IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 1, Ver. II (Jan – Feb. 2016), PP 61-70
www.iosrjournals.org

Use of Artificial Neural Networks Type MLP for the Prediction of Phosphorus Level from the Physicochemical Parameters of Sediments

Monyr Naoual[1], Abdallaoui Abdelaziz[1*], El Hmaidi Abdellah[2]
[1] *Department of Chemistry, Research Team "Analytical Chemistry and Environment", Faculty of Science/ University Moulay Ismail, Morocco.*
[2] *Department of Geology, Research Team "Water Sciences and environmental engineering", Faculty of Science/ University Moulay Ismail, Morocco.*
**Corresponding author: a.abdallaoui@gmail.com*

Abstract: *The present work is a contribution on the development of mathematical models for predicting the phosphorus contents based on the physicochemical properties of the sediments of the reservoir of the water dam Sidi Chahed (Meknes, Morocco). For that purpose, artificial neural networks (ANN) of type multilayer perceptron (MLP) was used. The data base used corresponds to 118 samples of superficial sediments taken from several stations, and distributed in space and time at the level of the reservoir of the water power plant Sidi Chahed. This data base of the neural network, which was collected between 2010 and 2012, consists of the phosphorus content (variable to explain or to predict) and physical and chemical parameters as explicative or predictive variables. The performance of the mathematical models provided by artificial neural networks of type PMC was compared to the multiple linear regression model (MLP). This comparison showed that neural stochastic models are more efficient compared to the model on the MLR standard method, for the prediction of the phosphorus. This result can be explained by the existence of a non-linear relationship between the investigated physical and chemical parameters and the phosphorus contents of sediments from the dam's reservoir. The obtained results showed that the most efficient model is that of type PMC with the configuration [14-7-1], which uses, as transfer functions, the hyperbolic tangent function in the hidden layer and in the output layer, and learning algorithm of type quasi Newton BFGS.*

Keywords: *Artificial neural networks, Multiple linear regressions, Physical and chemical parameter, Phosphorus, Prediction.*

I. Introduction

Water pollution by chemical pollutants has become one of the most serious pollution, which affect notably industrial and urban areas. One of these chemical elements responsible for the pollution is phosphorus, which comes mainly from domestic waste (sewage, water treatment plants, etc.) and agricultural waste (direct discharges and diffuse pollution).

Moreover, phosphorus is an essential element for the proper maintenance of aquatic ecosystems. In an unpolluted lake, phosphorus arriving in lake is consumed almost immediately without creating surpluses. However, the addition of phosphorus in the watershed or directly into the lake by human activities can cause significant environmental degradation (eutrophication, water pollution, etc.). On the other hand, phosphorus has the property to bind to sediment, but can be resuspended during high wind events in shallow lakes. Moreover, when there is a lack of oxygen, phosphorus can be released into the waters of Lake thus causing the degradation of their quality.

The present work concerns the modeling and predicting of total phosphorus content based on physicochemical parameters of surface sediments of the reservoir of the water dam Sidi Chahed, using two modeling tool: multiple linear regression and artificial neural networks (ANN) of type multilayer perceptron (MLP). The comparative study of these two methods will test the performances of each of these two modeling methods.

II. Presentation Of The Study Area

Sidi Chahed dam is built on the Oued Mikkes, about 30 km North West of Fez city, on the main road linking it to Sidi Kacem city (Fig. 1). Its construction was mainly intended to supply the city of Meknes drinking water. Its storage capacity is 170 million m^3. However, the reservoir water quality was found unfit for human consumption due to its relatively high salinity **[1]** since it's commissioning in February 1997.

Indeed, drainage of very rich land in Triassic of rock salt and gypsum by the Oued Mellah (course salt water), the main tributary of Oued Mikkès, whose waters flow directly into the reservoir, is the main origin of salinity in the waters of the dam Sidi Chahed **[2]**.

