Revue des sciences de l'eau
Journal of Water Science

Modélisation de la dose de coagulant par les systèmes à base d'inférence floue (ANFIS) application à la station de traitement des eaux de Boudouaou (Algérie)

Modelling coagulant dose with an adaptive neuro-fuzzy inference system (anfis) Application to water treatment plant of Boudouaou (Algeria)

Salim Heddam, Abdelmalek Bermad and Noureddine Dechemi

Volume 25, Number 1, 2012

URI: https://id.erudit.org/iderudit/1008532ar
DOI: https://doi.org/10.7202/1008532ar

See table of contents

Publisher(s)

Université du Québec - INRS-Eau, Terre et Environnement (INRS-ETE)

ISSN

1718-8598 (digital)

Explore this journal

Cite this article

Heddam, S., Bermad, A. & Dechemi, N. (2012). Modélisation de la dose de coagulant par les systèmes à base d'inférence floue (ANFIS) application à la station de traitement des eaux de Boudouaou (Algérie). *Revue des sciences de l'eau / Journal of Water Science*, *25*(1), 1–17. https://doi.org/10.7202/1008532ar

Article abstract

Coagulation is an important component of water treatment. Determining the optimal coagulant dosage is vital, as insufficient dosage will result in undesirable treated water quality. A number of chemicals have been used successfully in coagulation, particularly alum (Al_2SO_4•$18H_2O$). Traditionally, jar tests are used to determine the optimum coagulant dose. However, this is expensive and time-consuming and does not enable responses to changes in raw water quality in real time. An optimal modeling approach can be used to overcome these limitations.

The main purpose of this study was to investigate the applicability and capability of Adaptive-Network-Based Fuzzy Inference System (ANFIS) and Neural Network (ANN) methods for modeling coagulant dose. To verify the application of this approach, Boudouaou surface water, located in the northern part of Algeria, was chosen as the case study area. The data used for the determination of the models included six (6) input variables describing the raw water characteristics (temperature, pH, turbidity, conductivity, dissolved oxygen and the ultraviolet absorption). The data set was divided into two (2) subgroups, calibration and validation periods. Coagulant models having various input structures were trained and tested to investigate the applicability of the used methods. To obtain a more accurate evaluation of the results for the ANFIS models, the best fit model structures were also tested by artificial neural network (ANN) and multiple linear regression (MLR) methods. The results of three methods were compared, and it was observed that the ANFIS is preferable and can be applied successfully because it provides high accuracy and reliability for coagulant dosage modelling, according to the following performance evaluation criteria: determination coefficient (R^2), Root Mean Square Error (RMSE) and bias (B). The results are of practical importance: the coagulant dose changes according to the six variables representing the raw water characteristics, as there is no one dominant variable.

RSE

MODÉLISATION DE LA DOSE DE COAGULANT PAR LES SYSTÈMES À BASE D'INFÉRENCE FLOUE (ANFIS) APPLICATION A LA STATION DE TRAITEMENT DES EAUX DE BOUDOUAOU (ALGÉRIE)

Modelling coagulant dose with an adaptive neuro-fuzzy inference system (ANFIS)
Application to water treatment plant of Boudouaou (Algeria)

Salim HEDDAM[1], Abdelmalek BERMAD[2], Noureddine DECHEMI[2]*

[1]Maître Assistant, Faculté des sciences, Département d'Agronomie, Université 20 Août 1955, Route El Hadaik, BP 26, Skikda, 21000 Algérie
[3]Noureddine, DECHEMI, Professeur, Laboratoire Construction et Environnement, École Nationale Polytechnique, 10, avenue Hassen BADI, BP 160, El Harrach, Alger, 16182 Algérie

Reçu le 7 novembre 2008, accepté le 23 août 2011

RÉSUMÉ

La coagulation est l'une des étapes les plus importantes dans le traitement des eaux. La difficulté principale est de déterminer la dose optimale de coagulant à injecter en fonction des caractéristiques de l'eau brute. Un mauvais contrôle de ce procédé peut entraîner une augmentation importante des coûts de fonctionnement et le non-respect des objectifs de qualité en sortie de la station de traitement**.** Le sulfate d'aluminium ($Al_2SO_4.18H_2O$) est le réactif coagulant le plus généralement utilisé. La détermination de la dose de coagulant se fait au moyen de l'essai dit de « Jar Test » conduit en laboratoire. Ce type d'approche a le désavantage d'avoir un temps de retard relativement long et ne permet donc pas un contrôle automatique du procédé de coagulation.

