European Scientific Journal June 2015 edition vol.11, No.18 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431

UTILISATION D'UN MODELE HYBRIDE BASE SUR LA RLMS ET LES RNA-PMC POUR LA PREDICTION DES PARAMETRES INDICATEURS DE LA QUALITE DES EAUX SOUTERRAINES CAS DE LA NAPPE DE SOUSS-MASSA- MAROC

Manssouri T.
Sahbi H.
Laboratoire de Géo-Ingénierie et Environnement, Faculté des Sciences, Université Moulay Ismail, Zitoune, Meknès- Maroc.
Manssouri I.
Laboratoire de Mécanique, Mécatronique et Commande, École Nationale Supérieure d'Arts et Métiers, Université Moulay Ismail, Meknès– Maroc.
Boudad B.
Laboratoire de Géo-Ingénierie et Environnement, Faculté des Sciences, Université Moulay Ismail, Zitoune, Meknès- Maroc.

Abstract

This work describes a new approach to the prediction of the parameters (microbiological, physical-chemical) groundwater quality indicators in the water table of Souss-Massa Morocco. The originality of this work lies in the application of a hybrid model based on the Stepwise Multiple Linear Regression and Neural Networks Multilayer Perceptron type. During the first stage, conventional statistical models namely the Stepwise Multiple Linear Regression was applied to a database that consists of eleven vectors as input vectors of the model and three vectors as the model output vectors in order to optimize the explanatory variables. In a second step, the optimized data base in the first step was used to construct a non recurring multi-layer network, the weights of the network connections are determined using the gradient back propagation algorithm. The data used as a database (learning, testing and validation) of the hybrid model are those relating to the analysis of 52 groundwater samples collected at several stations distributed in space and in time, of the groundwater Souss-Massa Morocco. The dependent variables (to explain or predict), which are three in number, are the Electrical Conductivity EC, the amount of Fecal Coliforms CF and Organic Matter MO.

Keywords: Prediction, hybrid model, optimisation, Neural Network MLP Type, Stepwaise Multiple Linear Regression, quality indicators of groundwater, groundwater SOUSS-Massa-Morocco

Résumé:

Le présent travail, décrit une nouvelle approche à la prédiction des paramètres (microbiologique et physico-chimique) indicateurs de qualité des eaux souterraines dans la Nappe de Souss-Massa- Maroc.

L'originalité de ce travail réside dans l'application d'un modèle hybride basé sur la Régression Linéaire Multiple Stepwise et les Réseaux de Neurones de type Perceptron Multicouche. Durant la première étape, les modèles statistiques classiques à savoir la Régression Linéaire Multiple Stepwise a été appliquée à une base de données qui se compose de onze vecteurs comme des vecteurs d'entrées du modèle et trois vecteurs comme des vecteurs de sorties du modèle dans le but d'optimiser les variables explicatives.

Dans une seconde étape, la base de données optimisée durant la première étape a été utilisée pour construire un réseau multicouche non récurrent, les poids des connexions du réseau sont déterminés à l'aide de l'algorithme de rétro-propagation de gradient.

Les données servant comme base de données (Apprentissage, test et validation) du modèle hybride sont celles relatives à l'analyse de 52 échantillons des eaux souterraines prélevés au niveau de plusieurs stations, réparties dans l'espace et dans le temps, de la Nappe de Souss-Massa-Maroc. Les variables dépendantes (à expliquer ou à prédire), qui sont en nombre de trois, sont la Conductivité Electrique CE, la quantité de Coliformes Fécaux CF et la Matière Organique MO.

Mots clés: Prédiction, modèle hybride, optimisation, Réseau de Neurones de type PMC, Régression Linéaire Multiple Stepwise, indicateurs de qualité des eaux souterraines, Nappe Souss- Massa-Maroc

Introduction

L'eau est un élément indispensable à la vie au revêt de l'importance pour dénombrables activités humaines. L'eau déjà rare, est aussi soumise à l'augmentation continue des besoins, due à l'évolution rapide de la population, à l'amélioration du niveau de vie, au développement industriel et à l'extension de l'agriculture irriguée. Ces pressions sur les ressources en eau s'accompagnent d'une dégradation croissante et de plus en plus grave de leur qualité.

