ISSN (e): 2250 – 3005 || Vol, 04 || Issue, 5 || May – 2014 ||
International Journal of Computational Engineering Research (IJCER)

Elaboration of stochastic mathematical models for the prediction of parameters indicative of groundwater quality Case of Souss Massa – Morocco

Manssouri T. [1], Sahbi H. [1], Manssouri I. [2]
(1) Laboratory of Geo-Engineering and Environment, Faculty of Sciences, University Moulay Ismail, BP 11201, Zitoune, Meknes, Morocco.
(2) Laboratory of Mechanics, Mechatronics and Command, ENSAM, Moulay Ismail University, BP 4042, 50000, Meknes, Morocco.

ABSTRACT:
Groundwater is a real wealth which requires rational management, monitoring and control achievable by the various methods that can classify them according to their degree of water mineralization lasting quality. Indeed, to assess the quality of groundwater, the knowledge of a certain number of indicators, such as the Electrical Conductivity EC, Organic material OM and the amount of Fecal Coliforms FC is paramount.
This work seeks to analyze the prediction indicators of quality of groundwater Souss-Massa Morocco. Initially, methods based on neural models MLP (Multi Layer Perceptron) are applied for the prediction of quality indicators of groundwater.
The choice of the architecture of the artificial neural network ANN MLP type is determined by the use of different statistical tests of robustness, i.e. the AIC criterion (Akaike Information Criteria), the test RMSE (Root-Mean-Squarre error) and the criterion MAPE (Maximum Average Percentage Error). Levenberg Marquardt algorithms are used to determine the weights and biases existing between the different layers of neural network.
In a second step, a comparative study was launched between the neural prediction model MLP type and conventional statistical models, including total multiple linear regression.
The results showed that the performance of neural prediction model ANN - MLP is clearly superior than those established by the total multiple linear regression TMLR.

KEYWORDS**:** *Prediction, Neural Network MLP type, robustness tests, Multiple Linear Regression, indicators of quality of groundwater.*

I. INTRODUCTION

The management of water resources is nowadays one of the key global issues, both in agricultural and industrial activities and in terms of direct consumption of the population. Indeed, the regular growth in demand for water resources, for several decades, has had various problems, both qualitative and quantitative.

Water quality can be judged in relation to-three basic types of parameters which are:

- Organoleptic parameters which are designed to assess the quality if the water is pleasant to the senses of the observer either by sight or by smell or taste as well.
- Microbiological parameters: some microscopic beings live in water. Their concentration defines some quality of these waters; Fecal Coliforms (FC); Fecal streptococci (FS) can be found.
- The physico-chemical parameters: these parameters, global or specific, help to assess the ability of a water to the use for which it is destined. They can be divided into chemical parameters and physical parameters.

In addition, the quality of groundwater is defined by the components distributed and transported in the liquid medium. Concentration measurements of chemical indicators in water samples collected at various locations, positions near or distant from the source of infiltration are basic elements to establish and monitor quality.
In previous works, neural networks have found great success in the modeling and prediction, we include for example:

Perez et al (2001) [1] have proposed to predict the concentration of nitrogen NO2 and nitric oxide NO in Santiago based on meteorological variables and using the linear regression method and the method of network neural. The results showed that neural networks are the method that performs the prediction error the lowest compared to other methods.

In 2008, Nohair et al [2] use neural networks to predict changes in a stream based on climatic variables such as the temperature of the ambient air, the water flow temperature received by the stream. Two methods were applied: the first, iterative type uses the estimated day j to predict the value of the water temperature on day j + 1 value, and the second method much simpler to implement, is of estimating the temperature of all the days taken at one time.

Bélanger et al (2005) [3] also use multilayer neural networks to predict the temperature of the water from the hydrometeorological parameters based on the model of artificial intelligence and the traditional method of multiple linear regression. The results of this study show that artificial neural networks seem to give a fit to the data slightly better than that offered by the multiple linear regression.

In 2010, Cheggaga et al (2010) [4] show the possibility of using neural networks to non-recurring layers for extrapolation, prediction and interpolation of the wind speed in time and in space in 3D (radius r, height h, time t), based on neural network learning for a few days. This work has shown the possibility of using neural networks to non-recurring layers for extrapolation and interpolation prediction of wind speed.

El Badawi et al. (2011) [5] carried out a comparative study of the performance of two modeling methods used for the prediction of heavy metal concentrations in Moroccan river sediments using a number of physico-chemical parameters.