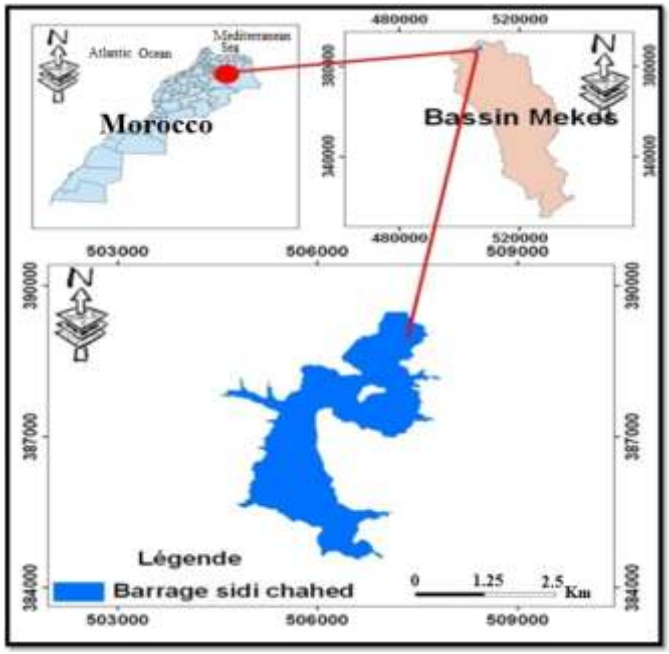

Figure 1 : Geographic situation of the dam Sidi Chahed relative to Morocco and its watershed.

III. Origin And Data Formatting

The data base used in this study is that relating to the analysis of 118 samples sediments taken from several station during several campaigns, and distributed in time and space, at the level of reservoir Sidi Chahed **[3;4;5;6;7]** It was collected between 2010 and 2012, and consists of the phosphorus content and physical and chemical parameters.

In total we chose fourteen independent variables (predictive) and a single dependent variable (to predict). These variables with their designations are recorded in "Table I ".

Furthermore, the input data which are the independent variables are untransformed raw values. To standardize the measurement scales, these data are converted into standardized variables. Indeed, the values X(i) of each independent variable (i) were standardized with respect to its mean and its standard deviation according to the relationship (1) **[8]**.

$$X_S(i) = \left(\frac{X(i) - X_m(i)}{\sigma_X(i)} \right) \qquad (1)$$

With:

$X_S(i)$: Standardized value relating to the variable i;

$X(i)$: Gross value relating to the variable i;

$X_m(i)$: Average value relating to the variable i;

$\sigma_X(i)$: Standard deviation relating to the variable i.

Table I. Variables used in this study and their designations.

Variables	Designations	Units	Variable Types
Depth	Dp	m	independent variables (predictive)
pH_{KCl}	pH_{KCl}	-	
pH_{Eau}	pH_{Eau}	-	
Water content	WC	%	
Conductivity	Cond	µs/cm	
Fine Fraction (< à 50µm)	FF	%	
Carbonates	$CaCO_3$	%	
Organic Materials	OM	%	
Total Organic Carbon	TOC	%	
Nitrogen	N_2	%	
Potassium	K	%	
Sodium	Na	%	
Calcium	Ca	%	
Magnesium	Mg	%	
Phosphorus	P	%	dependent variable (to predict)

The standardization of values for all variables aims to avoid exponential calculations,very large or very small, and limits the increase in the average variance **[9]**.

The corresponding values for dependent variables were also normalized in the interval [0, 1] to adapt to the requirements of the transfer function used by the neural networks (sigmoid function). The normalization was performed according to the relation (2) **[8]**:

$$Y_n = \left(\frac{Y - Y_{\min}}{Y_{\max} - Y_{\min}} \right) \qquad (2)$$

Y_n : Normalized value;

Y : Original value;

$Y_{\min}$: Minimum value;

$Y_{\max}$: Maximum value.

IV. Prediction Methods

Several methods are applied to address problems prediction and modelling of complex nonlinear systems. These methods are particularly useful when these systems are difficult to model using classical statistical methods **[10]**.