Il est donc essentiel de quantifier et d'analyser la quantité et la qualité des eaux souterraines et de trouver les moyens de gérer ces ressources pour en assurer la durabilité.

La qualité des eaux souterraines de la zone d'étude du bassin hydraulique de Souss massa (BHSM) a subit ces derniers années une certaine détérioration, à cause de l'utilisation intensive d'engrais chimiques et de fertilisants dans l'agriculture ainsi que d'une exploitation désordonnée. Ces élément modifient le chimisme de l'eau et la rendent impropre aux usagers souhaités.

Pour apprécier la qualité des eaux souterraines, la connaissance d'un certains nombre d'indicateurs tels que la conductivité électrique CE, la quantité de Coliformes Fécaux CF et la Matière Organique MO, est primordiale.

Dans la littérature, les réseaux de neurones ont trouvé un grand succès dans la classification, la modélisation et la prédiction, nous citons à titre d'exemple:

(Manssouri, 2013) Donnent deux méthodes de modélisation utilisées pour la prédiction des paramètres météorologique en général et le taux d'humidité en particulier. Dans un premier temps, des méthodes sont basées sur l'étude des réseaux de neurones artificiels de types PMC (Perceptron multi-couches) sont appliquées pour la prédiction du taux d'humidité de la région de Chefchaouen au Maroc. Dans un deuxième temps, la nouvelle architecture de réseaux de neurones de types PMC proposé a été comparée à celle du modèle de la régression linéaire multiple (RLM). Les modèles prédictifs établis par la méthode des réseaux de neurones PMC sont plus performants par rapport à ceux établis par la régression linéaire multiple, du fait que la bonne corrélation a été obtenue par les paramètres issus d'une approche neuronale avec une erreur quadratique moyenne de 5 %.

(Cheggaga, 2010) montrent la possibilité de l'utilisation des réseaux de neurones à couches non-récurrentes pour l'extrapolation, la prédiction et l'interpolation de la vitesse de vent dans le temps et dans l'espace en 3D (rayon r, hauteur h, temps t), à base de réseaux de neurones pour un apprentissage de quelques jours.

(Najjar, 2000) proposent une approche basée sur les réseaux de neurones artificiels pour l'évaluation de contamination d'un sol non saturé par déversement accidentel de polluant. L'étude vise l'analyse des risques de contamination des ressources en eau par déversement de polluant suite à un accident routier. Une base des données a été construite en utilisant une modélisation par les éléments finis, cette base de donnée est utilisée pour calibrer le modèle du réseau de neurones qui a servi à établir le risque de pollution de la nappe dans une zone concernée par un projet routier. L'étude réalisée montre que le modèle de réseau de neurones artificiels constitue un outil fort intéressant pour la modélisation du transfert de polluants dans un sol non saturé et pour l'étude d'impact de la construction des routes sur le sol et les ressources en eau.

(Ryad, 2002) ont travaillé sur l'application des réseaux de neurones de type RBF (Radial Basis Function) pour le problème de prédiction d'un système non linéaire. L'intérêt de cet article réside dans deux aspects : un apport au niveau de la topologie du réseau récurrent pour prendre en compte l'aspect dynamique des données. Le deuxième apport concerne l'amélioration de l'algorithme d'apprentissage.

(Manssouri, 2008) utilisent les réseaux de neurones à fonction de base radiale (RBF) pour la séparation de deux modes « normal et anormal » dans un système de distillation continue de méthylcyclohexane à partir d'un mélange binaire toluène\ méthylcyclohexane.

(D'Orazio, 2008) abordent le problème d'inspection et de détection automatique des anomalies dans les matières composites combinant les techniques ultrasoniques et les techniques de réseaux de neurones ; pour ce faire, D'Orazio et al. considèrent deux étapes pour aborder les problèmes de détections automatiques: la première étape, consiste en un prétraitement des différents signaux tirés de différentes structures composites étudiées ; ces dernières servent comme base d'apprentissage du réseau neuronal. La deuxième étape est l'utilisation du classifieur neuronal pour la détection des anomalies existantes dans les matières composites.