This study showed that predictive models established by the recent method, which is based on the principle of artificial neural networks, are much more efficient compared to those established by the method based on multiple linear regression. The performance of the method shows the existence of a non-linear relationship between the physico-chemical characteristics studied (independent variables) and metal concentrations in sediments of the watershed of OuedBeht.

Manssouri et al. (2013) [6] use two modeling methods for the prediction of meteorological parameters in general and the humidity in particular.At first, the methods are based on the study of artificial neural networks MLP types (multi-layer Perceptron) are applied for the prediction of moisture, of the Chaouene area in Morocco. In a second step the new architecture of neural networks proposed of MLP types, was compared to the model of Multiple Linear Regression (MLR).

Predictive models established by the method of neural networks MLP, are more efficient than those established by multiple linear regression, because good correlation was obtained with the parameters from a neural approach with a quadratic error from 5%.

The main objective of this paper is to apply stochastic mathematical models as a tool for prediction of parameters (microbiological, physical and chemical) quality indicators groundwater in Souss_Massa.

II. GEOGRAPHICAL LOCATION

The Hydraulic Basin Souss Massa (BHSM) is located between the High Atlas in the north and the ancient mountains of the Anti-Atlas to the south. It includes the catchments of Souss Massa Tamri, Tamraght and Atlantic coastal basins Tiznit-Ifni. The total area of the study area is almost 27 900 km ². The main plains of the region are the Souss (4500 km ²), the plain Chtouka (1260 km ²) and the plain of Tiznit (1200 km ²) see Figure 1.

The study area falls within the Prefectures of Agadir Ida Outanane and Inezgane Ait Melloul and provinces Chtouka Ait Baha, Taroudant and Tiznit. It also includes other provinces like Chichaoua and Essaouira.

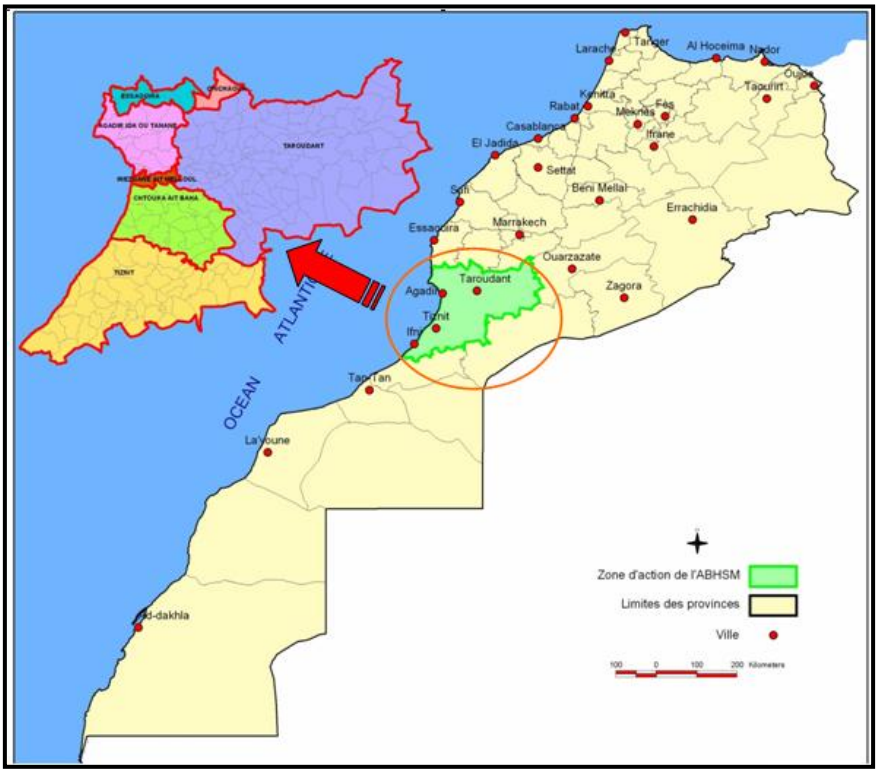

Figure 1.Location of the area of action of the Water Basin Agency of Souss Massa (ABHSM, 2004).

III. CLIMATE AND RAINFALL POTENTIAL

The climate of the region is predominantly arid, with strong sunshine (3089 hours / year). It is conditioned, including the influence of cold winds from the Atlantic Ocean (Canary Current) and Saharan latitude which gives the area a pre-Saharan climate to cool winters in the plains.

The temperatures are mild and regular, with an annual average ranging from 18.3 ° C to 20 ° C in Agadir at the dam on the Abdelmoumen river (Table 1). The maximum daily temperature reaches 49 ° C and the minimum temperature drops to 3 ° C below zero. The temperature ranges are also high and can reach 48 ° C according to a study prepared in 2004 by the Water Basin Agency of Souss Massa (BHSM Agency).

