Application of the Artificial Neural Networks of MLP Type for the Prediction of the Levels of Heavy Metals in Moroccan Aquatic Sediments

Hicham El Badaoui[1], Abdelaziz Abdallaoui[1], Imad Manssouri[2], L. Lancelot[3]

[1]laboratory Of Chemical Biology Applied To The Environment, Analytical Chemistry And Environment Team, Faculty Of Sciences, University Moulay Ismail, Bp 11201, Zitoune, Meknes –Morocco.

[2] Mechanical Engineering Department And Structures, Ensam, University Moulay Ismail, Bp 4042, 50000, Meknes, Morocco.

[3]laboratory Civil Engineering And Geo-Environment Polytech'lille, 59655 Villeneuve D'ascq, France.

ABSTRACT

The present work describes a new approach to the prediction of the concentrations of heavy metals in Moroccan river sediments relying on a number of physico-chemical parameters. The originality of this work lies in the application of neural networks the application of neural networks MLP type (Multilayer Perceptron). During the first step, the parameters of the neurons are determined by the method supervised. In a second step, the weights of output connections are determined using the algorithm of gradient back propagation. The data used as the basis for learning of the neuronal model are those related to the analysis of the sediment samples collected at the level of several stations, distributed in space and time, of the watershed of the river Beht of the region Khemisset in Morocco. The dependent variables (to explain or predict), which are three, are containing heavy metal (Cu, Pb and Cr) of sediments. A comparative study was established between the neuronal model for prediction of MLP type and conventional statistical models namely the MLR (Multiple linear regression).The performance of the predictive model established by the type MLP neural networks are significantly higher than those established by the multiple linear regression.

KEYWORDS *Heavy metals, Prediction, Physico-Chemical Parameters, ANN, Back propagation Gradient, MLP, MLR.*

I. INTRODUCTION

In Morocco, the metal contamination of aquatic ecosystems has attracted the attention of researchers from different backgrounds. It is indeed one of the aspects of pollution the greatest threat to these habitats. By its toxic effects, it can cause critical situations even dangerous. Unlike many toxic elements, heavy metals are not completely eliminated by biological means and consequently are subject to cumulative effect in the sediment. To predict the concentrations of these from a number of physico-chemical parameters, we refer to performing statistical methods, multiple linear regression and artificial neural networks [1], [2] and [3].We find in the literature several prediction methods proposed for the prediction of environmental parameters, we cite as examples some items proposed in the field of prediction using neural networks MLP and RBF type : Ryad et al. have worked on the application of neural network RBF (Radial Basis Function) for the prediction problem of a nonlinear system. The interest of this article lies in two aspects: a contribution at the recurrent network topology to accommodate the dynamic data and the second contribution for the improvement of the learning algorithm [4].

Perez et al. have proposed to provide for the concentration of NO_2 and nitric oxide NO in Santiago based on meteorological variables and using the linear regression method and neural network method. The results showed that the neural network MLP type is the method that achieves the lower prediction error compared to other linear methods [5].In this article, we used these methods to the prediction of the concentrations of heavy metals (Cu, Pb, Cr) in the sediments of the watershed of river Beht located in the north-west of Morocco, from a number of physico-chemical parameters.

II. MATERIAL AND METHODS

2.1. Data base

Our database consists of 104 samples [6] sediment collected at four sampling stations located upstream of the dam El Kansera (Fig.1). The independent variables are the physico-chemical characteristics determined in these samples of sediment organic materials (OM), water content (TE), the fine fraction (FF), pH, organic carbon (C), carbonates ($CaCO_3$), total phosphorus (P), calcium ($Ca^{2\ +}$), magnesium (Mg^{2+}), potassium(K^+), sodium (Na^+) and suspended solids (SS). When with dependent variables (to predict), they are three in number. These are the contents of heavy metals (Cu, Pb, and Cr) sediments.