(Nagendra, 2006) utilisent le modèle de réseaux de neurones stochastique déterministe qui s'appelle le ANN-VEE (Véhicule Exhaust Emission) pour modéliser le phénomène de dispersion de la série temporelle journalière NO_2 à partir de l'émission des véhicules dans deux stations différentes à Delhi dans les routes urbaines. Les modèles de NO_2 basés sur les réseaux de neurons artificiels ont été formulés en utilisant trois choix d'entrées : le premier choix consiste à choisir comme entrées les variables trafiques et météorologiques ; le deuxième choix consiste à choisir seulement les variables météorologiques, et le dernier ne considère que les variables trafiques. Les résultats montrent que le modèle formé par les variables météorologiques et trafiques est plus performant dans les deux stations concernées.

L'objectif principal de cet article, est d'appliquer un modèle mathématique stochastique hybride basé sur les Réseaux de neurones Artificiels et la Régression Linéaire Multiple Stepwise comme outil de prédiction des paramètres (microbiologique et physico- chimique) indicateurs de la qualité des eaux souterraines de la Nappe Souss-Massa.

Description de la zone d'étude de la nappe Souss Massa-Maroc :

Le bassin hydraulique de Souss-Massa est localisé dans la partie occidentale du grand sillon sud africain. Il couvre une superficie totale de l'ordre de 27 880 Km² et comprend quatre bassins : le bassin de Souss (subdivisé en trois sous bassins : Souss amont, médian et aval), le bassin de

Massa, le bassin de Tamri-Tamraght et le Bassin de Tiznite. Les principales plaines de la région sont la plaine du Souss (4 500 km²), la plaine des Chtouka (1 260 km²) et la plaine de Tiznit (1 200 km²).

Les nappes du Souss, Chtouka Tiznit et Sidi Ifni sont les plus importantes dans la zone d'action de l'ABHSM voir figure1.

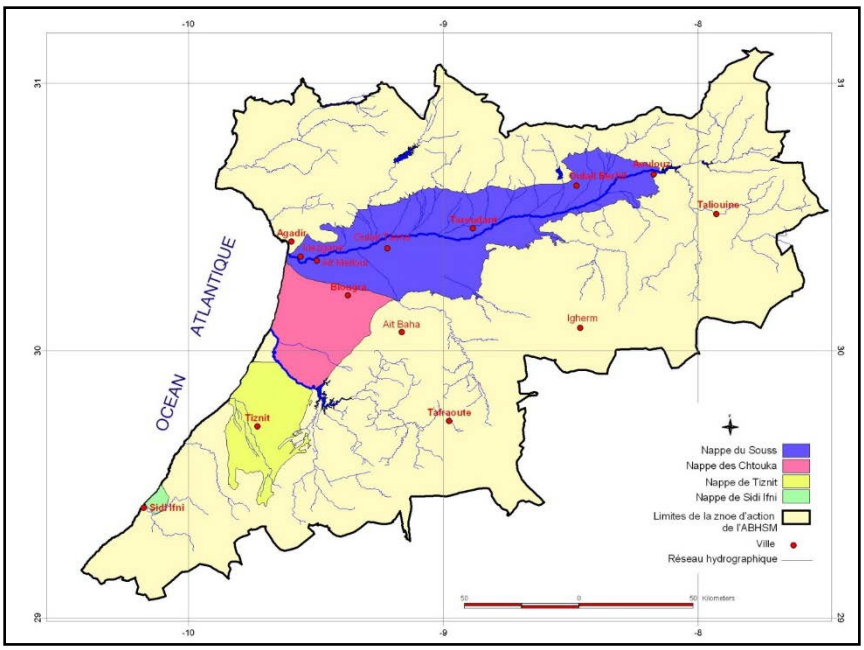

Figure 1. Principales nappes de la zone d'action de l'ABHSM Fournies par l'agence BHSM (2004)

Le climat de la région est à prédominance aride, avec un fort ensoleillement (3 089 heures/an).

La pluviométrie est très variable dans l'espace et dans le temps. Le nombre total de jours avoisine 30 jours par an en moyenne, le Haut Atlas étant plus arrosé avec un nombre de jours de pluie de l'ordre de 60 jours.

Résultats et discussion

Base de données

Les données utilisées dans le cadre de cette étude sont relatives à l'analyse chimique des paramètres indicateurs de qualité des eaux souterraines de souss massa effectuées à partir de 52 points de mesures (Cherkaoui, 2006) reparties sur l'ensemble des nappes souterraines gitant sous ce bassin.