In this study, we are interested in a comparative study of the analysis of two methods for predicting phosphorus levels from the physicochemical parameters of sediments of the dam Sidi Chahed, using two modeling tools: multiple linear regression and of artificial neural networks MLP type.

Multiple Linear Regressions

Linear regression consists in describing the relationships between a dependent variable Y and several variables called independent variables $X_1; X_2; ...; X_i$.

Indeed, the multiple linear regression which is a data analysis method, is commonly used for establishing predictive models to the phenomena observed in the aquatic environment **[11]**.This method is a technique for defining a polynomial function and determines the most significant input variables. The model is written:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i + \varepsilon(X) \qquad (3)$$

With:

- i : number of independent variable;
- ε(X) : random noise (error term or regression residual), but which depends a priori point X of the data space defined by the values of X_1 ;
- Y : dependent variable;
- $X_1, X_2, ... X_i$: independent variables;
- β_0 : ordinate at the originally estimated;
- $\beta_1; \beta_2; ... ; \beta_i$: model coefficients.

Artificial Neural Networks

The neural networks are assemblies units of calculations called formal neurons, whose original inspiration was a model of human nerve cell.

The neuron is the fundamental cell of an artificial neural network, which is strongly connected to the elementary processors operating in parallel. It's a computation unit which receives a number of inputs (Xi) directly from the environment or upstream neurons (Fig. 2). When information comes from a neuron, we associate it a weight which represents the ability of the neuron upstream to excite or inhibit the neuron downstream. Each neuron has one output **[12;13]**.

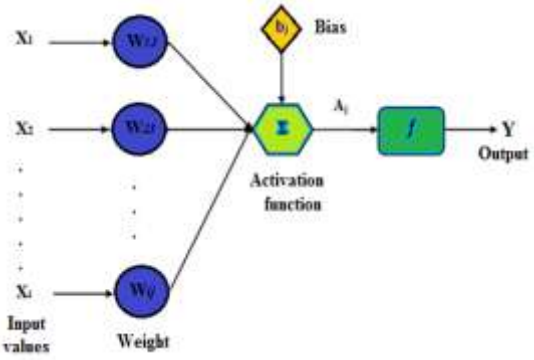

Figure 2 : Diagram showing the structure of an artificial neuron.

The connections between neurons which make up the network describe the topology of the model. In general, there are several neural network models are models Hopfield, Hamming, Carpenter, perceptron diaper and one multilayer perceptron **[14;15]**. To do more complex calculations, the most used networks are the multilayer networks (Fig. 3).

In multilayer networks, each neuron i receives a series of the signals of neurons j lying at the preceding layers (Fig.3). Each connection is assigned a weight.

The operation of an artificial neural network is governed by the following equation:

$$Y_j = \varphi\left(\sum_i W_{ij} X_i + \theta_j\right) \qquad (4)$$

- i : number of input neurons;
- j : number of hidden neuron;
- W_{ij} : *synaptic* weights;
- X_i : input values on the i variable;
- θ_j : bias (threshold) of neuron j.

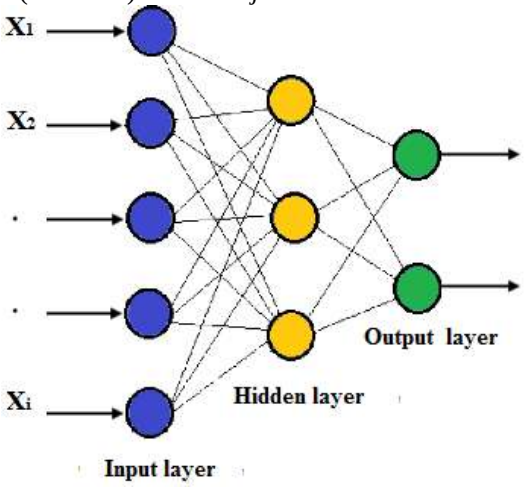

Figure 3 : Multilayer neural networks.