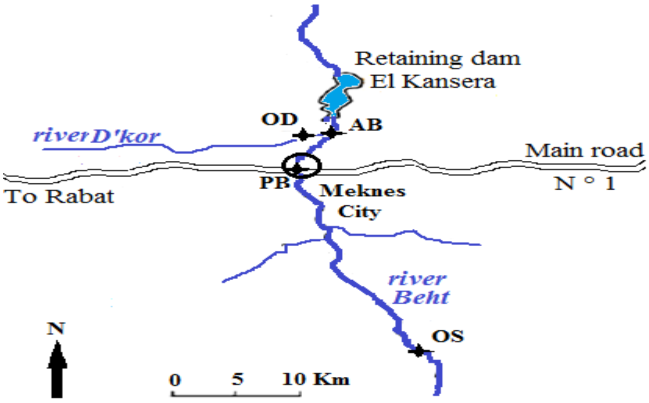

Figure 1. Location of sampling stations in the watershed of the river Beht.

2.2. Data modeling techniques

2.2.1. Methodology

Neural networks, with their parallel processing of information and their mechanisms inspired by nerve cells (neurons), they infer emergent properties to solve problems once described as complex. The mathematical model of an artificial neuron is shown in (Fig.2). A neuron consists essentially of an integrator which performs the weighted sum of its inputs. N the result of this sum is then transformed by a transfer function sigmoid f which produces the output of neuron.

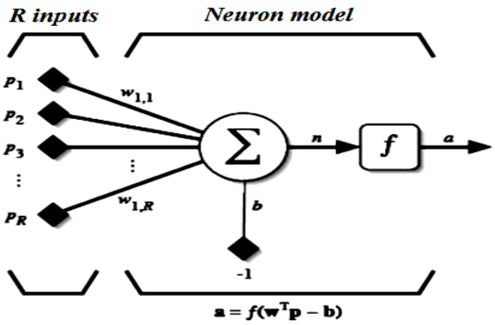

Figure 2. Model of an artificial neuron.

Learning is a phase during which the behavior of the network is changed until the desired behavior. It has several choices: error criterion to be achieved, the optimization algorithm of this criterion parameters of the algorithm and the range of values of random initial weights and thresholds [7]. Validation is the verification of the performance of a neural network-sample and generalization ability. Once the network is calculated, always conduct tests to verify that our system responds correctly. After that, comes the phase of implementation, which aims to demonstrate the performance and utility of this model [8].

2.2.2. Application of the model

Modeling data is performed in two steps. The first step is to compare the results obtained with the methods based on multiple linear regression and artificial neural networks, applying these methods to the data set on 104 samples. The second step is to justify the predictive quality of the models, using the same techniques on a set of data on 74 samples randomly selected among the 104 samples, which were the group for learning a model predictor of the dependent variable. The remaining 30 samples, which were not involved in the learning models, were used to test the validity and performance of the prediction models. Inputs (the physico-chemical parameters) and outputs (the contents of heavy metals) are normalized to a range [0 1] to adapt to the requirements of the transfer function used in this study (sigmoid function). The tool used for modeling, learning and visualization of the results obtained by neural networks is the MATLAB The dependent variable (to explain or predict) is the contents of heavy metals in sediments.

III. RESULTS AND DISCUSSION

3.1. Multiple linear regression (MLR)

We conducted an analysis by the MLR with all independent variables and we obtained the equations (1), (2) and (3) respectively for copper, lead and chromium.