Les variables dépendantes sont les paramètres indicateurs de la qualité des eaux souterraines déterminées dans ces points d'eau telles que la Conductivité Electrique CE, la Matière Organique MO et la quantité de

Coliformes Fécaux CF et les variables indépendantes qui sont le sodium Na^+,le potassium K^+,le calcium Ca^{2+},ion de magnésium Mg^{2+},ion d'hydrogénocarbonate HCO_3^-,ion de sulfate SO_4^{2-},Température de l'air Ta(°C), Température de l'eau Te (°C), Potentiel hydrogéne Ph, Streptocoques Fécaux SF,Coliforme Totaux CT

D'une manière générale, les bases de données doivent subir un prétraitement afin d'être adaptées aux entrées et sorties des modèles mathématiques stochastiques. Un prétraitement courant consiste à effectuer une normalisation appropriée (Basheer, 2000) qui tient compte de l'amplitude des valeurs acceptées par les modèles.

La normalisation de chaque entrée xi est donnée par la formule:

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

On obtient une base de données normalisée entre 0.1 et 0.9. Cette base de données est constituée de différentes variations des variables indépendantes (entrées du modèle) et dépendantes (sortie du modèle).

Régression Linéaire Multiple Stepwise (RLMS)

L'objectif principale de la Régression Linéaire Multiple pas à pas (Stepwise) est l'optimisation des variables indépendantes c'est-à-dire chercher un ensemble de variable qui donne une reconstitution satisfaisante pour les variables à expliquer.

Les objectifs d'une telle démarche sont multiples : économiser le nombre de prédicateurs, obtenir les formules assez simples, stables et d'un bon pouvoir prédictif en éliminant les variables redondantes qui augmentent le facteur d'inflation de la variance c'est-à-dire qui apportent peu d'information dans le modèle linéaire (Saporta, 1990).

En réalisant l'analyse par la Régression Linéaire Multiple pas à pas, on obtient les équations 1, 2 et 3 relatives respectivement à la conductivité électrique (CE), à la concentration de la Matière Organique (MO) et à la concentration des coliformes fécaux (CF).

CE = -0,0699 + 0,2946* [Na^+] + 0,1214* [$SO4^{2-}$] + -0,1646 *Ta + 0,5830 *CT (1)

R^2= 0,813 p< 0,0001

[MO] = 9,74654485232047E-02 + 0,732641083898005*[Na^+] (2)

R^2= 0,433 p< 0,0001

[CF] = 0,5167 + 0,4101*CT (3)

R^2=0,363 P= 0,003

Le modèle relatif à la conductivité électrique CE (1) est hautement significatif toutefois du fait que sa probabilité est inférieure à 0,001, les modèles (2) et (3) relatifs à la concentration des coliformes fécaux et à la concentration de la Matière Organique sont moins significatifs.

Modèle hybride basé sur les Réseaux de neurones Artificiels et la Régression Linéaire Multiple Stepwise.

Notre choix s'est porté sur un réseau multicouche non récurrent, en se basant sur l'algorithme d'apprentissage de rétro propagation. Le but de cet algorithme d'apprentissage est de minimiser RMSE (Root Mean Square Error). Le réseau de neurones artificiels se compose d'une couche d'entrée (input layer), d'une couche cachée (hidden layer) et d'une couche de sortie (output layer). Les variables d'entrées X= ([Na+], [SO42-], Ta, CT) (voir figure 2) qui sont les variables indépendantes, sont optimisées à l'aide de la Régression Linéaire Multiple Stepwise et normalisées entre 0.1et 0.9 puis présentées à la couche d'entrée du réseau qui contient quatre neurones. Ils sont d'abord multipliés par les poids IW et ensuite rajoutés au biais IB existant entre la couche d'entrée et la couche cachée. Les neurones de la couche cachée reçoivent les signaux pondérés. Après addition, ils les transforment en employant une fonction sigmoïde non linéaire. Le modèle mathématique est présenté à l'entrée de la couche de sortie. Ce modèle sera multiplié par les poids LW puis rajouté au biais LB existant entre la couche cachée et la couche de sortie et enfin transformé par une fonction sigmoïde non linéaire S (.).

