Predicting Nitrate Levels in the Saïss Water Table: A Comparative Study of Machine Learning Methods

Hajar Jaddi[1,*], *Abdellah* El-Hmaidi[1], *Habiba* Ousmana[1], *Mohamed* Berrada[2], *My Hachem* Aouragh[1], *Zineb* Iallamen[1], *Zahra* Kasse[3], *Anas* El Ouali[4], *M'hamed* Boufala[5], *Hind* Ragragui[1], *Jihane* Saouita[1]

[1] Water Science and Environmental Engineering Team, Faculty of Sciences, Moulay Ismail University, Meknes, Morocco.
[2] Faculty of Medicine and Pharmacy, Abdelmalek Essaâdi University, Tangier, Morocco.
[3] Anassi High School (Annex 2), Ministry of National Education, Meknes, Morocco.
[4] Department of Geomorphology and Geomatics, Scientific Institute, Mohammed V University, Rabat
[5] ONEE, National Office for Electricity and Drinking Water, Water Branch, Meknes, Morocco.

Abstract. The main goal of this study is to predict nitrate (NO^{3-}) levels in the Saiss basin water table as a function of various physicochemical parameters. To accomplish this, three machine learning approaches were utilized: multiple linear regression (MLR), super vector regression (SVR), and artificial neural networks (ANN). The independent variables were composed of six water quality parameters, including Ca^{2+}, Na^{2+}, EC, Cl^{-}, HCO^{3-}, and $SO4^{2-}$. The study utilized a dataset of 389 water samples collected between 1991 and 2017. The artificial neural network (ANN) was trained using the Levenberg-Marquardt (LM) algorithm, which was selected from various optimization algorithms. Additionally, during the training of the SVR model, it was observed that the RBF kernel outperformed the other kernels (linear, polynomial, and sigmoid kernel). The results were analyzed by the coefficient of determination (R2) and the mean square error (MSE). The results of the MLR method revealed R2 (0.523) and MSE (757.34). The ANN model with architecture [6-20-1] performed better than RLM with R2 = 0.836, MSE= 0.023 The SVR model result confirms what has been proved by ANN concerning the performance, with R2=0.902 and MSE= 4,364.

1. Introduction

Groundwater resources in some regions of Morocco constitute the only drinking water resource for human populations. They are dependent on a set of natural and anthropic factors. They are subject to daily disturbances, which lead to a deterioration of their hygienic quality [1]. Over the last few decades, the increase in urban, agricultural, and Industrial activities have heightened the risk of contaminating both surface and groundwater.

The control of water, which necessarily involves knowledge of its quality, is a fundamental aspect of human development [2]. The Saïss Plain ranks among Morocco's primary agricultural regions. Its water resources are provided largely by the karstic aquifer of the Lias of the Middle Atlas carbonate causse. The water resources of this aquifer are used to satisfy domestic, agricultural, and industrial needs and constitute a determining factor for the economic and social development of this region.

Linked to the water richness of the Saïss aquifer, the intensification of crops, through poor management of irrigation and increased use of fertilizers such as nitrates and pesticides, has resulted in a significant deterioration of soil and groundwater quality.

Artificial neural networks (ANNs) stand out among prominent spatial modeling techniques for their ability to manage non-linear relationships, classify unfamiliar data by learning from known examples during training, and navigate uncertainties. These capabilities open up new avenues for data mining [3]. Today they play a crucial role across various scientific disciplines, particularly in automating solutions for environmental challenges.

Artificial neural networks (ANNs) produce compelling outcomes owing to their capacity for learning [4], parallel processing, and ability to handle inherent or nonlinear complexities within systems [5]. In recent years, ANNs have been utilized in the fields of hydrology, ecology, and environmental sciences.

Artificial neural networks have been used by several authors for modeling and forecasting environmental parameters, in the literature. Many studies have shown that the ANN model outperforms other regression and machine learning models in predicting water quality. [6] developed a study was conducted to predict organic

* Corresponding author: hajar.jaddi@edu.umi.ac.ma

carbon levels in the Terminal Quaternary deposits of the Alboran Sea [7] using these stochastic models for the prediction of a parameter (electrical conductivity) indicative of groundwater quality at fractured aquifers of the Ivory Coast. [7] established probabilistic mathematical models to predict concentrations of heavy metals in surface waters using artificial neural networks and multiple linear regression. [8] designed a neural model to predict nitrate and sulfate concentrations in lake waters of the Middle Atlas Mountains in Morocco.[9] predicted carbonate and biogenic and detrital element contents by artificial neural networks in Atlantic mud volcano deposits in northern Morocco. [10] predicted NO_2 and nitric oxide NO concentrations in Santiago, Chile. [11] Assessed how well Artificial Neural Networks (ANN) and Local Linear Regression (LLR) techniques performed in predicting monthly reservoir levels.; etc.