[Cu] = 749.3716 - (22.0209 x MO) - (5.4306 x TE) + (1.8362 x FF) - (58.3936 x pH) + (12.8937 x C) + (0.0737 x P) - (2.2918 x $CaCO_3$) + (0.0014 x MES) - (0.5204 x Ca^{2+}) - (0. 6761 x Mg^{2+}) + (0. 0748 x Na^+) - (3. 4102 x K^+)

N= 74; R^2 = 0. 64; $p < 0.005$ (1)

[Pb] = 258.4691 - (7.1084 x MO) - (2.0632 x TE) + (0.6295 x FF) - (16.0839 x pH) + (6.1819 x C) + (0.0308 x P) – (1.2017 x $CaCO_3$) - (0.0064 x MES) - (0.2109 x Ca^{2+}) - (1. 4860 x Mg^{2+}) + (0. 0261 x Na^+) - (1.4544 x K^+)

N= 74; R^2 = 0. 69; $p < 0.005$ (2)

[Cr] = 22.3011 + (2.4554 x MO) + (0.0590 x TE) - (0.2643 x FF) - (1.3483 x pH) + (2.0420 x C) + (0.0215 x P) + (0.1012 x $CaCO_3$) - (0.0014 x MES) + (0.0386 x Ca^{2+}) - (0. 4798 x Mg^{2+}) - (0.0065 x Na^+) + (0. 7157 x K^+)

N= 74; R^2 = 0.87; $p < 0.005$ (3)

From these equations, we noticed that the three models for copper (1), chromium (3), and lead (2) are significantly important, because their probabilities are less than 0.005 (0.5 %).In fact, the model for chromium (3) is the most efficient, compared to models for copper (1) and lead (2). The correlation coefficient between observed and predicted concentrations of chromium is higher (R^2 = 0.87). Note on one hand, that the signs of the coefficients for the variables in the model (1) of copper, are almost similar to those of the model (2) lead. The coefficients for the variables (MO, TE, pH, $CaCO_3$, Mg^{2+} and K^+) are negative, whereas those for variables (FF, C, P, Na^+, $Ca^{2+)}$ are positive. However, they differ in the variable (MES). This analogy of signs shows the probable existence of a strong correlation between observed and copper levels observed in lead.

3.2. Artificial Neural Networks (ANN)

Since the laws of behavior of the environment are nonlinear and to model this type of problem, we are interested particularly to a typical neural network model known Multilayer Perceptron (MLP). To create the optimal structure of the neural network, we conducted several learning by varying the network parameters such as the activation function, number of hidden layers, and number of neurons in each layer, the learning function, the number of iteration and learning step [9]. However, we programmed the neural network, using the toolbox of Matlab neural network included in the MATLAB software.

3.2.1. Choice of the architecture model

Our choice is focused on a multi-layer non-recurring network, based on the learning algorithm of back propagation [10]. The purpose of this learning algorithm is to minimize the mean square error E (MSE). The network consists of three layers of neurons, called input layer, output layer and hidden layer. To determine the best number of neurons in the hidden layer, we varied the number between two and fifteen choosing the best combination of layer and the best distribution in each case. Based on the error values of MSE test shown in Table 1, we note that the minimum mean squared error is achieved when NHN = 3 and we noticed that increasing the number of hidden layers increases the load calculations without any performance gain. Finally, we can choose 3 neurons in the hidden layer of the network in this study to predict the contents of heavy metals (Fig. 3).

Table 1. Variation of the mean square error (MSE) as a function the number of neurons in the hidden layer on copper, chromium and lead.

Heavy metals	Cu	Pb	Cr
NNC	MSE Test	MSE Test	MSE Test
2	0.06029	0.05062	0.07651
3	**0.00236**	**0.00283**	**0.00246**
4	0.01441	0.01341	0.08665
5	0.01631	0.09103	0.04356
6	0.01759	0.07722	0.05156
7	0.00535	0.07837	0.09467
8	0.01791	0.07098	0.08605
9	0.07321	0.08576	0.07727
10	0.02634	0.09562	0.09471
11	0.05432	0.03452	0.08713
12	0.01944	0.05979	0.06242
13	0.01532	0.09611	0.05848
14	0.07095	0.03381	0.06563
15	0.06549	0.05148	0.07195

