Development of Stochastic Mathematical Models for the Prediction of Heavy Metal Content in Surface Waters Using Artificial Neural Network and Multiple Linear Regression

Rachid El Chaal[1,*] , and *Mouley Othman* Aboutafail[1]
[1]Data Analysis, Mathematical Modeling, and Optimization Team, Department of Computer Science, Logistics, and Mathematics, Ibn Tofail University, National School of Applied Sciences ENSA, Kenitra 14 000, Morocco

Abstract. The principal purpose of this study is to build stochastic neuronal models, for the prediction of heavy metal, contents in the surface waters of the Oued Inaouen catchment area of the TAZA region, according to their Physico-chemical parameters; we have carried out a comparative study: the multiple linear regression (MLR) method and the artificial neural network (ANN) approach. The following statistical indicators were used to evaluate the performance of the stochastic models developed by neural network and MLR: The sum of the quadratic errors (SSE) and the determination coefficient (R^2), also through the study of fit graphs. The results show that the predictive modelling using artificial neural networks is very effective. This performance shows a non-linear relation between the studied Physico-chemical characteristics and the heavy metal contents in the surface waters of the Oued Inaouen catchment area.

1 Introduction:

The sustainable management of natural resources and the protection of our environment is considered in the environmental charter of our country Morocco [1], which is an introduction to the development and management of nature's resources. Water pollution can result from the contamination of wastewater, various discharges of industrial and domestic products, etc., among which metals occupy a very privileged place, as they are widely used by man in several fields.

All these discharges, in particular discharges of toxic micro-pollutants, i.e., heavy metals, cause water pollution with all the risks that this entails in terms of hygiene, biological life and environmental protection.

Indeed, water transports these heavy metals and inserts them into the food chain (algae, fish, etc.). These metals are most often present in trace amounts, but their toxicity develops through bioaccumulation in organisms. As a result, they are becoming an increasingly worrying problem[2].

The toxicity of heavy metals has led public authorities to regulate emissions by setting limit levels. Aware of this problem, the Moroccan public authorities and the national scientific community are increasingly interested in environmental studies, with a view to assess risks and protect our ecosystem [3].

The present study proposes to develop mathematical models for the prediction of heavy metal contents (zinc, Copper, and manganese) from a particular environmental parameter. The investigation concerns the surface waters of the Oued Inaouen catchment area.

To achieve this modelling objective, a comparative study is being conducted between two prediction models, namely multiple linear regression, and artificial neural networks, using the Statistica10 software.

2 Materials and methods

2.1 Database:

Our database consists of 100 samples (observations) of surface water taken in the province of Taza. From 2014 to 2015, the collection, transport and conservation of water samples refer to the protocol and procedures defined by the National Drinking Water Office. A part analysis a part was done there, and another part was made at the Regional University Interface Center (CURI) laboratory supported by

*Corresponding author: rachid.elchaal@uit.ac.ma

the University Sidi Mohamed Ben Abdellah (USMBA) of Fez.

2.2 Selection of Inputs:

The independent (explanatory) variables are the Physico-chemical characteristics determined in these samples, which are sixteen: Temperature (T ° C), pH, dissolved oxygen (DO), Conductivity (Cond), total dissolved solids (TDS), Bicarbonate (HCO3), Total Alkalinity (as CaCO3), Magnesium (Mg), Sodium (Na), Potassium (K), Chlorides (Cl), Calcium (Ca), Sulfates (SO4), Nitrate (NO3), Phosphorus (P) and Ammoniacal nitrogen (NH4),

The three dependent variables (to be predicted) are the levels of heavy metals: Manganese (Mn), Zinc (Zn) and Copper (Cu).

The database distribution is as follows: 70% of the data samples, selected randomly from the entire database, for the training phase of a forecast model of the dependent variable.

The remaining 30% of the samples were used to verify network performance while training the network and to avoid over-learning. The aim is to test the predictive validity and effectiveness of these models (15% for the test and 15% for the validation).

2.3 Data Formatting [4]–[6]

The normalization is a method of preprocessing data that helps reduce the complexity of models. The input data (16 independent variables) are raw, untransformed values. They have very different orders of magnitude. To normalize the measuring scales, the data are converted into a standardized variable. Indeed, the values of each independent variable (i) have been normalized to its means and standard deviation according to the relation:$\boldsymbol{X(v_i)}$

$$\mathbf{X_s(v_i)} = \left(\frac{\mathbf{X(v_i)} - \overline{\mathbf{X}}\mathbf{(v_i)}}{\boldsymbol{\sigma}\mathbf{(v_i)}}\right) \qquad i \in \{1.....16\} \qquad \textbf{(1)}$$

With:

$\mathbf{X_s}(\boldsymbol{v_i})$: Standardized value relating to the variable i.