The performance of artificial neural networks (ANN) is compared with that of the support vector regression (SVR) model using four different types of kernels: linear, polynomial, radial basis function (RBF), and sigmoid.

The aim of this study is to develop statistical models utilizing multiple linear regression, artificial neural networks, and super vector regression, these models are intended for predicting nitrate (NO^{3-}) levels based on physicochemical parameters that serve as indicators of groundwater quality.

2. Study area

Located in Northern Morocco (Fig.1), the Saïss aquifer, which is characterized by a low topography varying between 250 m and 600 m, covers a total area of approximately 2100 km^2. Saïss basin is bounded on the North by the Prerifains wrinkles; on the East by the Wadi Sebou; on the west by the Wadi Beht, and on the south by the northern limit of the middle-Atlasic Causse. The study area comprises a Miocene basin underlaid by Liassic formations and overlaid by Plio-Quaternary formations [12]. It is an intra-montane continental basin filled with deposits of upper Miocene to Quaternary age overlying a Mesozoic basement [13].

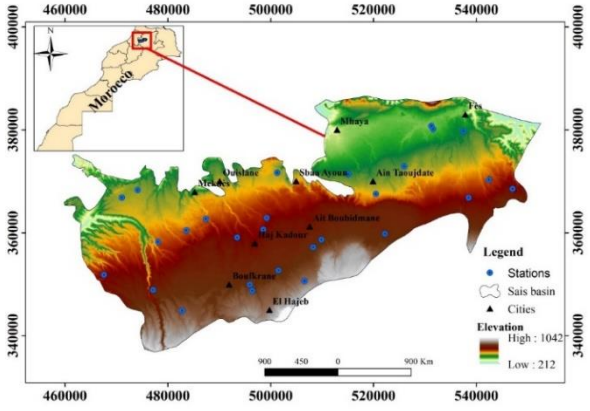


Fig. 1. Geographical location of the Saïss basin.

3. Materials and methods

3.1 Data used

The database used in this study consists of 389 water samples, collected by the Sebou hydraulic basin agency (SAHB) between 1991 and 2017. The independent variables are represented by six physic-chemical parameters: calcium (Ca^{2+}), sodium (Na^{+}), electrical conductivity (E.C), chlorides (Cl^{-}), bicarbonates (HCO_3^{-}), and sulfates (SO_4^{2-}). The dependent variable (to be predicted) is denoted by nitrates (NO_3^{-}).

3.2 Multiple Linear Regression MLR

It extends the simple regression model to incorporate multiple explanatory variables. It consists of finding a linear equation linking the variable to be modeled Y (variable to be explained or endogenous) to the input variables (explanatory or exogenous variables), X. Let p represent the number of explanatory variables, and N denotes the number of observations.

The linear equation we are looking for is of the form (equation 1):

$$Y= \beta 0+ \beta 1x1+ \beta 2x2+ \ldots\ldots \beta pxp \qquad \text{Equation 1}$$

The parameters β are referred to as partial regression coefficients or weights. They quantify the influence of each variable on the studied quantity. We note that the number of parameters to be determined for a linear regression-based model (RLM) is (p+1).

3.3 Artificial Neural Networks (ANN)

Inspired by the neural system of the human brain, where specific actions are linked to specific reasons, artificial neural networks (ANNs) employ this technique. It emulates this sophisticated neural networking system by capturing intricate relationships between inputs and outputs, mirroring the complex structure of biological neurons. This method consists of a set of elementary processors, the neurons, which are largely connected and can exchange information through the connections that link them. The connections are directional and each of them is associated with a real value called the weight of the connection. This representation reflects the biological inspiration behind the first wave of interest in formal neurons from 1940 to 1970 [14].

Artificial neural networks provide an alternative to traditional mathematical modeling methods. They are considered nonparametric and nonlinear statistical models that effectively address decision support, diagnosis, and prediction problems. etc [15, 16]. The idea is to present the ANN with input and output data, allowing it to learn the relationship between the two through a process called learning. The latter consists in minimizing the error between the model output and the

desired output by adjusting the model parameters (the weights). The learning process generally unfolds as follows in five steps (Fig. 2):

1- Weighting the inputs by parameters called weights (W).

2- total of the weighted inputs [17] (equation 2):

$$A_j = \sum_{i=1}^{n} w_{ij} I_i \qquad \text{Equation 2}$$

With wij is the synaptic weight and Ii the input values.

3- Activation of the neuron through the activation function f (Equation 3)

$$S_j = f\left(\sum_{i=1}^{n} w_{ij} I_i + b_j \right) \qquad \text{Equation 3}$$

Where bj is the bias of neuron j.

4- Calculation of the error between the measured output and the one calculated by the ANN [18] (equation 4):

$$E = \frac{1}{N} \sum_{i=1}^{n} (Y_{ri} - Y_{di})^2 \qquad \text{Equation 4}$$

Where Y (r_i): the values of the desired output and Y_(d_i) the values of the decided output.