Table1. Annual average temperatures. (Agency ABHSM 2004)

Poste	Agadir	Barrage Abdelmoumen	Barrage Aoulouz	Taroudant	Barrage Youssef Ben Tachfine
T (° C)	18,3	20,6	19,9	19,7	19,9

The average annual evaporation ranges from 1400 mm in the mountains and near the Atlantic coast and 2000 mm in the plains. The minimum mean monthly evaporation is recorded in January with an average of 35 mm in the mountains and 100 mm in the plains, while the maximum was recorded in July with an average of 240 mm in the mountains and 270 mm in plain.

The humidity is quite high throughout the year on the ocean fringe, which allows maintaining a dense natural vegetative cover (effect of dew, fog and mist). The monthly average relative humidity is about 65%. Its maximum in July (73%), while the minimum occurs in December (58%).

The winds in the area are generally of two types:

Hot winds from EAST known as "Chergui" that occur in late spring to mid-autumn;

A sea breeze whose influence is felt in the coastal zone and to a depth of 25 to 30 km inland.

Rainfall is highly variable in space and time. The total number of rainy days is around 30 days per year on average, over the High Atlas is sprinkled with a number of rainy days in the order of 60 days.

Two rainfall seasons are distinguished in the region, namely:

- ✓ A wet season from November to March, during which the region receives 70-75% of the annual rainfall;
- ✓ A dry season, from April to October during which the region receives 25-30% of the annual rainfall.

The inter-average rainfall in the region varies from one watershed to another, as shown in Table 2.

	Souss	Massa	Tamraght	Tamri	Tiznit
Precipitations (mm)	280	265	390	370	145

Table2. Average annual rainfall. (Agency BHSM 2004)

I. Results and Discussion

Iv.1 Development of the Basis Learning and Test

Data Collection

The objective of this step is to collect data, both to develop different prediction models and to test them.

In the case of applications on real data, the objective is to collect a sufficient number of data to form a representative database that may occur during use of different prediction models. The data used in this study are related to the chemical analysis of parameters indicative of groundwater quality Souss Massa, made from 52 measurement points distributed over the entire groundwater heeling under the basin.

The dependent variables are the chemical characteristics determined in these water points such as Electrical Conductivity EC Organic material OM, the amount of Fecal Coliforms FC; Independent variables are sodium Na^+, potassium K^+, calcium Ca^{2+} the magnesium ion Mg^{2+}, the hydrogen ion HCO_3^-, sulfate ion SO_4^{2-}, the air temperature Ta (° C), the water temperature Te (° C), the potential hydrogen Ph, Fecal streptococci FS, Total Coliforms TC (Figure.1)

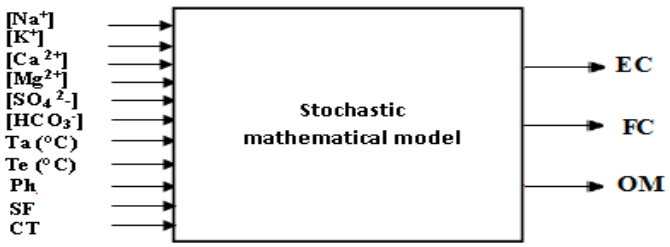

Figure.2: The inputs / outputs of the stochastic mathematical model

Data Analysis

After data collection, the analysis is necessary to determine the discriminate characteristics to detect and differentiate the data. These characteristics are the input of the neural network.

This determination of characteristics has an impact on both the size of the network (and thus the simulation time), on system performance (power separation rate prediction), and development time (learning time). Data filtering can be used to exclude those aberrant and / or redundant.

Data standardization

In general, the database must undergo pretreatment to be adapted to the inputs and outputs of stochastic mathematical models. A common pretreatment is to conduct a proper normalization that takes into account the magnitude of the values accepted by the models.

The normalization of each input xi is given by the formula:

$$x_i^{k}{}_{new} = 0{,}8 * \frac{x_i^{k}{}_{old} - \min(x_i)}{\max(x_i) - \min(x_i)} + 0{,}1$$

A standardized data between 0.1 and 0.9 is obtained. This database consists of different variations of the independent variables (model inputs) and dependents (model output).

Iv. 2. Total Multiple Linear Regression

To construct the linear mathematical model, the database should be divided into two bases: the learning that reflects 60% of the total database and the test-validation that reflects 40% of the total data base.
To build the first model we used the TLMR (Total Multiple Linear Regression).