V. Results and Discussion

To develop mathematical models for predicting phosphorus levels from the physicochemical parameters of sediments of the dam Sidi Chahed we followed a statistical approach which is based on two methods: multiple linear regression (MLR) and artificial neural networks (ANN).

Multiple Linear Regression

Statistical analysis by the method of multiple linear regression was performed using Xlstat 2009 on all of the database software. Analysis by this method is to find the total phosphorus levels depending all the independent variables which are the physicochemical parameters.

$$\begin{aligned}\%P = {} & 0.84 - 1.88\times10^{-3}\times Dp + 3.89\times10^{-2}\times pH_{KCL} - 0.10\times pH_{eau} \\ & - 3.71\times10^{-3}\times WC - 5.43\times10^{-5}\times Cond + 7.43\times10^{-4}\times FF \\ & - 3.26.\times10^{-3}\times CaCO_3 + 4.45\times10^{-2}\times OM - 4.17\times10^{-2}\times TOC \\ & + 8.43\times10^{-2}\times N_2 - 0.11\times K + 4.8\times10^{-3}\times Na - 4.85\times10^{-3}\times Ca \\ & - 4.76\times10^{-2}\times Mg\end{aligned} \qquad (5)$$

{R^2= 0,476 ; Prob < 0,0001}

With:

R^2: coefficient of determination,

Prob : probability.

From this result, the model explains 47.6% of the variance (R^2 = 0.476), but the probability is very low value (Pr <0.0001). This means that the model is significant. Figure 4 shows the relationship between the observed and estimated levels of total phosphorus in sediment from the dam Sidi Chahed, the model established by the method of the RLM.

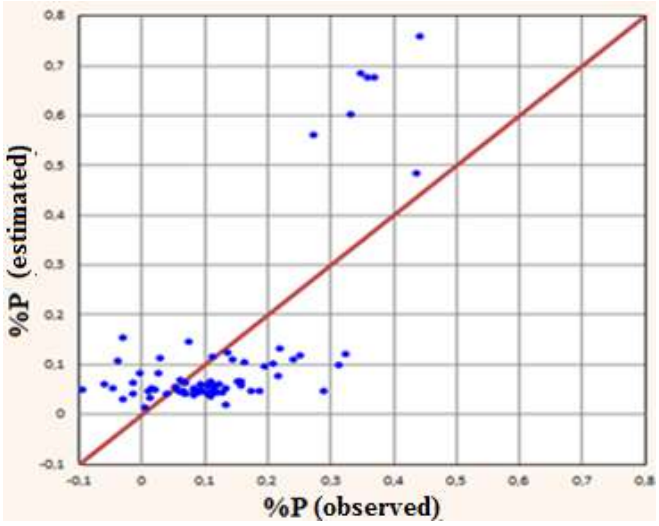

Figure 4 : Relationship between observed phosphorus levels and those estimated by multiple linear regression method (MLR)

Artificial Neural Networks

To establish the mathematical model for the prediction of the phosphorus with the method based on the principle of artificial neural networks (ANN), the computer software Statistica version 10 was used. This software uses a back-propagation algorithm with a supervised learning.

Statistica is an automated software tool, it also contains an automatic network search tool that can treat a large number of neural architecture of different complexity and can select the best set of specific architecture for a given problem. It saves time by automatically driving the heuristic search. This research concerns the type of network size and network architectures, activation functions, and even error functions in some cases. Use of this software has recently been described in detail by **[16]**.

To demonstrate the quality of predictive models, the data used in this study are divided into three groups. The first group corresponds to 60% of the total data. This group will be used to drive the system.

The second group corresponds to 20% of the total data and it will be used to validate the network. The remaining 20% that dit not participate in the learning models will be used as an independent test of network

generalization. It should be noted that these three groups of data were drawn from a random manner among the total database.

Several architectures were simulated for learning and validation. The tool ANS Statistica gives architects having the maximum correlation value and the minimum value of the mean squared error (MSE).