5- Optimizing Weights to Minimize Error: Learning Algorithms.

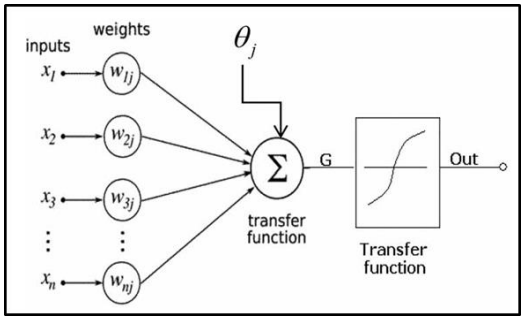


Fig. 2. Learning process and operation of the NNA

3.4 Data normalization

The database, in general, needs to be preprocessed to match the inputs and outputs. A common preprocessing is to perform an appropriate normalization, which considers the output values. This database is normalized between [-1;1] which aligns with the requirements of the hyperbolic tangent function used by neural networks. This normalization was performed according to the relation (equation 5):

$$Y_N = (y_{max} - y_{min}) \left(\frac{x_i - x_{min}}{x_{max} - x_{min}} \right) - y_{min} \qquad \text{Equation 5}$$

Y_N : Normalized values.

x_i : Non-normalized values.

x_{max} and x_{min} : Maximum and minimum of non-normalized values.

y_{max} and y_{min} : Maximum and minimum of the normalized values.

3.5 Evaluation of the Performance of Regression Models

To choose the best and most efficient regression model, several robustness metrics are commonly used. For this database, the statistical indicators used include the root mean square error (RMSE), mean absolute error (MAE), and correlation coefficient (R), represented by equations 6, 7, and 8, respectively

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (NS_{mes}(i) - NS_{est}(i))^2} \qquad \text{Equation 6}$$

$$MAE = \sum_{i=1}^{N} |NS_{mes}(i) - NS_{est}(i)| \qquad \text{Equation 7}$$

$$R = \sqrt{1 - \frac{\sum_{i=1}^{N} (NO3_{mes} - NO3_{est})}{\sum_{i=1}^{N} (NO3_{mes} - NO3_{est})}} \qquad \text{Equation 8}$$

NS_mes (i) et NS_est (i) represent the values of the target vector and output neuron prediction vector of our network. N represents the number of test samples studied and N_W represents the total weights and biases used for each architecture. The most well-known kernel functions are listed as follows:

Linear kernel: K(xi, xj)= xTi xj

Polynomial kernel: K(xi, xj) = (xTi xj +γ)d, γ > 0

RBF kernel: K(xi, xj) = exp (-γ ||xi -xj||)2, γ > 0

Sigmoid kernel: k(xi, xj) = tanh (γxTi xj +r), γ > 0

Variables xi and xj serve as inputs, and γ is the regularization parameter. The Lagrange multipliers are denoted as αi ¼ αi _ αi. The accuracy of prediction is based on the selection of three parameters, that is, γ, ε, and C, whose values are determined using a firefly algorithm.

4. Results and discussion

4.1 Developing the MLR Model Framework

Multiple linear regression analysis, for the total base containing 389 samples, consists of dealing with the relationships between the quantitative dependent variable to be explained (Yi: nitrate levels (NO^{3-}) and

several independent quantitative explanatory variables (X: Ca, Na, EC, Cl, HCO_3, and SO_4). It also attempts to determine the equations of a non-linear polynomial fit for the analysis of the relationships between two quantitative variables, to interpret the correlation coefficient, and to verify the role of each variable (Fig.3).

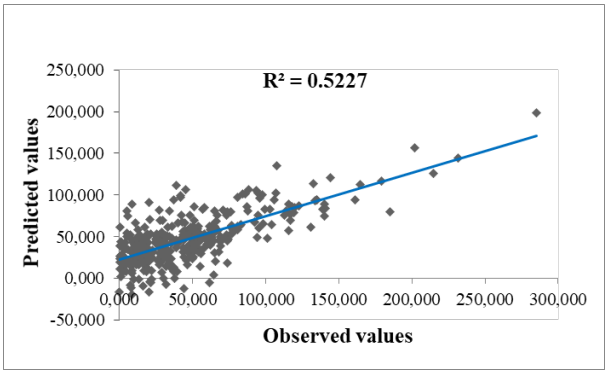


Fig. 3. Relationship between observed nitrate levels and those estimated by multiple linear regression.