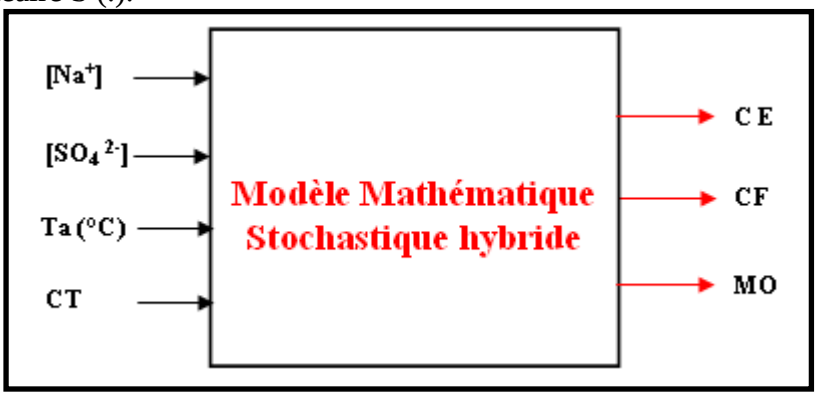

Figure.2: Les entrées /sorties du modèle mathématique stochastique hybride

Les divers modèles du réseau de neurones artificiels utilisés dans ce travail ont été développés et implémentés avec le langage de programmation C++ sur une machine I3 PC, 2.4 GHz et 3 Go de RAM.

Choix de l'architecture du modèle hybride

Pour choisir la « meilleure » architecture de réseau de neurones, plusieurs tests statistiques sont généralement employés, dans notre cas, nous

avons utilisé les tests statistiques Root Mean Square Error RMSE (voir figure 3), Maximum Average Percentage Error MAPE (voir figure 4) et le coefficient de détermination R^2 (voir figure 5).