The algorithms that we have chosen to study are those of type BFGS quasi-Newton. The algorithms are available in its Matlab optimization toolbox. This type of BFGS algorithms takes its name from the initials of mathematicians C. G. Broyden, R. Fletcher, D. Goldfarb et D. F.S Hanno, who discovered independently in the late 60. This is one solution often used when one wishes a descent directions algorithm.

Table II presents the ten best architectures found by the tool ANS Statistica. It shows correlation coefficients and the error for each learning, test and validation according to the number of neurons in the hidden layer and the network topology. It also indicates the activation functions for the hidden layer and the output layer and the algorithm use.

The architecture of the most relevant neural network model for predicting phosphorus content is of type [14 -7 -1] with a coefficient of determination $R^2 = 0.999$ and a mean square error of $0{,}03\times10^{-4}$ The model is composed of:

- 14 neurons in the input layer containing the independent variables: physical and chemical parameters;
- 7 neurons in the hidden layer;
- 1 neuron in the output layer: phosphorus levels.

Table II. Summary of best performance obtained for each architecture models established by artificial neural networks type MLP with BFGS algorithm for predicting phosphorus

Network Architecture	**Activation functions**		**Coefficients of determination**			**Mean square errors x 10^{+4}**		
	Hidden layer	**Output layer**	**learning**	**test**	**validation**	**learning**	**test**	**validation**
[14-9-1]	Logistic	Identity	0,9992	0,9997	0,9973	0,15	0,05	0,03
[14-5-1]	Identity	Identity	0,9991	0,9999	0,9969	0,17	0,02	0,04
[14-7-1]	**Tanh**	**Tanh**	**0,9999**	**0,9985**	**0,9983**	**0,03**	**0,02**	**0,02**
[14-13-1]	Exponential	Identity	0,9993	0,9996	0,9973	0,15	0,01	0,05
[14-13-1]	Tanh	Identity	0,9993	0,9999	0,9977	0,12	0,04	0,03
[14-5-1]	Identity	Identity	0,9991	0,9999	0,9970	0,17	0,02	0,04
[14-5-1]	Identity	Identity	0,99901	0,9999	0,9972	0,18	0,02	0,04
[14-13-1]	Identity	Identity	0,9991	0,9999	0,9969	0,17	0,02	0,04
[14-14-1]	Identity	Identity	0,9991	0,9999	0,9971	0,17	0,02	0,04
[14-5-1]	Identity	Identity	0,9991	0,9999	0,9969	0,17	0,02	0,04

The model uses the same hyperbolic tangent activation function for the hidden layer and the output layer. The figure 5 shows the architecture of the developed model of neural networks.

Neural networks thus provide an improvement of 52.2% compared to multiple linear regression. This relatively significant improvement indicates a nonlinear relationship between the physicochemical parameters and the phosphorus content of the sediments of the reservoir of the dam Sidi Chahed.

The figure 6 shows the relationship between observed phosphorus levels and those estimated by the model established by artificial neural networks (ANN) for the entire database.

The figure 7 describes training of the network. It shows that after number of iterations 128, the desired result is achieved. With 7 hidden neurons, the two curves relating to the evolution of the mean square error of the two phases properly converge on the minimum mean square error (MSE).

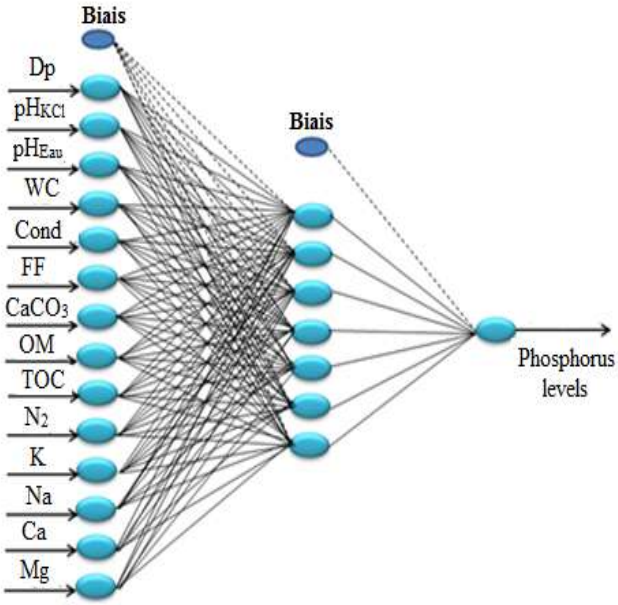

Figure 5 : Architecture of the neural network with three layer configuration [14-7-1] for predicting phosphorus levels.