4.2 Development of the ANN model

Different types of learning algorithms chosen to form our artificial neural network (ANN) are 12 in number, performed by the MATLAB software, and are shown in figures 4 and 5. The results obtained indicate that after the execution of these various algorithms, the most performing is the Levenberg-Marquardt (LM) with minimal levels of MSE and high values of coefficient of determination (R^2). The algorithms include the quasi-newton backpropagation (BFGS), fletcher-Reeves conjugate gradient backpropagation (CGF), one-step secant backpropagation (OSS), Powell-Beate conjugate gradient backpropagation (CGB) and variable learning rate backpropagation (GDX) represent suitable performances for the prediction of groundwater nitrate content of Saïss. The algorithms namely, bayesian regularization train lm (BR), Resilient backpropagation (RP), Scaled conjugate gradient backpropagation (SCG), Batch gradient descent with momentum (GDM), polak-ribiere conjugate gradient backpropagation (CGP) and Batch gradient descent (GD), remain insignificant with low coefficients of determination R2 and high root mean square errors MSE (Fig. 4 and 5). Thus, the results obtained by these algorithms show inconceivable and less reliable performances compared to the others.

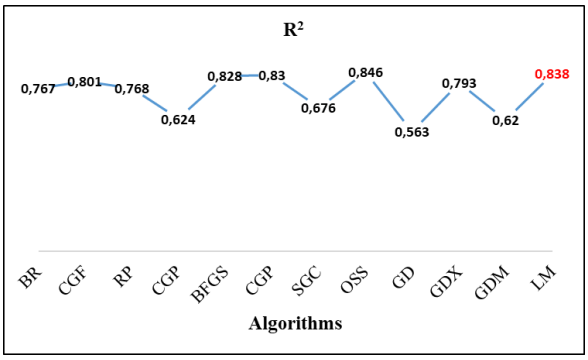


Fig. 4. Comparing Algorithm Performance on ANN Model via R^2.

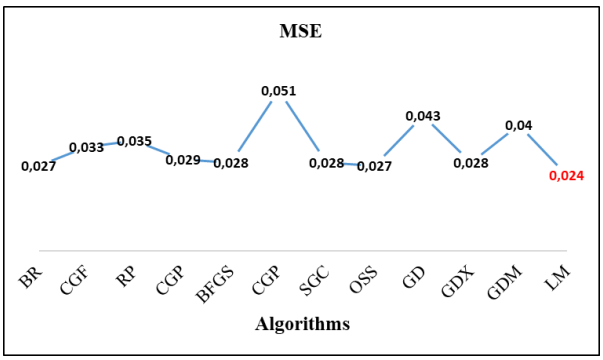


Fig. 5. Comparing Algorithm Performance on ANN Model via MSE.

The Levenberg-Marquardt (LM) backpropagation algorithm was employed to efficiently and robustly train the artificial neural network, focusing on single hidden layer networks due to their capability to predict all outputs.

To enhance the model's performance, several preliminary tests on the whole database of 389 samples have been performed. Thus, The neural network architecture is adjusted iteratively, primarily focusing on modifying the number of hidden layers, hidden neurons, and/or the number of training cycles or iterations.

To show the predictive quality of ANN, the data used during this study are randomly divided into three groups. The first group, consisting of 70% of the total data, will be used to train the system. The second group, containing 15% of the total data, will be called to validate the network and the remaining 15% that did not participate in training the model will be used as an independent test to approve the generalization of the model.

Evolution of the mean square error (MSE) for nitrate levels as a function of the number of neurons in the hidden layer. It shows that after the fifth iteration, the desired result is achieved (Fig. 6). Simulations were conducted to ascertain the optimal number of neurons required for predicting nitrate levels. They demonstrated that the error stabilizes despite increasing the number of neurons. This stability is evident in the network, particularly with 20 neurons in the hidden layer (Fig. 7).

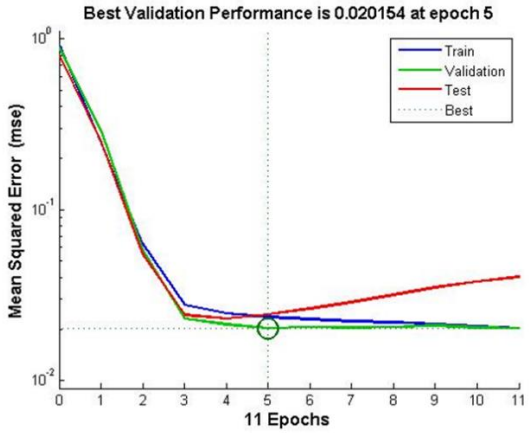


Fig. 6. Evolution of the MSE for nitrate levels with 20 neurons in the hidden layer.

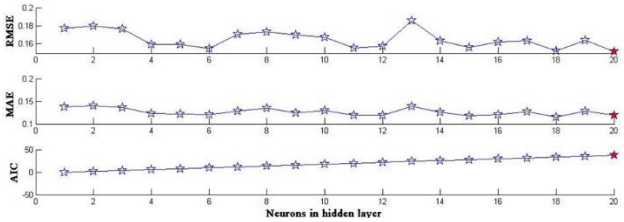


Fig. 7. Performance Metrics as a Function of Hidden Layer Neurons: RMSE, MAE, and AIC.