Le présent article décrit un modèle neuro flou de type Takagi Sugeno (TK), développé pour la prédiction de la dose de coagulant utilisée lors de la phase de clarification dans la station de traitement des eaux de Boudouaou qui alimente la ville d'Alger en eau potable. Le modèle ANFIS (système d'inférence flou à base de réseaux de neurones adaptatifs), qui combine les techniques floues et neuronales en formant un réseau à apprentissage supervisé, a été appliqué durant la phase de calage et testé en période de validation. Les résultats obtenus par le modèle ANFIS ont été comparés avec ceux obtenus avec un réseau de neurones de type perceptron multicouche (MLP) et un troisième modèle à base de regression linéaire multiple (MLR). Un coefficient de détermination (R^2) de l'ordre de 0,92 en période de validation a été obtenu avec le modèle ANFIS, alors que pour le MLP, il est de l'ordre de 0,75, et que pour le modèle MLR, il ne dépasse pas 0,35. Les résultats obtenus sont d'une grande importance pour la gestion de l'installation.

Mots clés : modélisation coagulant, réseaux de neurones, modèle neuro floue, analyse en composantes principales, régression linéaire multiple, station de traitement.

*Auteur pour correspondance :
Téléphone: 06 61 74 51 22
Courriel : *heddamsalim@yahoo.fr*

ABSTRACT

Coagulation is an important component of water treatment. Determining the optimal coagulant dosage is vital, as insufficient dosage will result in undesirable treated water quality. A number of chemicals have been used successfully in coagulation, particularly alum (Al_2SO_4•$18H_2O$). Traditionally, jar tests are used to determine the optimum coagulant dose. However, this is expensive and time-consuming and does not enable responses to changes in raw water quality in real time. An optimal modeling approach can be used to overcome these limitations.

The main purpose of this study was to investigate the applicability and capability of Adaptive-Network-Based Fuzzy Inference System (ANFIS) and Neural Network (ANN) methods for modeling coagulant dose. To verify the application of this approach, Boudouaou surface water, located in the northern part of Algeria, was chosen as the case study area. The data used for the determination of the models included six (6) input variables describing the raw water characteristics (temperature, pH, turbidity, conductivity, dissolved oxygen and the ultraviolet absorption). The data set was divided into two (2) subgroups, calibration and validation periods. Coagulant models having various input structures were trained and tested to investigate the applicability of the used methods. To obtain a more accurate evaluation of the results for the ANFIS models, the best fit model structures were also tested by artificial neural network (ANN) and multiple linear regression (MLR) methods. The results of three methods were compared, and it was observed that the ANFIS is preferable and can be applied successfully because it provides high accuracy and reliability for coagulant dosage modelling, according to the following performance evaluation criteria: determination coefficient (R^2), Root Mean Square Error (RMSE) and bias (B). The results are of practical importance: the coagulant dose changes according to the six variables representing the raw water characteristics, as there is no one dominant variable.

Key words: coagulant modelling, neural networks, Fuzzy Inference System, principal components analysis, multiple linear regression, water treatment plant.

1. INTRODUCTION

Au cours des dernières années, l'Algérie a connu une forte demande en eau potable, amplifiée par une forte croissance démographique dans les grands centres urbains. Cette problématique de l'eau a été résolue par de gros investissements mettant en service d'importantes infrastructures de production et de distribution d'eau potable. Etant donné que le traitement des eaux est une étape indispensable et réglementée, les stations de traitement de l'eau potable figurent parmi les infrastructures auxquelles une grande importance a été donnée.

Deux principaux objectifs sont ciblés lors de la gestion et l'exploitation d'une station de traitement des eaux : un objectif de qualité et un objectif de coût (BAXTER, 1998). L'atteinte de l'objectif de qualité exige la production d'eau en quantité suffisante, de façon continue et avec une qualité répondant aux normes de potabilité en vigueur (normes chimiques, microbiologiques, etc.) lesquelles sont de plus en plus sévères. Cela implique la conception d'une station de traitement sophistiquée correctement dimensionnée et incluant un large ensemble d'équipements mécaniques et hydromécaniques. Au passage de l'eau à traiter dans un tel système, la qualité de l'eau obtenue se trouve modifiée tant au niveau physique que chimique ou microbiologique. La nature et l'étendue de ces modifications dépendent des caractéristiques de l'eau à l'entrée de la station et du degré d'interaction entre les différentes composantes du processus mis en jeu.

En raison du grand nombre de variables, ainsi que de la complexité des phénomènes biologiques, physiques et chimiques impliqués dans les procédés de traitement de l'eau potable, il est souvent très difficile de quantifier au préalable les interactions et les relations qui existent entre les entrées (variables) et les sorties (paramètres de qualité) de ces procédés (BAXTER *et al.,* 2002**).** La coagulation est une étape très importante dans la production de l'eau potable à partir d'eau brute. Elle a pour but la déstabilisation des colloïdes et leur agglomération ainsi que celle des particules fines en suspension (AMIRTHARAJAH et O'MELIA, 1990)**.** La consommation d'agent coagulant fait de cette étape de traitement l'opération la plus coûteuse dans la chaîne de traitement.