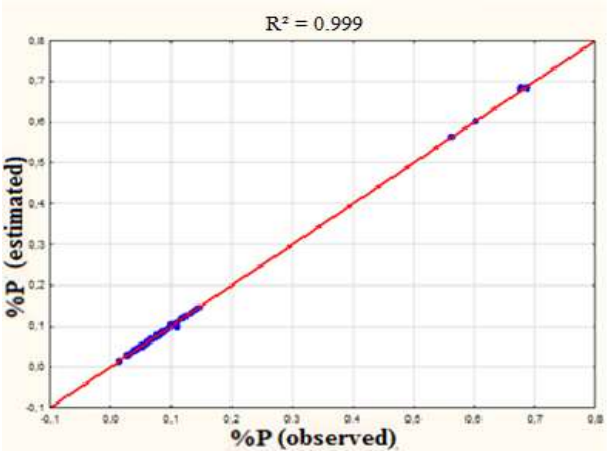

Figure 6 : Relationship between observed phosphorus levels values and those estimated by the ANN model established for the prediction of phosphorus.

Comparison Between The Two Models

Comparison of the results obtained by the different models shows that the model established by ANN neural networks is the most efficient than the model established by multiple linear regression MLR. This performance is due to the importance of the coefficient of determination is 0.999 for the first model and was 0.476 for the second, during the learning phases

Tables III and IV respectively represent the coefficients of determination and the mean square errors for the three samples of training, validation, and test, which are obtained by the models established respectively by the ANN and the MLR.

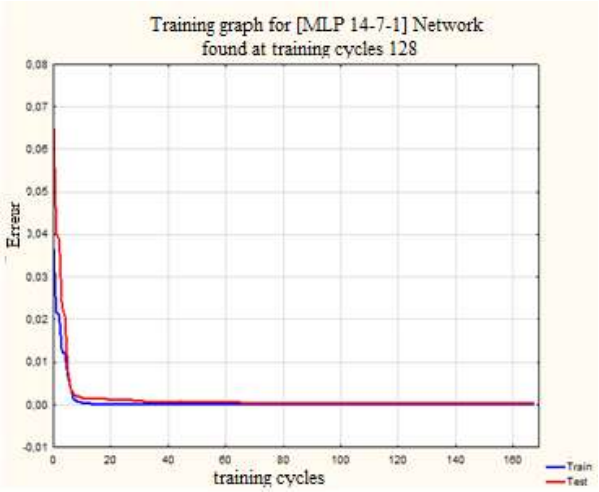

Figure 7 : Evolution of the mean square error with 7 neurons in the hidden layer for the case of phosphorus.

Table III. Coefficient of determination and the mean square error (MSE) obtained by the model established by the ANN for the three phases of learning, validation and test.

Phase	R^2	MSE×(10^{+4})
Learning	0.999	0.03
Validation	0.998	0.02
Test	0.998	0.02

Table IV. Coefficient of determination and the mean square error (MSE) obtained by the model established by the MLR for the three phases of learning, validation and test.

Phase	R^2	MSE×(10^{+1})
Learning	0.476	0.19
Validation	0.655	0.03
Test	0.772	0.01

These results show that the model established by neural networks keeps almost the same performance during the three phases of learning, validation and testing, as opposed to the model established by the multiple linear regressions, which shows a degradation of performance while passing of learning, validation and test.