The analysis by this method consists in finding a polynomial function, ie the value of each indicator given as a linear function of all the independent variables.

By performing the analysis by multiple linear regression with all the variables, we have obtained the equations 1, 2 and 3 respectively on the Electrical Conductivity EC, Organic Material OM and the amount of Fecal Coliforms FC from independent variables which are sodium Na^+, potassium K^+, calcium Ca^{2+}, magnesium ion Mg^{2+}, hydrogen ion HCO^{3-}, ion SO_4^{2-}, sulfate air Temperature Ta (° C) water temperature Te (° C), hydrogen potential Ph, Fecal Streptococci FS and Total Coliforms TC.

CE = -2,48716025993775E-02+0,333562079485328*[Na^+]+0,269703130591743*[K^+]-4,80221916946264E-02*[Ca^{2+}]-8,76866239678415E-02*[Mg^{2+}]-0,140917153044832*[HCO_3^-]+0,165424140021363*[SO_4^{2-}]-0,17659530501312* Ta +1,32439297208373E-02* Te +3,95855888033802E-02* Ph -1,67707116677135E-03* SF +0,557636121731416* CT **(1)**

R^2= 0,844 p < 0,0001

MO = -0,467919439118611+1,2145039855662*[Na^+]-1,20427213233719E-02*[K^+] -0,292644994647529*[Ca^{2+}]-0,564030336075162*[Mg^{2+}]+0,370232286748363*[HCO_3^-]-0,049368993591269*[SO_4^{2-}]+1,95020345380644E-02* Ta +3,59121215403218E-02* Te +6,57040240587299E-03* Ph -5,11243331413514E-02* SF +0,661995092580923* CT **(2)**

R^2= 0,584 p =0,057

CF = 0,302564389883223-0,124689894191023*[Na^+] +1,26022839275979E-02*[K^+] +9,11518997299649E-02*[Ca^{2+}] +4,67050632639108E-02*[Mg^{2+}]-1,85733043744043E-02*[HCO_3^-] -2,09539963169211E-02*[SO_4^{2-}] +3,26042977324909E-02* Ta +4,75125209774659E-02* Te +6,20911785378519E-02* Ph +0,172064089619721* SF +0,418787632774451* CT **(3)**

R^2= 0,541 p =0,067

The model on the Electrical Conductivity EC (1) is highly significant based on the fact that the probability is less than 0.0001. However other models are less significant than the first-mentioned model.

For the model (2) which connects the Organic Matter MO with all the independent variables, it is less significant since its probability is 0.057 and for the model (3) which connects the amount of Fecal Coliforms FC with all the independent variables as less significant because its probability is 0.067.

We conclude that the model (1) seems the most effective compared to the other two models. Indeed, the coefficients of determination of the first model is 0, 844, while for model (2) and (3) the coefficient of determination is 0.584 and 0.541.

IV.3-The Development Of Models Of Artificial Neural Network Mlp

Neural networks are powerful techniques of nonlinear data processing, which have proven themselves in many domains. Therefore, the artificial neural networks have been applied in various domains of prediction.

The various models of artificial neural network used in this work have been developed and implemented with the programming language C + + on an I3 PC machine 2.4 GHz and 3 Go of RAM.

The artificial neural network consists of an input layer, a hidden layer and an output layer. Input variables x = (x1, x2... x11) and independent normalized between 0.1 and 0.9 and then presented to the input layer of the neural network which contains eleven neurons.

They are first multiplied by the weight IW, and then added to bias IB through the input layer and the hidden layer. Neurons in the hidden layer receive the weighted signals. After addition, they transform them using a nonlinear sigmoid function S (.). Given by the equation:

$$S(n)=\frac{1}{1+\exp(-n)}$$

The following mathematical model S (IW * X + IB) is presented to the input of the output layer. This model will be multiplied by the weight WL then added to the bias LB that exist between the hidden layer and the output layer and then converted by a nonlinear sigmoid function S (.)

Finally, we obtain the mathematical model of artificial neural network as follows:

S {LW* S (IW*X+IB) + LB}.

The back propagation algorithm was used to form the artificial neural network in a fast and robust manner.

The analysis was restricted to networks that contain a single hidden layer, since this architecture is able to predict all outputs.

IV.4. Test Of Robustness

To select the "best" architecture of neural network, several statistical tests are commonly used; in our case we used statistical tests Root Mean Square Error RMSE, Maximum Average Percentage Error MAPE and Akaike Information Criteria AIC.