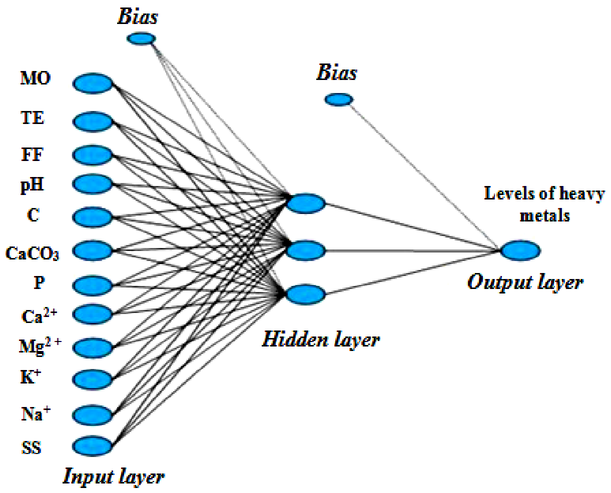

Figure 3. Architecture of ANN of topology [12, 3, 1] used in this study.

The curves shown in Figure 4 represent the evolution of the quadratic error of learning, validation and test according to the number of iterations. We note that the error learning, validation and test are very low and after 64 iterations, we have the stability of the network. Beyond this value it is necessary to stop learning an optimal number of times equal to 64 iterations (MSE = 0.00236). This phase allowed us to determine the optimal structure of our neural network. At the end of 64 iterations, the desired result is achieved with three hidden neurons; the three curves (error learning, validation and test) converge correctly.

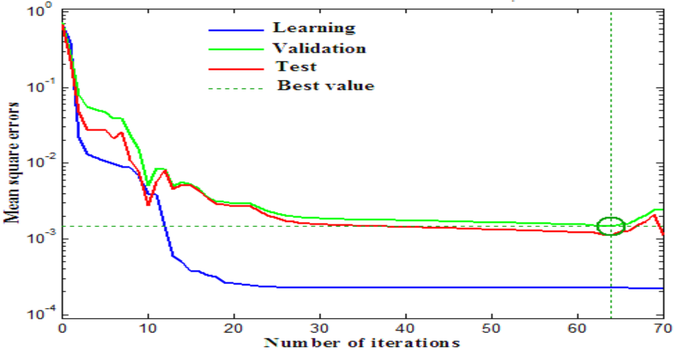

Figure 4. Evolution of the mean square error in the event of the copper during the learning, validation and testing with a network configuration [12- 3- 1].

The network has been driven up to the stage of learning, this has been met after 70 iterations, it is interesting to continue learning until they reach this stage for testing in order to reduce the gradient and therefore more perfect our network (Fig. 5).

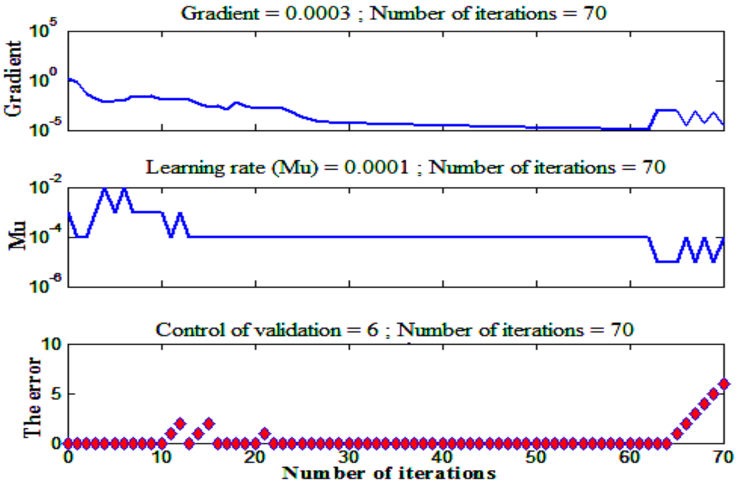

Figure 5. Variations of the gradient of the error, the learning rate and the validation error (for copper) as a function number of iterations.