The architecture of the most suitable neural network model for predicting nitrate levels is [6-20 -1] (Figure 8). This model is formed by 6 neurons for the input layer (Ca^{2+}, Na^{+}, EC, Cl^{-}, HCO_3^{-} and SO_4^{2-}), there are 20 neurons in the hidden layer, and only one neuron exists in the output layer specifically for NO_3^{-}.

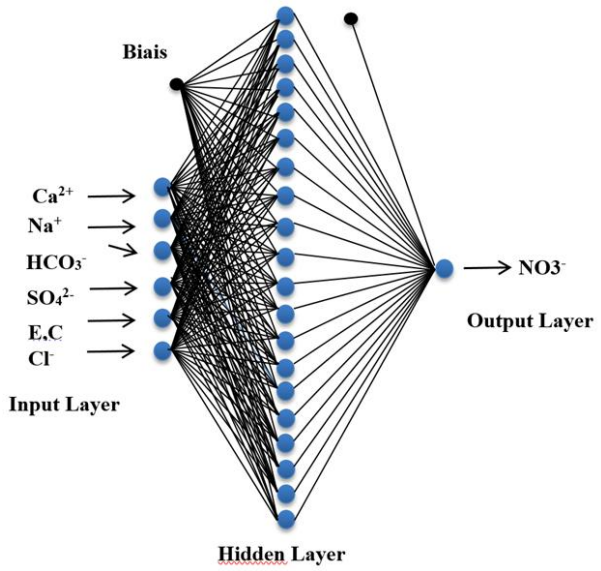
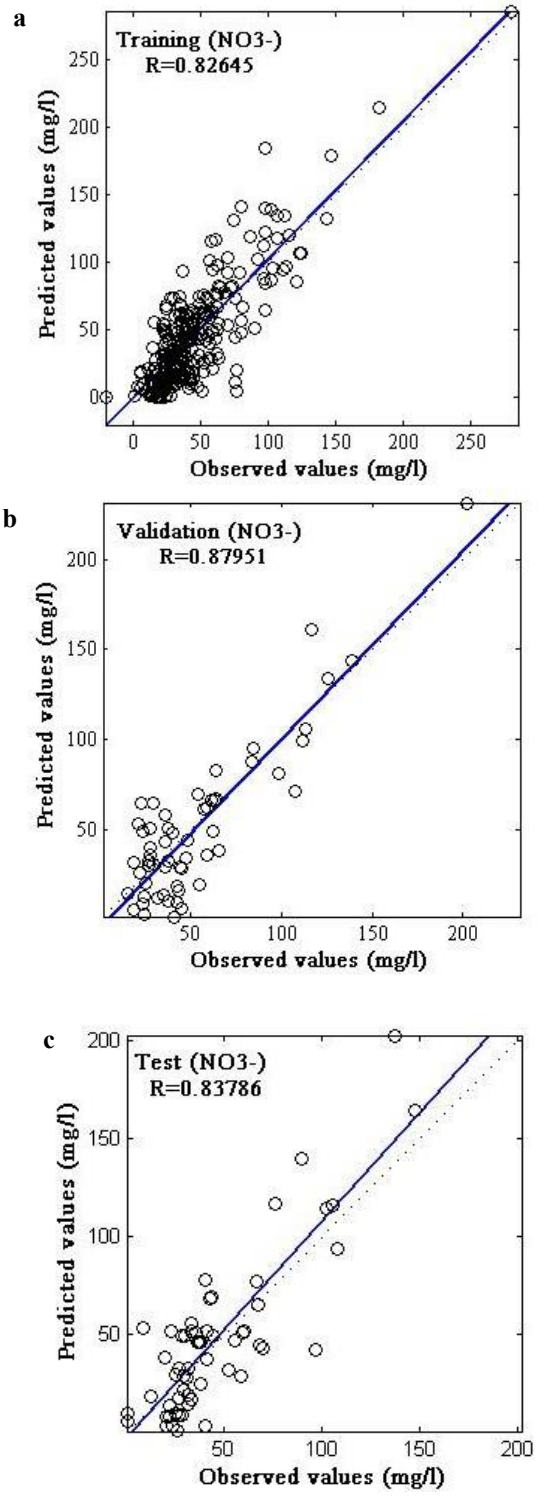


Fig. 8. Neural network configuration Architecture [6-20-1].

4.3 Learning and validation

The modeling conducted by the neural networks demonstrated the predictive quality of the models. A total of 398 samples were used in this study. Of these, 70% (x samples) formed the dataset for developing a predictive model for the dependent variable, while the remaining 30% (y samples) were set aside to test the validity and predictive performance of these models.

This model predicts the nitrate levels, once the architecture [6-20 -1] of the ANN model is chosen. The results of this prediction (approximated values) are compared to the values measured by laboratory analysis (observed values). This comparison indicates a good agreement between these values for the learning, validation, and testing phase (Fig. 9 and 10).