This last criterion we try to minimize, is proposed by Akaike [7], it is derived from the information theory, and relies on the measurement of Kullback and Leibler. It is a model selection criterion that penalizes models for which the additions of new variables do not provide enough information to the model. These tests are given respectively by the following equations:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{i=N}(E_{pi}-E_{ai})^2}{N}}$$

$$MAPE=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|E_{pi}-E_{ai}\right|}{E_{ai}}*100$$

$$AIC=\ln\left(\frac{N}{2}*L_F\right)+\frac{2N_w}{N}$$

With LF is the cost function (mean square errors).

$$L_F=\frac{1}{N}\sum_{i=1}^{i=N}(E_{pi}-E_{ai})^2$$

Ea and Ep are the values of the target vector and vector for predicting output neuron of the network. N represents the number of samples studied tests and Nw is the total weight and bias used for each architecture.

Figures (2-4) RMSE, MAPE and AIC-after 1500 iterations for different number of neurons in the hidden layer, give the opportunity to choose 19 neurons in the hidden layer. We then get the architecture [11-19-3] as "best" configuration of the neural network Due to its good predictive ability.

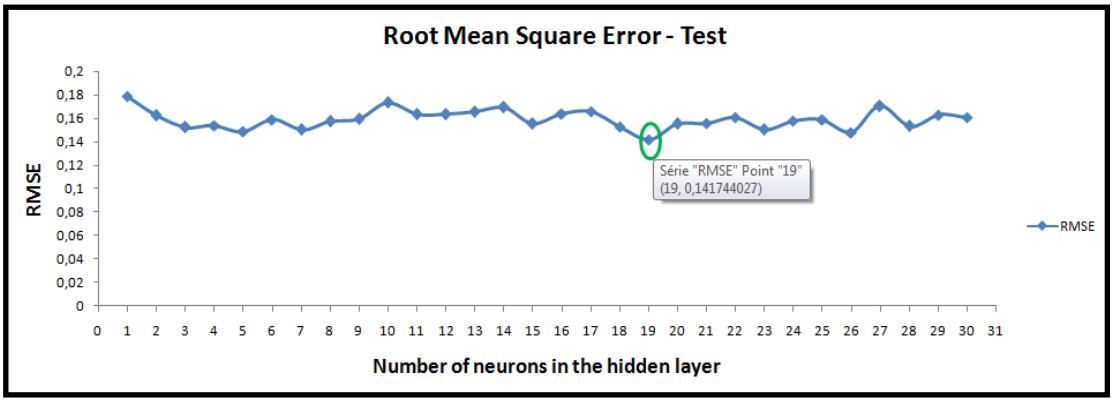

Figure 3. Robustness test: Root Mean Square Error

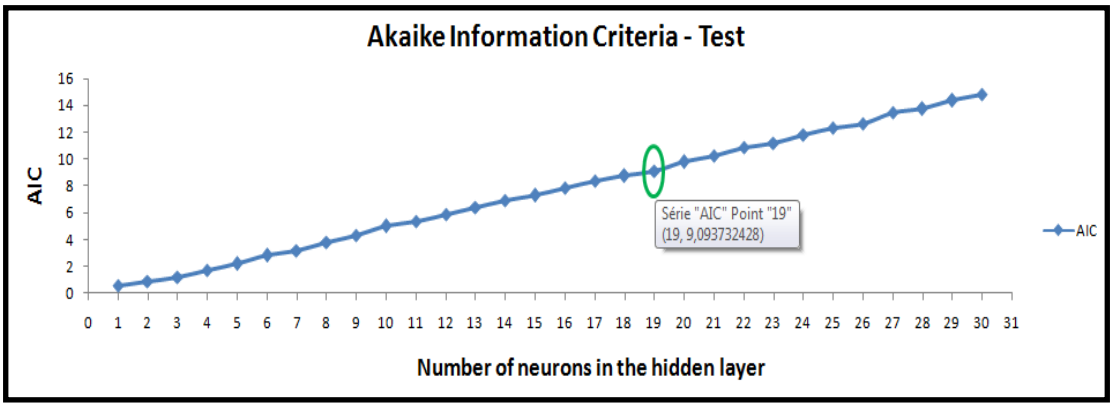

Figure 4. Robustness test: Maximum Average Percentage Error