Based on the results in Figure 5 we can conclude the different values of learning parameters found in this study:

✓The learning parameters are as follows:

- Maximum number of iterations (Epochs) = 70
- Mean square error (MSE) = 0.0023
- Rate of learning (Mu) = 0.0001
- Gradient minimum = 0.0003

To compare results between different numerical methods (Neural and multiple linear regression), two performance indices were calculated for each series: The coefficient of determination (R^2) and mean square error. The correlation coefficient (R^2) is the total error on the dependent variable y (The contents of heavy metals) explained by the model. This coefficient is expressed by [11]:

$$R^2 = 1 - \frac{\sum_{i,j=1}^{N}(Y_j - Y_{aver})}{\sum_{i,j=1}^{N}(Y_i - Y_{moy})}$$

The mean square error E is defined by the following equation) [10]:

$$MSE = E = \frac{1}{2}\sum_{i=1}^{N}(Y_j - Y_i)^2$$

Y_j : The output obtained by the network. Y_i : The target (desired output).

Y_{aver} : The average of measured values. N : Number of samples.

The coefficients of determination, calculated by the ANN were significantly higher (greater than 0.98), whereas the coefficients calculated by the MLR, they are lower (between 0.37 and 0.69). On the other hand, the coefficients of determination obtained by testing the validity of the models established by the ANN are clearly similar to those related to learning. However, the coefficients of determination relating to test the validity of models for the MLR, are widely different from those obtained during learning (Table 2).

Table 2. Coefficients of determinations obtained by MLR and ANN relating to copper and lead.

	Cu		Pb		Cr	
Method	Learning	Test	Learning	Test	Learning	Test
MLR	0.64	0.37	0.68	0.52	0.87	0.42
ANN	0.99	0.98	0.99	0.99	0.99	0.99

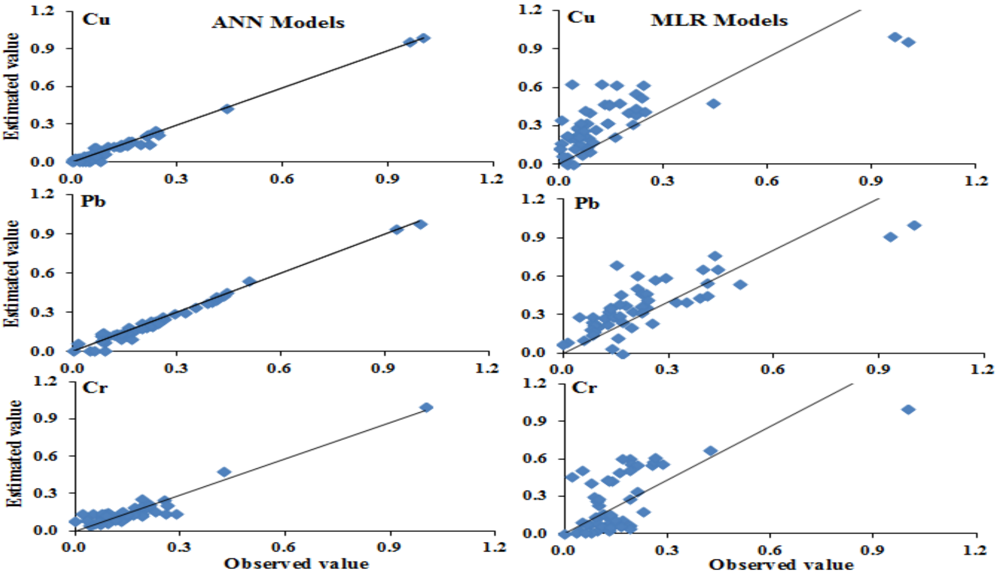

Figure 6. Relations between the observed and estimated levels of Cu, Pb and Cr established by MLR and ANN.