$\mathbf{X}(\boldsymbol{v_i})$: Observed value relating to variable i.

$\overline{\boldsymbol{X}}(\boldsymbol{v_i})$: Average value relating to variable i.

$$\overline{\boldsymbol{X}}(\boldsymbol{v_i}) = \frac{1}{100}\sum_{k=1}^{100} \boldsymbol{X_k}(\boldsymbol{v_i}) \qquad \textbf{(2)}$$

$\boldsymbol{\sigma}(\boldsymbol{v_i})$: standard deviation relating to the variable i,

$$\boldsymbol{\sigma}(v_i) = \sqrt{\frac{1}{100}\sum_{k=1}^{100}(\mathbf{X_k(v_i)} - \overline{\mathbf{X}}\mathbf{(v_i)})^2} \qquad \textbf{(3)}$$

The purpose of normalizing the values for all variables is to avoid very large or minimal exponential calculations and to limit the increase in variance with the mean.

The values for the dependent variables were also normalized to the range of values [0;1] to meet the requirements of the transfer function used by the neural networks. This normalization was performed using the following relationship:

$$\mathbf{Y_n} = \left(\frac{\mathbf{Y} - \mathbf{Y_{min}}}{\mathbf{Y_{max}} - \mathbf{Y_{min}}}\right) \qquad \textbf{(4)}$$

With:

$\mathbf{Y_n}$: Standardized value.

$\mathbf{Y}$: Original value.

$\mathbf{Y_{min}}$: Minimum value.

$\mathbf{Y_{max}}$: Maximum value.

2.4 Data modelling techniques

Modelling approaches that have been developed to establish the best relationship between Manganese(Mn), Zinc(Zn) and Copper (Cu)and significant elements are based on two methods[7]:

- Multiple Linear Regression
- MLP type Networks of artificial neurons.

3 Results and Discussion

The results reported in Table 1 are found by the STATISTICA neural network tool. They show the determination coefficients, the number of iterations and the activation functions of the two layers depending on the topology; the algorithm used is Broyden-Fletcher-Goldfarb-Shanno (BFGS).

We modified the architecture of the network, by playing on the total quantity of hidden neurons and, or the number of training cycles (number of iterations). To do this, we have consecutively modified the number of hidden neurons (1, 2, 3, ..., 15). The results showed that the minimum mean square error, the number of iterations and the maximum determination coefficient are reached when the number of hidden neurons equals eight for Cu; 11 for Mn and 15 for Zn, for the best predictive model of heavy metal concentrations.

Figure 1 describes the training of the network in the case of Copper. After 102 steps, the required result is attained. With eight neurons in the cache layer, the two curves (learning error and test) converge correctly.

Table 1. Variation of the root mean square sum of errors (RMSE), number of iterations and the determination coefficient R^2 as a function of the number of hidden layer neurons(NHN) for Copper, Manganese and Zinc.

	NHN= 8			**NHN = 11**			**NHN = 15**		
Metals	R^2	NI	RMSE	R^2	NI	RMSE	R^2	NI	RSME
Cu	0.99	102	0.0001	0.98	199	2.07	0.95	81	5.96
Mn	0.93	75	4.25	0.99	228	0.003	0.96	123	1.13
Zn	0.97	58	0.98	0.94	12	2,98	0.99	217	0,001

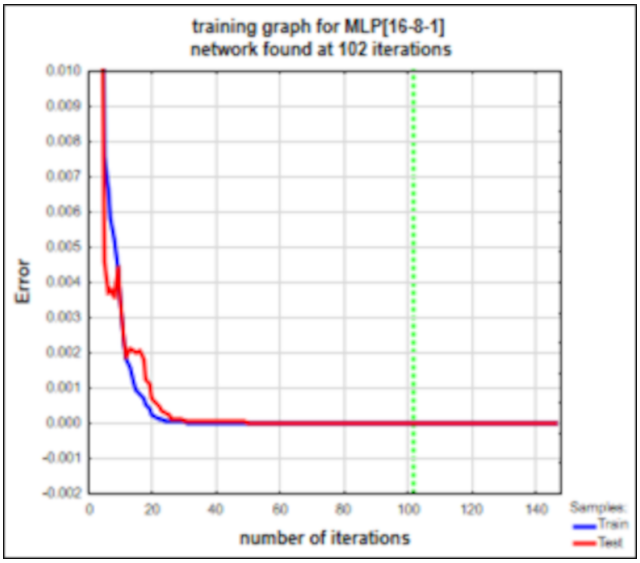


Fig.1. Development of the RMSE in the case of copper with eight neurons in the cache layer

The network has been trained to the point of overlearning; this, phenomenon has been encountered in 102 iterations.