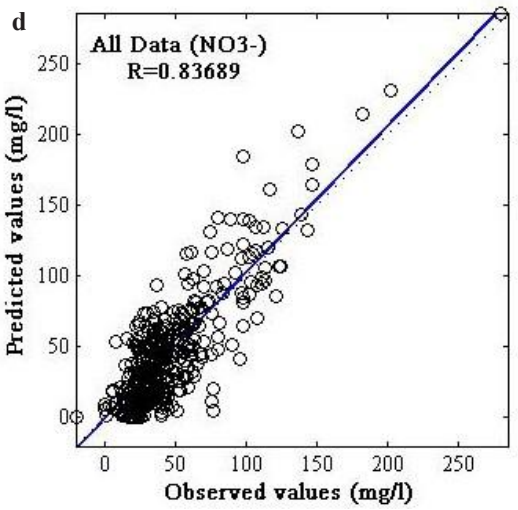


Fig. 9. Observed and predicted nitrates levels using MLP-ANN model: a) Training Phase, b) Validation Phase, c) Testing Phase, and d) Complete Dataset

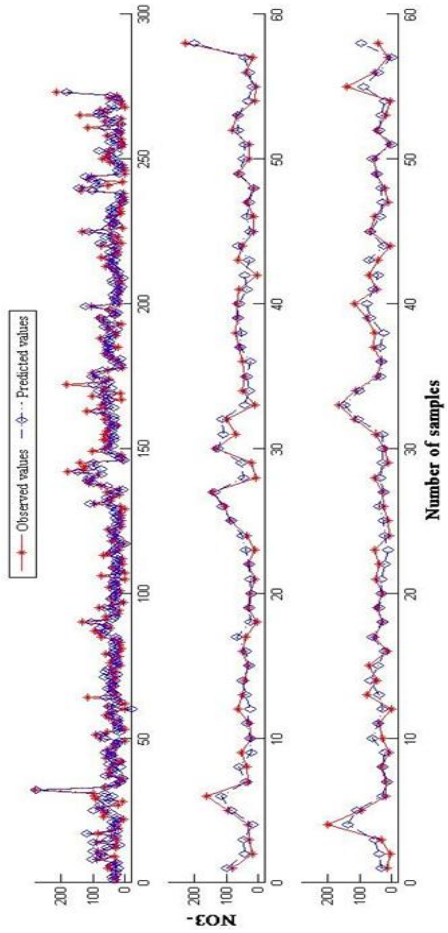


Fig. 10. Nitrate comparison: measured vs. predicted in the three phases of the MLP -ANN model [6-20 -1].

The box plots of the nitrate levels for the observed and estimated or predicted values during the three phases: learning, validation, and testing (Fig.11).

The nitrate content values in the groundwater of Saïss, both observed and predicted, are almost similar across the three phases. These results confirm what was obtained precisely during the comparison of nitrate contents by the MLP- ANN model.

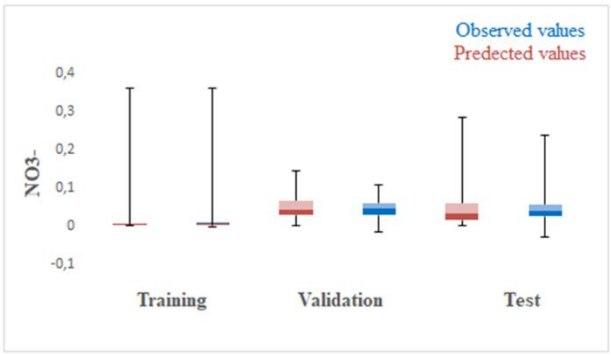


Fig. 11. Boxplots Comparing Measured and Predicted Nitrate Levels Across Training, Validation, and Testing Phases.

4.4 Development of SVR model

Figures 12 and 13 show the proposed SVR model is tested for the four types of Kernels, namely linear kernel, polynomial kernel, radial basis function (RBF) kernel, and sigmoid kernel.

The results indicate that the Radial Basis Function (RBF) kernel outperforms others, demonstrating a high determination coefficient (R) and low mean square error (MSE) in predicting nitrate levels.

These results align with findings from previous studies that evaluated the performance of NNA learning algorithms and SVR kernels in predictive modeling. For instance [19] in their study on « performance assessment of artificial neural networks and support vector regression models for stream flow predictions ». Tarbela Dam, located on the Indus River in Pakistan, was analyzed using artificial neural networks (ANNs) and four variants of support vector regression (SVR) models, which included linear, polynomial, radial basis function, and sigmoid kernels. The study highlighted that the SVR model utilizing the radial basis function kernel outperformed other models, including ANNs, in accurately predicting stream flows.