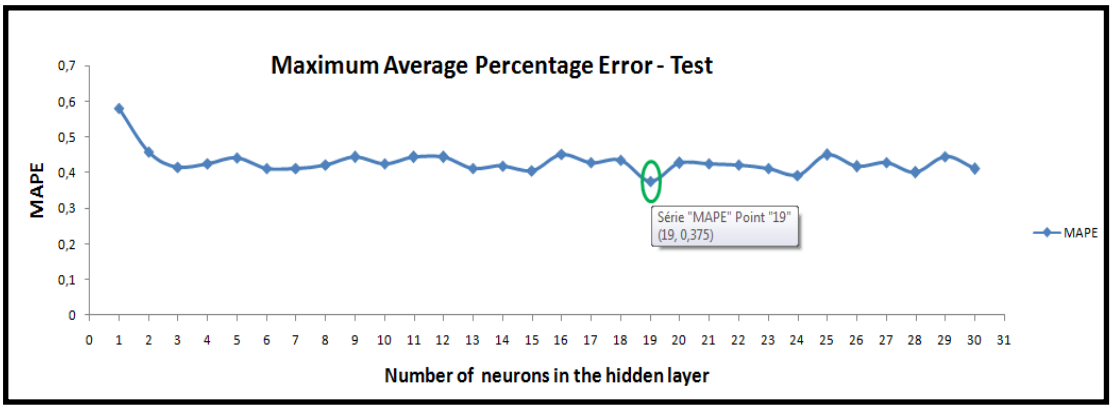

Figure 5. Robustness test: Akaike Information Criterion

IV.5-Learning And Validation

The artificial neural network MLP "Multi Layer Perceptron" consists of an input "input layer" containing eight neurons, a hidden layer "hidden layer" containing eight neurons and an output layer "output layer "containing three neurons (Figure 5).

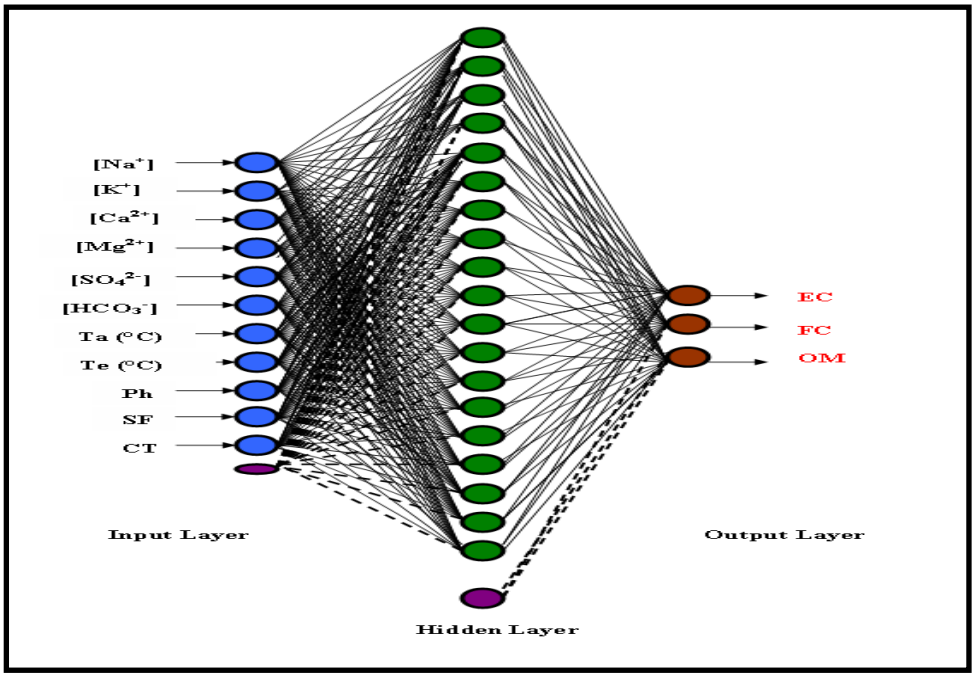

Figure 6: Neural Network architecture.

The basis learning consists of eleven vectors x1, x2, ..., x11, independent and normalized between 0.1 and 0.9 are: sodium Na^+, potassium K^+, calcium Ca^{2+}, magnesium ion Mg^{2+}, bicarbonate ion HCO^{3-}, sulfate ion SO_4^{2-}, air temperature Ta (° C), water temperature Te (° C), potential hydrogen Ph, Fecal Streptococci FS and Total Coliforms TC.

The basis learning, of the neural network consists of 32 samples. The weights and biases of the network were adjusted using the Levenberg Marquardt.

Once the architecture, weights and biases of the neural network have been set, we must know if this neural model is likely to be generalized.

The validation of neural architecture [11-19-3] is therefore to assess its ability to predict parameters indicating groundwater quality Souss Massa (EC, OM and FC) using the weights and biases calculated during learning to apply to another database of 20 samples test compounds, that is to say 40% of the total data.