This figure shows, for both heavy metals studied, that the values estimated by the models established by the neural networks are much closer to the observed values, whereas the values estimated by the models established by multiple linear regression are widely further observed values, this shows a very good correlation between simulated and observed values with a very good correlation coefficient. This proves the predictive power of the models established by the neural network in the prediction of contents of heavy metals from the physico-chemical parameters of sediment in the watershed of river Beht. Previous studies have shown that the model developed in this study produced very good results compared with the method of multiple linear regression. For example Bouras and collaborators [8], showed an outstanding performance for the ANN model, this model can give better results compared to the linear method, especially for problems of prediction. Bélanger et al. [12], treat a comparative study of the performance of two modeling methods used to predict the temperature of the water, the results of this study show that artificial neural networks seem to fit the data little better than that offered by the multiple linear regression. The results we obtained from the models developed by ANN proved its accuracy, they are very close to the actual measurements.

3.2.2. Study of residuals

The error committed by the models established by each individual method on a sample of model construction is called residue [13]. Thus the study of the relation between metal contents estimated by mathematical models and their Residues (Yj - Yi) ensures the performance of the model, and it also allows to verify empirically, among other things, the validity of model assumptions.

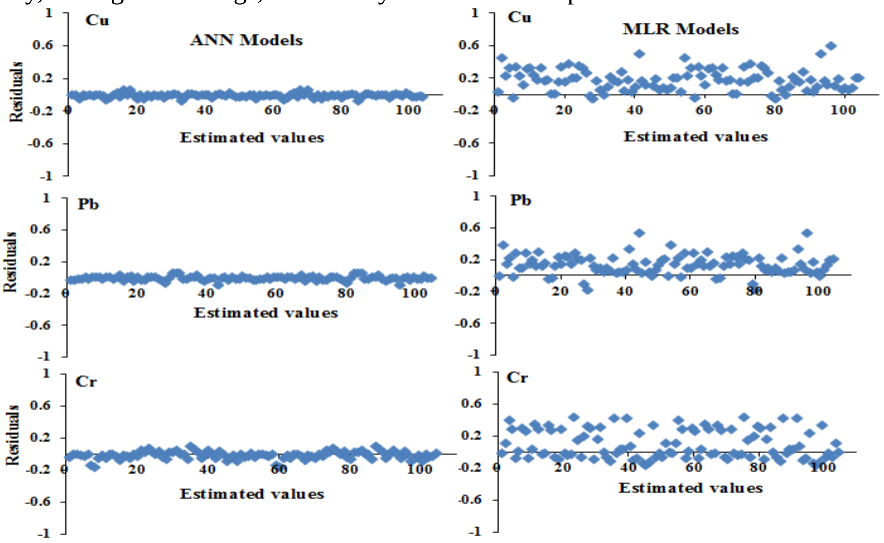

Figure 7. Relations between the estimated levels of Cu, Pb and Cr with the models established by the MLR and ANN and their residues.

Figure 7 presents the relations between the contents of heavy metals estimated using models established by neural networks (ANN) and those of multiple linear regressions (MLR) and their residues.

These figures show for the three heavy metals studied that the residuals obtained by neural networks are clearly less dispersed (closer to zero) and a significant improvement in the distribution of residues compared to those of multiple linear regression. This proves the predictive power of the models established by the neural network in the prediction of contents of heavy metals from the physico-chemical parameters of the sediments of river Beht. In general, the results are very satisfactory and justify the use of the approach by neural networks in the prediction of levels of heavy metals in sediments. This is in accord with the results of some recent studies demonstrated that multiple linear regression models are less efficient compared to those established by model neural network [12], [14] and [15].

IV. CONCLUSION

In this work we used neural networks to demonstrate that the contents of heavy metals in sediments are parameters which does not act alone but is explained by other physicochemical parameters. This study showed that the predictive models established by artificial neural networks are much more efficient compared to those established by the method based on multiple linear regression, of the fact that good correlation was obtained with the parameters from a neural approach, in addition to a better choice of network architecture that has been achieved through preliminary tests. The performance of neural networks demonstrates the existence of a non-linear relationship between the physico-chemical characteristics studied (independent variables) and the contents of heavy metals in sediments of the watershed of river Beht.