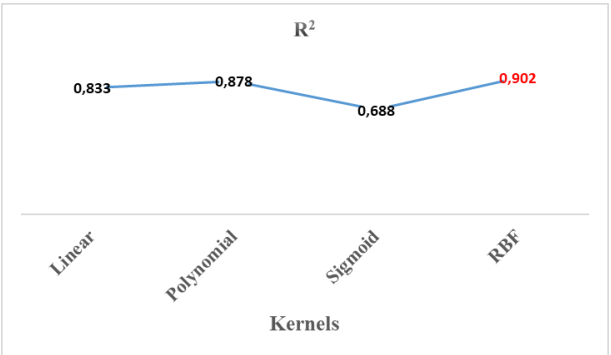


Fig. 12. Performance evaluation of kernels on SVR model by R^2.

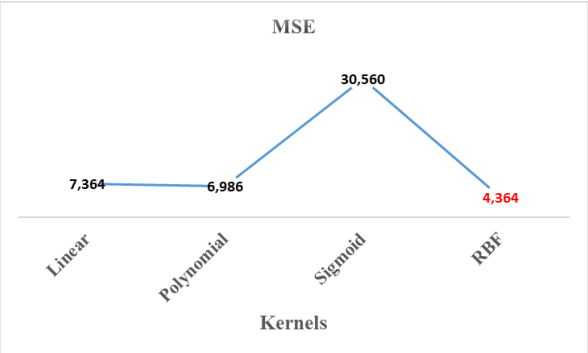


Fig. 13. Performance evaluation of kernels on SVR model by MSE.

The modeling carried out by the neural networks affirmed the predictive reliability of the models. For prediction, the nitrate levels are used during the learning and testing phases. The superposition is very good between the measured and predicted values using the SVR model in the train, overall, and, test phases (Fig. 14).

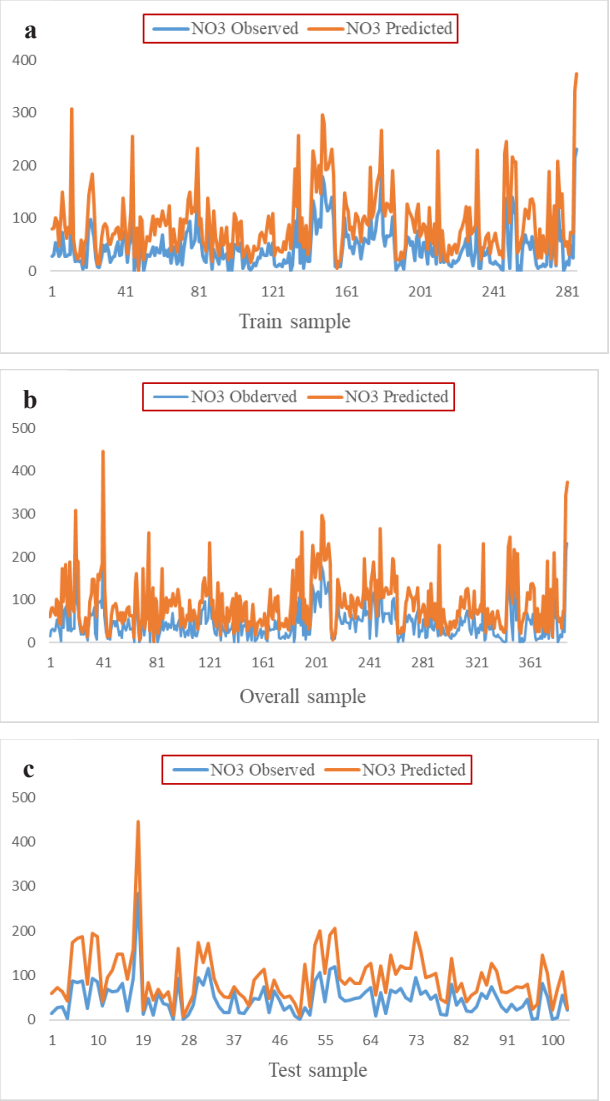


Fig. 14 Comparison of nitrate levels observed versus those predicted using the SVR model in the train (a), overall (b), and test phases (c).

The first part of 286 samples (75% of 389 samples) is used for training and building the model. The remaining portion of 103 samples (25% of a total of 389) is used for testing.

5. Performances comparison of the RLM with the RNA and the SVR methods

Comparing coefficients of determination and statistical metrics among three prediction models: multiple linear regression (MLR), artificial neural networks (ANN), and support vector regression (SVR). are illustrated in Table 1. The coefficient of determination R2=0.836 and R2=0.902 calculated by the multilayer perceptron neural networks and the super vector regression, when predicting nitrate levels is much closer to 1 than that calculated by the multiple linear regression with R2=0.523. Similarly, the mean square error MSE=0.023 and MSE=4.364, established by the ANN and SVR models, is very small compared to that calculated by the MLR with MSE =757.345.

These results show that the model established by neural networks and super vector regression clearly performs well compared to that established by the MLR method. Also, the parameters are non-linear due to the fact that the coefficients are very high in the case of the analysis with the ANNs and SVR and lower in the case of the MLR (Table 1).