The ANN model [11-19-3] gave a correlation coefficient for the testing and validation of 0.913117, which is equivalent to a mean square error of 0.14174, 0.375 for the average of the absolute errors and relative percentage of 9.09373 for the AIC.

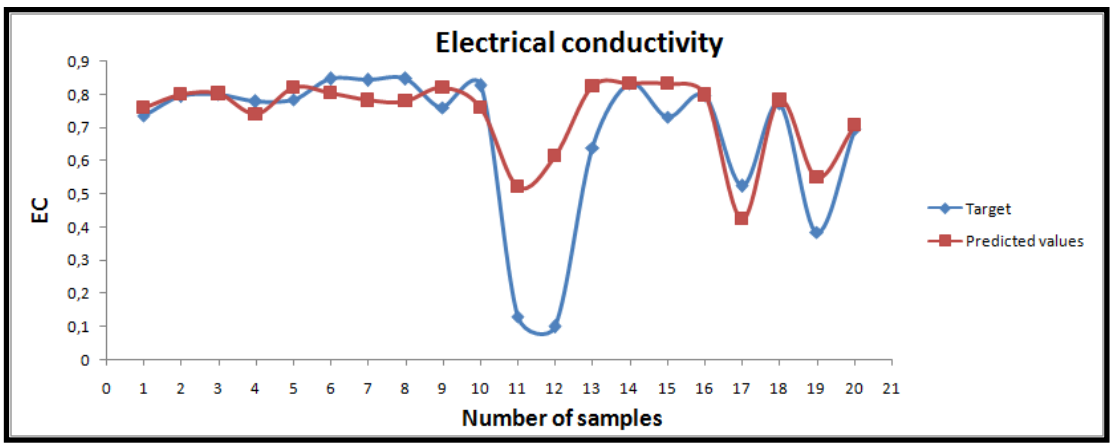

Figure 7: Result of test for predicting the Electrical Conductivity EC of the model ANN-MLP [11-19-3]

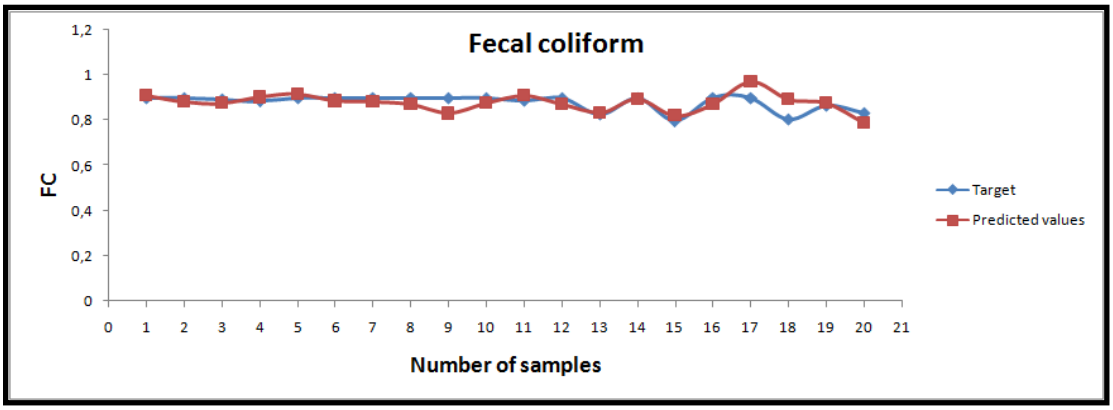

Figure 8: Results of prediction test the amount of Fecal Coliforms FC of the model ANN-MLP [11-19-3]