The models established by neural networks allow improvements of up to 82% in the explanation of variance compared to those found by multiple linear regression. It is 77% for Manganese, 82% for copper and 59% for Zinc. The coefficients of determination increase respectively from 0.224 to 0.99, from 0.172 to 0.99 and from 0.40 to 0.99. The results mentioned in Table 2 show that the models established by the ANN are better than those found by the MLR method. Indeed, the coefficients of determination calculated for the models established by the ANN are significantly higher (above 0.9), in comparaison the coefficients calculated for the models established by the MLR are lower (between 0.17 and 0.40). This can be easily seen by referring to Fig.3 and the following Table 2.

Table 2. Coefficient of determination was obtained between the observed heavy metal contents predicted by the MRL and the ANN.

	Copper	Manganese	Zinc
MLR	$R^2 = 0.172$	$R^2 = 0,224$	$R^2 = 0,40$
ANN	Architectures of the selected models		
	[16 - 8 -1]	[16-11 -1]	[16 - 15 -1]
	Learning		
	$R^2 = 0,99$	$R^2 = 0,99$	$R^2 = 0,99$
	SSE = 0,00012	SSE = 0,0039	SSE=0,0017
	Test		
	$R^2 = 0,99$	$R^2 = 0,99$	$R^2 = 0,99$
	SSE = 0,0082	SSE = 4.29	SSE = 3.90

Based on the determination coefficients found when checking the validity of the models developed by the ANN are close to those relating to learning. On the other hand, the determination coefficients relating to testing the validity of the models concerning MLR are somewhat different from those obtained during training. This shows that the concentrations of heavy metals in the waters of the Oued Inaouen catchment area are linked to the Physico-chemical characteristics of the environment by non-linear relations. What is commonly found in the aquatic environment[8], [9]

Fig.2 shows the ANN architecture we used for copper prediction. It is a topology network [16-8-1].

.

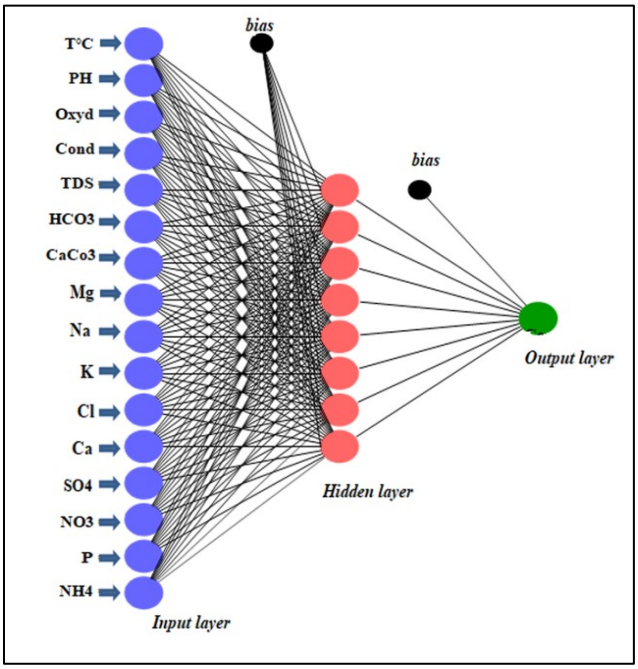
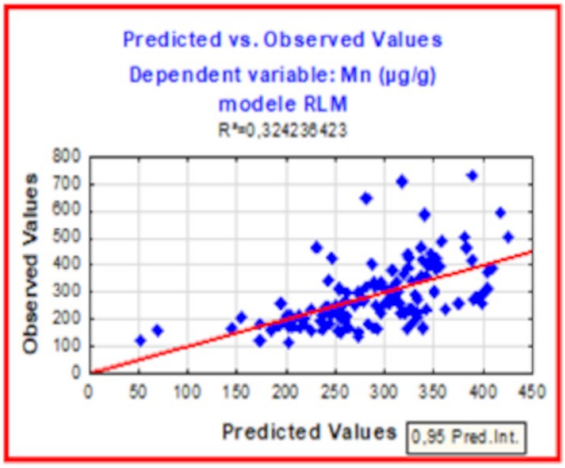
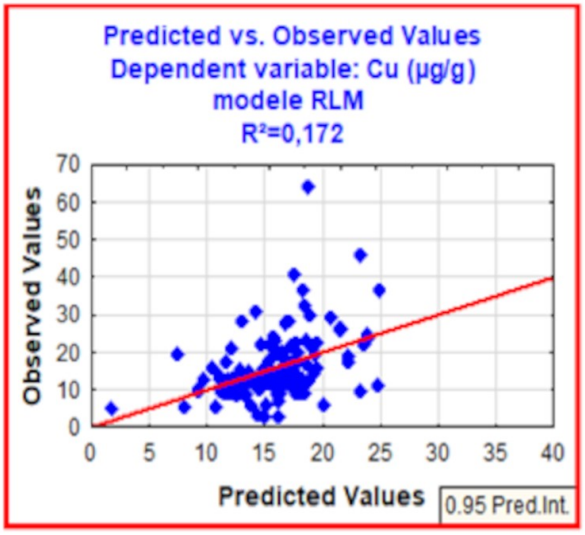