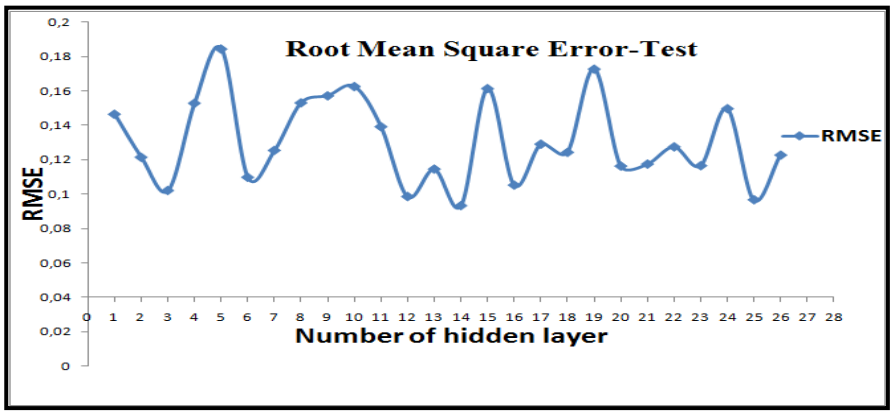

Figure 3: Erreur quadratique moyenne

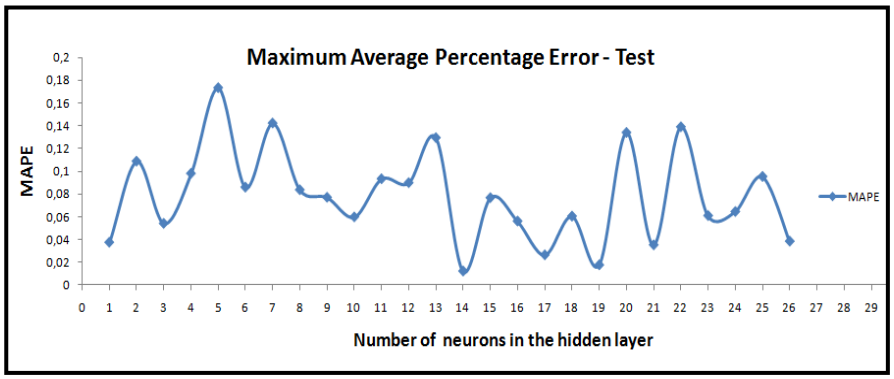

Figure 4: Pourcentage d'erreur moyenne maximale

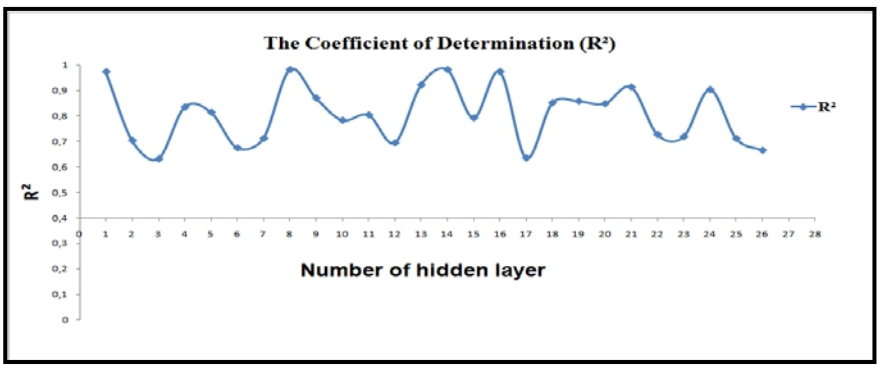

Figure 5. Le coefficient de détermination (R^2)

Après 1500 itérations pour les différents nombres de neurones de la couche cachée, nous avons eu la possibilité de choisir 14 neurones pour la couche cachée. Nous obtenons alors l'architecture [4-14-3] (voir figure 6) comme « meilleure » configuration du réseau de neurones vu sa bonne capacité de prédiction.

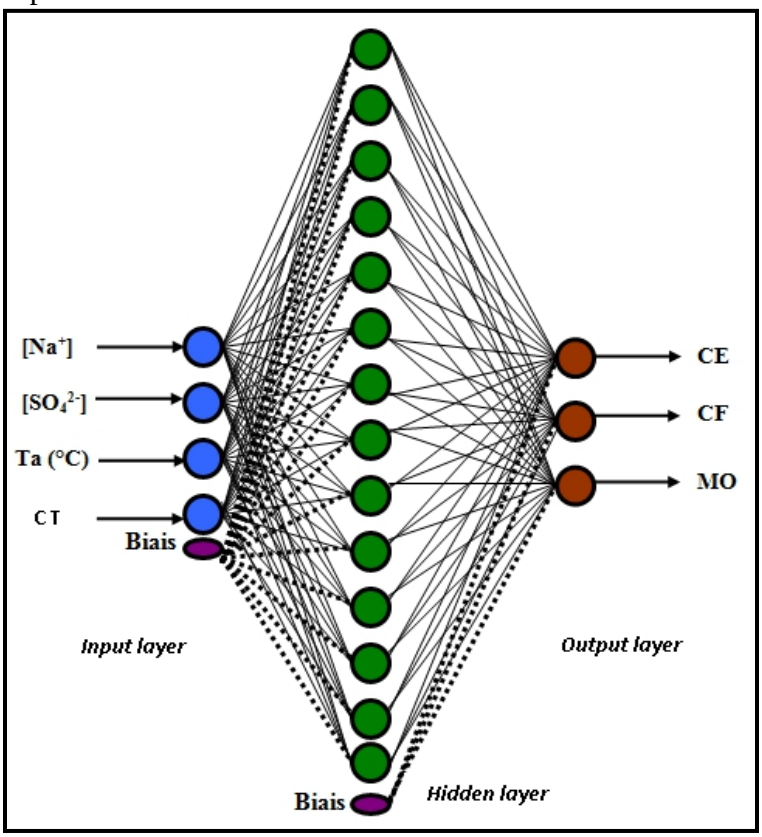

Figure 6: Réseau de neurones architecture [4-14-3]

Apprentissage et validation

La base d'apprentissage se compose de quatre vecteurs au lieu de

Onze vecteurs, indépendants et normalisés entre 0,1et 0,9 qui sont: le sodium Na^+, ion de sulfate SO_4^{2-}, Température de l'air Ta(°C) et Coliforme Totaux CT.

La base d'apprentissage du réseau neuronal est constituée de 32 échantillons. Les poids et les biais du réseau ont été réajustés à l'aide de l'algorithme de rétro propagation du gradient.