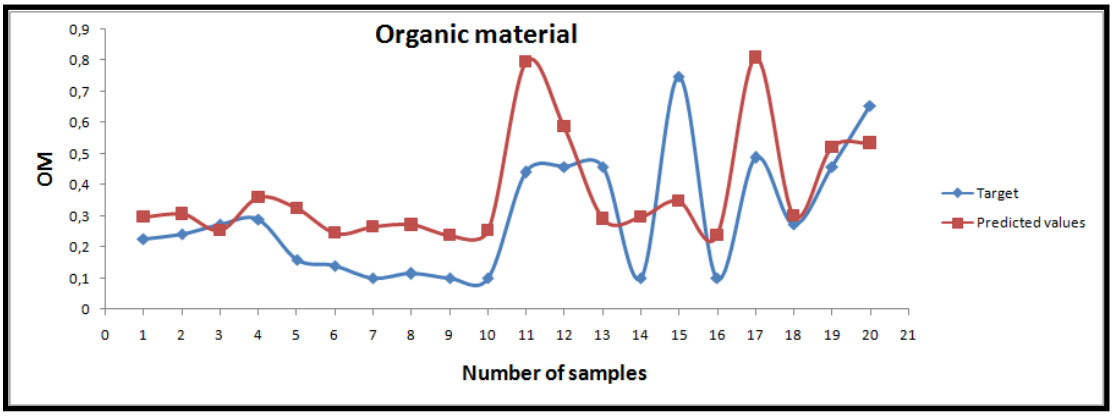

Figure 9: Test Result for predicting Organic Matter OM of the model ANN-MLP [11-19-3]

We evaluate the quality of prediction of the neural architecture [11-19-3] by the correlation coefficient. In our case the correlation coefficient is 91.3117%.

IV.5-Comparisons And Discussions

To assess the performance of the ANN-MLP method [11-19-3], we compared this method with other methods namely Multiple Linear Regression MLR.

The correlation coefficient calculated by ANN-MLP [11-19-3] is significantly higher (91.3117%), unlike the correlation coefficients calculated by the MLR are lower (between 0.541 and 0.844). On the other hand, the correlation coefficients obtained by testing the validity of the models established by the ANN (91.3117%) are significantly similar to those related to learning (99.1858%). However, the correlation coefficients of the tests of the validity of models for the MLR, are widely different from those obtained during training (see Table 3).

Method	EC		OM		FC	
	Learning	**Test**	**Learning**	**Test**	**Learning**	**Test**
RLM	0,844	0,6896	0,584	0,215	0,541	0,325
RNA	0,991858	0,913117	0,991858	0,913117	0,991858	0,913117

Table 3: Correlation coefficients obtained by MLR and ANN-MLP [11-19-3] on the Electrical Conductivity, Organic Matter and the amount of Fecal Coliforms.

V. CONCLUSION

We have been interested in a comparative study of the performance of two modeling methods used for the prediction of quality indicators of groundwater with hydraulic basin Souss Massa such as Electrical Conductivity EC, OM Organic Matter and the amount of Fecal Coliforms FC from a number of physico-chemical and microbiological parameters.

Both modeling methods are used; multiple linear regression and artificial neural networks (formal). The data is used for the analysis of water samples taken at several water points, distributed in space and time, of the study area.

The results obtained in this work show a significant capacity for learning and the prediction for indicators of water quality with a coefficient of determination of 91.3117%, equivalent to a mean squared error of 0 ,14174. Of 0.375 for the average absolute relative errors in percentage and 9.09373 for the AIC criterion for basic testing data used in addition to a better choice of the network architecture achieved through statistical tests of robustness. In multiple linear regression, the results are less significant with a coefficient of determination between 21.5% and 68.96%. This shows that the parameters are associated with indicators of groundwater quality in a non-linear relationship.

REFERENCES

[1]. Perez, P., Trier, A. (2001). Prediction of NO and NO2 concentrations near a street with heavy traffic in Santiago, Atmospheric Environment, 35: 1783-1789.
[2]. Nohair.M. (2008).Utilisation de réseaux de neurones et de la régularisation bayé sienne en modélisation de la température de l'eau en rivière. Revue des sciences de l'eau / Journal of Water Science, vol. 21, n° 3, 2008, p. 373-382., Faculté des sciences et techniques de Mohammedia, Maroc.11p.
[3]. Bélanger M. (2005). Estimation de la température de l'eau en rivière en utilisant les réseaux de neurones et la régression linéaire multiple. Revue des sciences de l'eau, Rev.sci.Eau 18/3: 403-421.
[4]. Cheggaga N., Youcef Ettoumi F. (2010). Estimation du potentiel éolien. Revue des Energies Renouvelables SMEE'10 Bou Ismail Tipaza 99 – 105.
[5]. Abdallaoui A., El Badaoui H. (2011). Prédiction des teneurs en métaux lourds des sédiments à partir de leurs caractéristiques physico-chimiques. Journal Physical and Chemical News. 58: 90-97.
[6]. El Badaoui H., Abdallaoui A., Manssouri I., Ousmana H., «The prediction of moisture through the use of neural networks MLP type », IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 11, Issue 6 (May. - Jun. 2013), pp 66-74, 2013.
[7]. AkaikeH..Information theory and the extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. (Eds: V.N.Petrov and F. Csaki). AcademiaiKiadó, Budapest, (1973), pp 267-281.