Fig.2. Architecture of the three-layer configuration neural network [16-8-1] used in this study.

Fig.3.a, and Fig.3.b shows the relationships between observed and estimated heavy metal contents for the models developed by the ANN and MLR methods. They show the performance of the ANN method, which is due to the high coefficient of determination of 0.99 observed for Manganese and the two other metals (Copper and Zinc). By comparing the models established by the MLR with those found by the ANN, we can see that the latter are the most efficient. This can easily be seen by referring to fig.3.





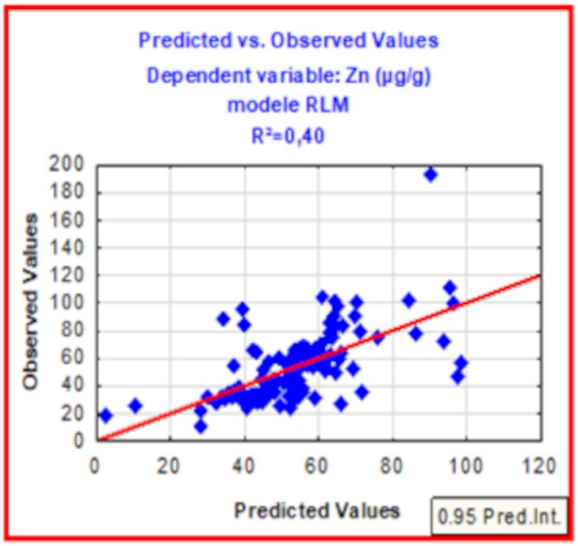
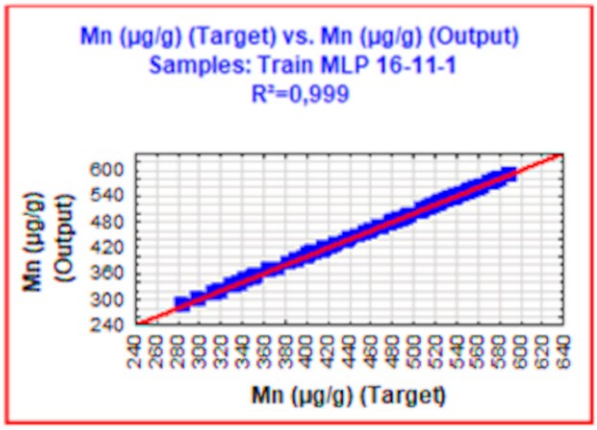
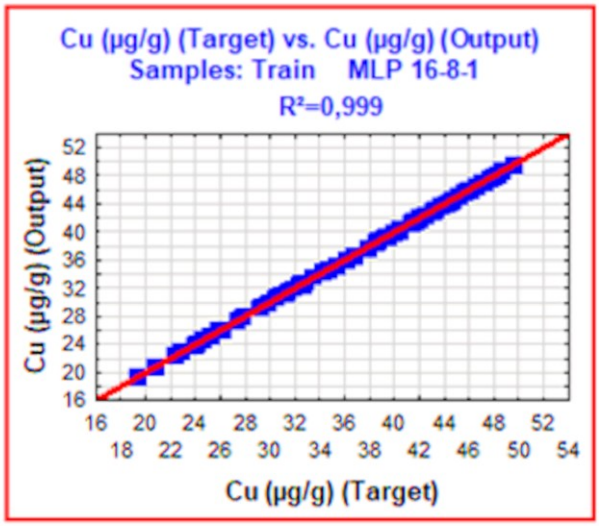


Fig.3.a Relationship between estimated Copper, Manganese and Zinc contents using multiple linear regression (MLR) models and observed values.