Indeed, the coefficients of determination calculated by the ANN model type are significantly higher, for against the coefficients calculated by the MLR model type is lower. The parameters are thus non-linear because the coefficients are very high in the case of analysis with ANN and lower with the.

Residues Study

The error made by the models established by each method on an individual of sample of construction the model is called residue. The study of the latter is designed to test the validity of a model. Residues or "observed errors" are defined as the differences between the observed values and the values estimated by a model, they have the distinction of representing the portion not explained by the MLR model.

The residues analytical methods are mainly graphical analysis methods. The figure 8 shows the residuals relating to the model established by artificial neural networks and those relating to the model established by multiple linear regression based on the estimated values.

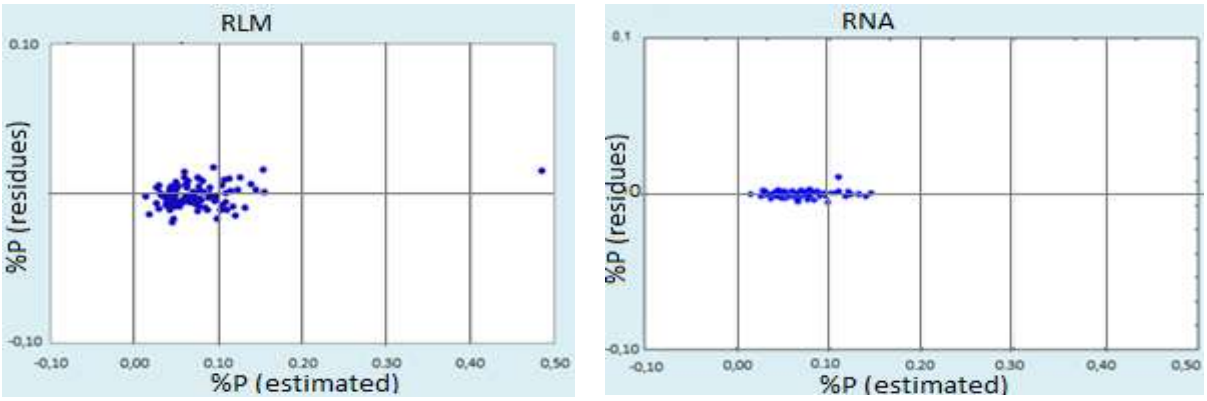

Figure 8 : Residues relating to the models established by multiple linear regression and artificial neural network depending of estimated values.

This figure shows that for phosphorus study, residues obtained by the model established by the neural network method are less dispersed (close to zero) against those obtained from the model established by multiple linear regression are more dispersed.

VI. Conclusion

Statistical analysis applied on the physicochemical parameters for the prediction of the phosphorus level was conducted using both modeling methods: the multiple linear regression and neural networks.

The results of these treatments have shown that predictive models established by the recent method, based on the principle of artificial neural networks are more performing compared to those established by the method based on multiple linear regressions. This performance seems to be due to the fact that phosphorus levels in the sediments of the dam Sidi Chahed, are linked to the physicochemical characteristics of the medium with non-linear relationships. Also, residues graphs showed the power of neural networks in data modeling.

On the other hand, we showed that the architecture for the establishment of the most powerful model with artificial neural networks to predict the phosphorus levels content of sediments of the dam reservoir Sidi Chahed, is one formed of three layers of configuration neurons [14-7-1], which uses, as transfer functions, the hyperbolic tangent function in the hidden layer and in the output layer, and learning algorithm of type quasi Newton BFGS.

References

[1] O. El Fatni, L. Gourari, D. Ouarhache, M. Chaffai, H. El Arabi, K. Boumir, and A. El Khanchou, Problème de salinité de la retenue du barrage Sidi Chahed: Etat des lieux et perspectives, 3rd International Days of Environmental Geosciences, El Jadida, Maroc, 2005.

[2] A. Dehbi, Etude des paramètres physico-chimiques et du phosphore total des sédiments superficiels de la retenue du barrage Sidi Chahed, Région de Meknès, Thése de Troisième siècle, Faculté des Sciences de Meknès, Chemie, 2011.