Une fois que l'architecture, les poids et les biais du réseau neuronal ont été fixés, il faut savoir si ce modèle neuronal est susceptible d'être généralisé.

La validation de l'architecture neuronale [4-14-3] consiste donc à juger sa capacité de prédiction des paramètres indicateurs de qualité des eaux souterraines de souss massa (CE, MO et CF) en utilisant les poids et les biais calculés durant l'apprentissage, pour les appliquer à une autre base de données tests, composée de 20 échantillons c'est-à-dire 40% de la totalité des données.

Les figures (7-8-9) montrent les résultats de prédiction des paramètres indicateurs de la qualité des eaux souterraines de souss massa (CE, MO et CF), il est remarquable d'après ces figures que les données tests qui se composent de 20 échantillons, sont parfaitement bien modélisées.

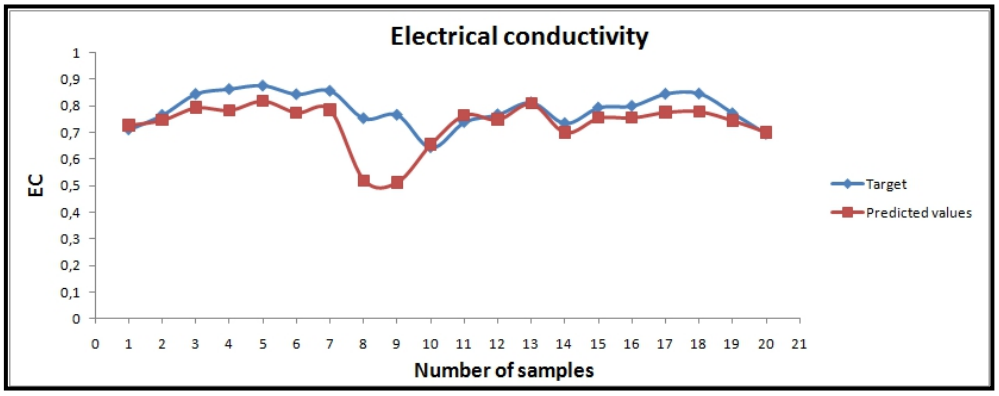

Figure 7: Résultat de test de prédiction de la Conductivité Electrique CE du modèle hybride

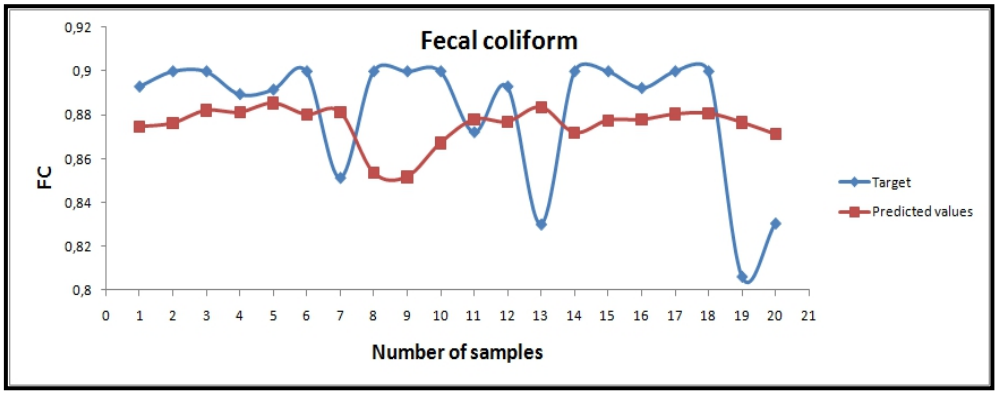

Figure 8: Résultat de test de prédiction la quantité de Coliformes Fécaux CF du du modèle hybride