Tab.1 Nitrate levels obtained by MLR, ANN, and SVR, along with their coefficients of determination and statistical indicators.

Methods	Nitrates (NO^{3-})		
	R^2	MSE	RMSE
MLR	0.523	757.345	27.520
ANN	0.836	0.023	0.152
SVR	0.902	4.364	2.089

6. Conclusion

The Saïss aquifer is located in one of Morocco's major agricultural regions. Its water resources are used to satisfy domestic, agricultural, industrial, and tourist needs. They are a determining factor for the economic and social development of this region.

Neural networks are highly potent mathematical and computational models adept at handling non-linear data. They hold significant importance across various scientific domains, particularly in automating solutions to environmental challenges.

The physics-chemical dataset was successfully modeled to predict nitrates in the groundwater of the Saïss basin.

A performance comparison of the methods MLR, ANN, and SVR was conducted.

The results obtained by the multiple linear regression method remain insignificant with a coefficient of determination of R^2= 0.523 and a very high mean square error MSE= 757.345. However, the prediction established by artificial neural networks and super vector regression shows very significant results with R^2= 0.836 and MSE= 0.023 with a better configuration [6-20-1], R^2= 0.902, and a mean square error MSE= 4.364 successively.

Modeling with artificial neural networks shows a strong correlation between observed and predicted values. This shows that the studied parameters in the data set are related to the modeled parameters by a nonlinear relationship and that the prediction of nitrates by MLP neural network with supervised learning with The "Levenberg-Marquardt" algorithm delivers superior results in terms of speed, convergence, and overall performance. Support vector regression is a powerful method for regression of the prediction of the nitrates levels by super vector regression with RBF kernel demonstrating the superior performance of SVR With a high coefficient of determination (R^2) and a low mean square error (MSE). This result confirms the findings of artificial neural networks (ANN) regarding the correlation between observed and predicted values.

References

1. Direction de la Recherche et de la Planification de l'Eau (DRPE), Gestion des ressources en eau au Maroc, p. 23 (2004).
2. S.G., Eblin, A.P. Sombo A.P. G. Soro, N. Aka, O. Kambiré, & N. Soro, J. Ap. Biosc., **75**, 6259-6271 (2014).
3. V. Nykänen, Nat. Resour. Res. **17** (1), 29-48, (2008).
4. He L. M. L., & He Z. L., USA. Water Research, **42**(10), 2563-2573 (2008).
5. W.C. Liu, & W.B. Chen, Computers & Geosciences, **45**, 13-25 (2012).
6. A. EL Hmaidi, H. El Badaoui, A. Abdallaoui, & B. EL Moumni, E.J. S. R. **107**, 400-413 (2013).
7. M.A. Kouassi, J.W.S. E.T. **1**(2) (2016).
8. H. El Badaoui, A. Abdallaoui, I. Manssouri, & L. Lancelot, J. H.M.E.R. **G** (2), 31-36 (2012).
9. A. Lalilti, B. El Moumni, A. El Arrim, H. El Malki, N. Alaoui Mhammedi, H. Ousmana, M. Berrada & A. El Hmaidi, B.I.S. Rabat, sect. Sc. Ter., **39**, 121-134 (2017).
10. P. Perez, & A. Trier, At. Env. **35**(10), 1783-1789 (2001).
11. A. Samani, J. Zubovits, & D. Pelewes, Phy. Med. & Bio. **52**, 6, p. 1565 (2007).
12. A. Essahlaoui, H. Sahbi, L. Bahi, & N. El-Yamine, J.A.E.S. **32**(4), 777-789 (2001).
13. D. Fassi, N.M.S.G. **389**. Rabat, Morocco, 527p (1999).
14. W.S. McCulloch, & W. Pitts, B. M. B. **5**, 115-133 (1943).
15. S. Huo, Z. He, J. Su, B. Xi, & C. Zhu, P. E.S. **18**, 310-316 (2013). https://doi.org/10.1016/j.proenv.2013.04.040.
16. N.D. Kaushika, R.K. Tomar, & S.C. Kaushik, Sol. En., **103**, 327-342 (2014). https://doi.org/10.1016/j.solener.2014.02.015.
17. E. El Tabach, L. Lancelot, I. Shahrour, & Y. Najjar, Math. Com. Mod., **45**(7-8), 766-776 (2007). https://doi.org/10.1016/j.mcm.2006.07.020.
18. M. Bouamar, & M. Ladjal, In 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, LASS, Laboratoire d'Analyse des Signaux et Systèmes, Université de M'sila, Algeria (25-29) (2007).
19. A.R. Ghumman, S. Ahmad, H.N. Hashmi, Env. Mon. Ass. 190. No. **12**, 1–20 p (2018). Doi 10.1007/s10661- 018-7012-9.