[3] D. Abrid, A. El Hmaidi, A. Abdallaoui, O. Fassi FihriI, and A. Essahlaoui, Impact de la pollution sur les eaux de l'oued Boufekrane (Meknès-Maroc): Etude physicochimique et bactériologique Physic. Chemical", Journal Physical and Chemical News, 58, 2011, 98-104.

[4] D. Abrid, A. El Hmaidi, A. Abdallaoui, and A. Essahlaoui, Apport du système d'information géographique à l'évaluation de la contamination métallique contenue dans les sédiments de la retenue du barrage Sidi Chahed (Meknès, Maroc), International Conference of GIS-Users, Taza GIS-days, Fès, Proceeding Book , 2012a, 396-398.

[5] D. Abrid, A. El Hmaidi, A. Abdallaoui, O. Fassi FihriI, and A. Essahlaoui, Etude de la contamination métallique des sédiments de la retenue du barrage Sidi Chahed (NE de Meknès, Maroc), Journal of Hydrocarbons Mines and Environmental Research, 3(2), 2012b, 55-60.

[6] D. Abrid, A. El Hmaidi, A. Abdallaoui, and A. Essahlaoui, Variation Spatiale des Concentrations en Éléments Traces Métalliques dans les Sédiments de la Retenue du Barrage Sidi Chahed (Meknès, Maroc), European Journal of Scientific Research, 106(4), 2013, 503-511.

[7] A. El Hmaidi, D. Abrid, A. Abdallaoui, and A. El ouali, Caractérisation sédimentologique et physico-chimique des sédiments de carottes de la retenue du barrage Sidi Chahed, NE de Meknès, Maroc, Journal of Hydrocarbons Mines and Environment Research, 3(2), 2012, 91-96.

[8] H. El Badaoui, and A. Abdallaoui, Prédiction des teneurs en métaux lourds toxiques des sédiments de l'oued Beht à partir des paramètres physico-chimiques, Journal Physical and Chemical News, 58, 2011, 90-97.

[9] F. Schwartz, Méthodologie de conception de systèmes analogiques. Utilisation de l'inversion ensembliste, Thèse de Doctorat d'Université, Université de Strasbourg, 2010.

[10] J. David, Développement d'une méthode d'étiquetage des jets de quarks b avec des muons de basses impulsions transverses, Thèse de doctorat, Université de la méditerranée Aix-Marseille II, 2010.

[11] A. Schmitt, B. Le Blanc, M.M. Corsini, C. Lafond and J. Bruzek, Les réseaux de neurones artificiels, Bulletins et mémoires de la Société d'Anthropologie, Paris, 2001, 1-2.

[12] K.D. Fausch, C.L. Hawkes, and M.G. Parsons, Models that predict the standing crop of stream fish from habitat variables 1950-85, General Technical Repport PNW-GTR-213, Department of Agriculture, Forest Service, Pacific Northwest Research Station, 1988.
[13] H. El Badaoui, Simulation numerique par l'utilisation des technique d'intelligences artificelles pour la modélisation des données météorologiques, Thèse de Doctorat, Faculté des Sciences de Meknès, 2014.
[14] B. El Mahdi, *Réseaux de neurones artificiels appliqués à la méthode électromagnétique transitoire InfiniTEM.* Mémoire, Université du Québec en Abitibi-Témiscamingue, Sciences appliquées, 2011.
[15] R.P. Lippmann, An introduction to computing with neural nets, IEEE Acoustics, Speech and Signal Processing Magazine, 1987, 4-22.
[16] M.T. Hagan, H.B. Demuth and M. Beale, Neural network design (PWS Publishing Company, Boston, Massachusetts, 1996).
[17] R. El Chaal, Prédiction des teneurs en métaux lourds dans les sédiments superficiels du barrage Sidi Chahed en utilisant les réseaux de neurones artificiels et la régression linéaire multiple, Mémoire de fin d'études master, Faculté des Sciences de Meknès, Geology, 2013.