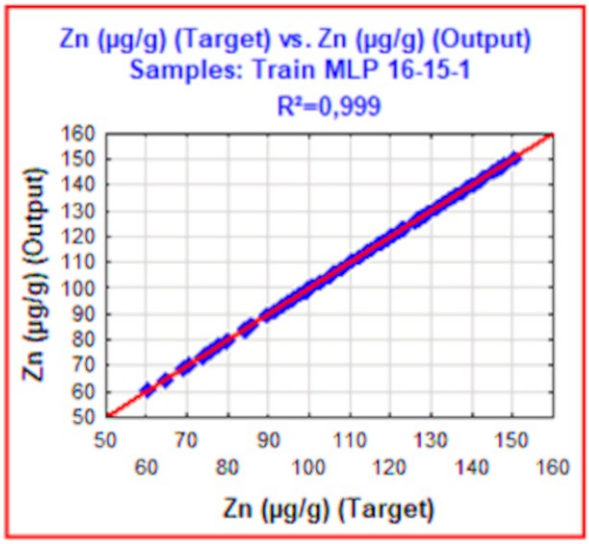


Fig.3.b Relationship between estimated Copper, Manganese and Zinc contents using artificial neural network (ANN) models and observed values.

4 Conclusion

The study of the prediction of water-soluble heavy metal concentrations from physico-chemical parameters revealed that the predictive models developed by the current method, which is founded on the ANN technique, are more efficient than those found by the MLR method. This performance appears to be caused by the fact that the concentrations of heavy metals in the waters of the Oued Inaouen catchment are related to the Physico-chemical characteristics of the water by non-linear connections. This leads us to consider, in the future, developing more aspects concerning this work. We propose to predict heavy metal contents by including other parameters in our database and to work with different types of neural networks such as the RBF (radial-based function) type to try to generalize these forecasting models to different Moroccan aquatic environments.

References

1. K. E. Nicolai, "A Green Gambit: *The Development of Environmental Foreign Policy in Morocco*," *J. North African Stud.*, pp. 1–27,(Dec. 2020), DOI: 10.1080/13629387.2020.1865931.
2. M. L. Sall, A. K. D. Diaw, D. Gningue-Sall, S. Efremova Aaron, and J.-J. Aaron, *"Toxic heavy metals: impact on the environment and human health, and treatment with conducting organic polymers, a review," Environ. Sci. Pollut. Res.*, vol. **27**, no. 24, pp. 29927–29942, (2020), DOI: 10.1007/s11356-020-09354-3.
3. S. B. Norton, D. J. Rodier, W. H. van der Schalie, W. P. Wood, M. W. Slimak, and J. H. Gentile, *"A framework for ecological risk assessment at the EPA," Environ. Toxicol. Chem.*, vol. **11**, no. 12, pp. 1663–1672, (Dec. 1992), DOI: https://doi.org/10.1002/etc.5620111202.
4. S. G. K. Patro and K. K. sahu, *"Normalization: A Preprocessing Stage," Iarjset*, pp. 20–22, (2015), doi: 10.17148/iarjset.2015.2305.
5. Z. Bayatzadeh Fard, F. Ghadimi, and H. Fattahi, *"Use of artificial intelligence techniques to predict distribution of heavy metals in groundwater of Lakan lead-zinc mine in Iran," J. Min. Environ.*, vol. **8**, no. 1, pp. 35–48, (2017), DOI: 10.22044/jme.2016.592.
6. A. Abdallaoui and H. El Badaoui, *"Comparative study of two stochastic models using the physicochemical characteristics of river sediment to predict the concentration of toxic metals," J. Mater. Environ. Sci.*, vol. **6**, no. 2, pp. 445–454, (2015).
7. A. E. H. H. Alayat, H. El Badaoui, A. Abdallaoui, D. Abrid, *"DEVELOPMENT OF MATHEMATICAL MODELS FOR PREDICTING THE IRON CONCENTRATIONS OF LAKE OUBEIRA WATERS (NE ALGERIAN)," J. Fundam. Appl. Sci.*, vol. **10**, no. 1, pp. 83–96, (2018), DOI: http://dx.doi.org/10.4314/jfas.v10i1.6.
8. H. E. B. A Abdallaoui, *"Prédiction des teneurs en métaux lourds des sédiments à partir de leurs caractéristiques physico-chimiques," J. Phys. Chem. News*, vol. **58**, pp. 90–97, (Apr. 2011).
9. H. El El Badaoui, A. Abdallaoui, and I. Manssouri, *"Elaboration de modèles mathématiques stochastiques pour la prédiction des teneurs en métaux lourds des eaux superficielles en utilisant les réseaux de neurones artificiels et la régression linéaire multiple," J. Hydrocarb. Mines Environ. Res.*, vol. **3**, no. 2, pp. 31–36, (2012).