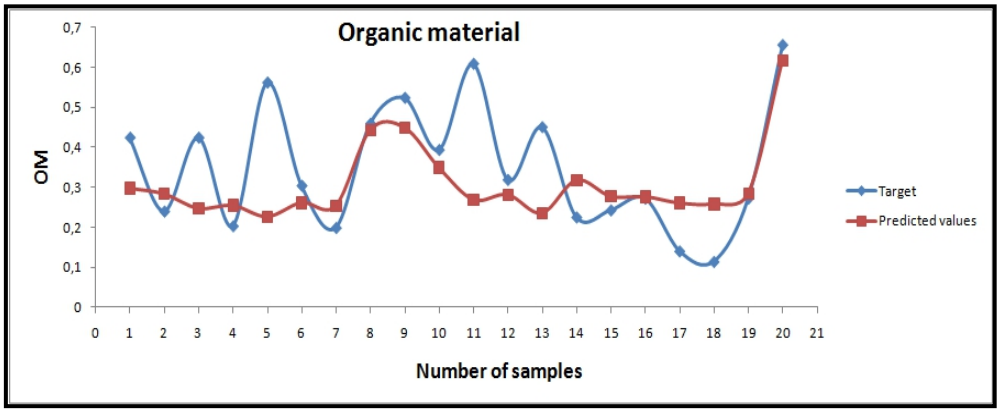

Figure 9: Résultat de test de prédiction de la Matière Organique MO du modèle du hybride

On évalue la qualité de prédiction du modèle hybride [4-14-3] par le coefficient de détermination donné par l'équation suivante:

$$R^2 = \frac{Cov^2(Y^{sim}, Y^{obs})}{V(Y^{sim}).V(Y^{obs})}$$

Dans notre cas le coefficient de détermination est de 98%.

Conclusion

Dans ce travail, nous nous sommes intéressés à construire un modèle de prédiction basé sur les réseaux de neurones artificiels de type perceptron multicouches pour prédire les paramètres microbiologique et physico-chimique, indicateurs de la qualité des eaux souterraines de la nappe de Souss- Massa-Maroc, à partir d'un certain nombre de variables optimisées par la méthode de la régression linéaire multiple pas à pas, le coefficient de détermination obtenu du modèle RNA-MLP [4-14-3] est de l'ordre de 98%, ce coefficient est significativement plus élevé et très satisfaisant par rapport aux études faites précédemment (Manssouri, 2014).

References:

El Badaoui H., Abdallaoui A., Manssouri I., Ousmana H., «The prediction of moisture through the use of neural networks MLP type», IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 11, Issue 6 (May. - Jun. 2013), pp 66-74, 2013.

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

Najjar Y. and Zhang X., « Characterizing the 3D Stress-Strain Behavior of sandy Soils », A Neuro-Mechanistic Approach, volume 96, p 43-57, 2000.
Ryad Z., Daniel R., Zerhouni N. «Réseaux de neurones récurrents à fonctions de base radiales: RRFR», Revue d'Intelligence Artificielle. Volume X – n°X/2002, p 1 à 32, 2002.
Manssouri, I., Chetouani, Y and El Kihel, B. «Using neural networks for fault detection in a distillation column», Int. J. Computer Applications in Technology, Vol. 32, No. 3, pp.181-186, 2008.
D'Orazio, T. Leo, M. Distante, A. Guaragnella, C. Pianese, V. Cavaccini, G.; «Automatic ultrasonic inspection for internal defect detection in composite materials». NDT & E International, Volume 41, Issue 2, pp: 145-154, 2008.
Nagendra, S.M., Khare, M. «Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions, Ecological Modelling», 190: 99-115, 2006.
Cherkaoui Dekkaki Hinde «Evaluation de la vulnérabilité et de la sensibilité des eaux souterraines à la pollution moyennent du SIG et de la télédétection, Application à la nappe phréatique du Souss au niveau des champs captant Sud et Ahmar Boudhar» Thèse de Doctorat à L'Ecole Mohammadia d'Ingénieurs (EMI) sous l'encadrement de M.Pr.H. SAHBI, 2006.
Basheer, I. A. and Hajmeer, M. «Artificial neural networks:fundamentals, computing, design, and application». Journal of Microbiological Methods, vol. 43, p. 3–31, 2000.
Saporta G. , Probabilités, analyse des données et statistique. Ed. Technip, Paris, p.493, 1999.
Manssouri T., Sahbi H., Manssouri I., - Elaboration of stochastic mathematical models for the prediction of parameters indicative of groundwater quality Case of Souss Massa – Morocco- IJCER International Journal of Computational Engineering Research, ISSN (e): 2250 – 3005, Vol, 04, Issue 5, pp 31-40, 2014.