REFERENCES

[1] J. Garcia, M. L. Mena, P. Sedeno and J. S. Torrecilla. Application of artificial neural network to the determination of phenolic compounds in olive oil mill Waste water. Journal of Food Engineering, 544 – 552, 2006.

[2] J. Rude. Développement d'un modèle statistique neuronal pour la description fine de la pollution par le dioxyde d'azote. Thèse d'Etat, L'université paris XII-Val de Marne, Paris, 140p, 2008.

[3] A. Serghini, A. El Abadi, L. Idrissi, Mouhir, M. Fekhaoui et E. Zaïd, Evaluation de la contamination métallique des sédiments du complexe zones humides de la ville de Mohammedia (Maroc). Bulletin de l'Institut Scientifique. Section Science de la Vie. Rabat, n°23,77-81, 2001.

[4] Z. Ryad, R. Daniel, N. Zerhouni. Réseaux de neurones récurrents à fonctions de base radiales: RRFR. Revue d'Intelligence Artificielle. Vol X. 1-32, 2002.

[5] P. Perez, A. Trier. Prediction of NO and NO_2 concentrations near a street with heavy traffic in Santiago, Atmospheric Environment, 35: 1783-1789, 2001.

[6] A. Abdallaoui, Contribution à l'étude du phosphore et des métaux lourds contenus dans les sédiments et leur influence sur les phénomènes d'eutrophisation et de la pollution. Thèse Doctorat d'Etat. Faculté des Sciences Meknès, 255p, 1998.

[7] A. Darbellay, M. Slama. Do neural networks stand a better chance. International Journal of Forecasting, 16: 71–83, 2000.

[8] F. Bouras, Y. Djebbar, H. Abida. Estimation de L'envasement des Barrages: une Approche non Paramétrique. Journal International Network Environmental Management Conflicts, Santa Catarina – Brazil, 113-119, 2010.

[9] A. Zouidi, A. Chaari, M. Stambouli and F. Fnaiech. Nonlinear continuous time modeling of a high pressure mercury vapor discharge lamp using feed forward back-propagation neural networks. Yasmine Hammamet, Tunisie, 2004.

[10] N. Cheggaga, F.Youcef Ettoumi. Estimation du potentiel éolien. Revue des Energies Renouvelables SMEE'10 Bou Ismail Tipaza, 99 – 105, 2010.

[11] M. Nohair, A. St-Hilaire et T. B. Ouarda, Utilisation de réseaux de neurones et de la régularisation bayé sienne en modélisation de la température de l'eau en rivière. Journal of Water Science. Vol 21.373-382, 2008.

[12] M. Bélanger, N. El-Jabi, D. Caissie, F. Ashkar, J-M. Ribi, Estimation de la température de l'eau en rivière en utilisant les réseaux de neurones et la régression linéaire multiple. Rev.sci.Eau, 18/3 : 403-421, 2005.

[13] C. Voyant. Prédiction de séries temporelles de rayonnement solaire global et de production d'énergie photovoltaïque à partir de réseaux de neurones artificiels. Thèse pour l'obtention du grade de docteur en physique, Université de Corse-Pascal Paoli école doctorale environnement société / UMR CNRS 6134 (SPE), 257 pages, 2011.

[14] A. Abdallaoui, H. El Badaoui, L. Lancelot. Development of a neural statistical model for the prediction of the concentrations of heavy metals (Fe, Mn and Al) in surface waters: case of the Oued Boufekrane, (Meknes, Morocco). Colloque International Eau et climat, Regards croisés Nord/Sud, Rouen/ France, 2012.

[15] H. El Badaoui, A. Abdallaoui, L. Lancelot. Application des réseaux de neurones artificiels et des régressions linéaires multiples pour la prédiction des concentrations des métaux lourds dans les sédiments fluviaux marocains 20ème édition du colloque Journées Information Eaux, Université de Poitiers / France, 2012.