayani K, M'Hiri A. Etude comparative de la perméabilité mesurée par les méthodes de Reynolds, Porchet et Mûntz dans le cas de sols argileux affectés par la salinité. Revue Sécheresse 1996; 3(7): 209-15.

Received: October 19, 2009 Revised: July 09, 2010 Accepted: July 12, 2010