Les produits chimiques impliqués dans le processus de coagulation (coagulants) sont d'origine minérale à base d'aluminium ($Al_2(SO_4)_3.18H_2O$) (DEMPSEY *et al.,* 1985) ou de fer ($FeCl_3$) (LEFEBVRE et LEGUBE, 1993), des polymères naturels ou des polymères de synthèse. Dans le cas de notre étude, le coagulant utilisé au niveau de la station de traitement des eaux de Boudouaou (prés de la ville d'Alger) est le sulfate d'aluminium. Ainsi, la mise en œuvre se déroule en deux étapes (Équation 1) (CARDOT, 1999) :

$$\underset{}{Al_2(SO_4)_3} \xrightarrow[\text{étape1}]{} \underset{\text{intermédiaires polychargés}}{Al_x(OH)_y(SO_4)_z} \xrightarrow[\text{étape2}]{} \underset{\text{précipité}}{Al(OH)_3} \quad (1)$$

L'étape 1 est une phase d'hydrolyse, tandis que durant la phase 2 il y a formation du précipité $Al(OH)_3$.

L'ajout du coagulant dans l'eau a les effets suivants : (i) réduction de la charge hydrostatique par son adsorption à la surface des particules; (ii) réduction de la charge diffuse. De ce fait, les principaux facteurs influençant l'efficacité de la coagulation sont le pH (STUMM et MORGAN, 1962), la turbidité initiale (EDWARDS et AMIRTHARAJAH, 1985) et la température de l'eau (MOHTADI et RAO, 1973). D'autres variables caractérisant l'eau brute influent considérablement sur le processus de coagulation à savoir, la conductivité de l'eau, l'absorbance à 254 nm (UV_{254}) ainsi que l'oxygène dissous (OD) (LIND, 1994a, 1994b)**.** L'absorbance à 254 nm exprime la capacité de l'eau à absorber un rayonnement UV_{254} à une longueur d'onde de 254 nanomètres. Cette mesure permet au professionnel de s'assurer que la désinfection aux ultraviolets est possible (WEISHAAR *et al.*, 2003).

Lors de la phase de coagulation, on cherche, d'une part, à maximiser la déstabilisation des particules et des colloïdes organiques pour faciliter leur agglomération et leur enlèvement subséquent, par un procédé de séparation solide-liquide et, d'autre part, à minimiser la concentration en coagulant résiduel. La minimisation des coûts de l'opération se fait par une coagulation que l'on juge optimale. Elle correspond au dosage du coagulant qui assure l'atteinte de tous les objectifs de qualité (EDZWALD et TOBIASON, 1999). Afin d'évaluer les conditions optimales de coagulation et de floculation, des essais dits de « Jar-Test » (JT) sont conduits à l'échelle de laboratoire. Ceux-ci, menés dans une large gamme de conditions opératoires, permettent de déterminer le type de coagulant, son dosage, le pH et les conditions d'agitation qui maximisent la réduction de la turbidité (KRASNER et AMY, 1995).

Ce type d'approche a l'inconvénient d'avoir un temps de réponse relativement long. En effet, on ne modifie la dose de coagulant qu'une fois un événement apparu. De plus, elle ne permet pas de suivre finement l'évolution de la qualité de l'eau brute (BAXTER *et al.,* 1999). On voit ici tout l'intérêt de disposer d'un contrôle automatique et efficace de ce procédé pour un meilleur rendement de traitement et une réduction des coûts d'exploitation. Au cours des dernières années une nouvelle approche a été développée, qui est la régulation du procédé de coagulation basée sur les variables descriptives de la qualité de l'eau brute. Cette technique impose de trouver un modèle reliant la dose optimale de coagulant à ces différentes variables (VALENTIN, 2000).

La modélisation par régression entrée/sortie a déjà fait l'objet de nombreuses applications dans ce domaine (VAN LEEUWEN *et al.,* 1999). Les approches proposées reposent le plus souvent sur des modèles régressifs linéaires**.**

BAZER-BACHI *et al.* (1990) ont proposé deux modèles mathématiques basés sur des équations polynomiales, reliant la dose optimale du coagulant (le sulfate d'aluminium) aux variables descriptives de la qualité de l'eau brute à savoir : la turbidité, la résistivité, la teneur en matière organique, la température et la nature de la suspension minérale. D'autres modèles linéaires ont été proposés (CRITCHLEY *et al., 1990;* ELLIS *et al.,* 1991; GIROU *et al.,* 1992; RATNAWEERA et BLOM, 1995). Le modèle de GIROU *et al.* (1992) est basé sur la concentration en ions calcium, les bicarbonates, les sulfates, la turbidité initiale, la température et le pH. Les données utilisées dans le modèle développé par RATNAWEERA et BLOM (1995) sont le débit de la rivière, le temps de sédimentation, la température, la turbidité, le pH et la conductivité, alors que le modèle proposé par CRITCHLEY *et al. (1990*) inclut la couleur, le débit de la rivière, le pH, la conductivité et la température. Ces études ont montré l'intérêt de cette approche mais également les limites de la modélisation linéaire pour ce type de problème.

Les progrès importants réalisés au cours des dernières années dans le domaine de l'intelligence artificielle ont permis de réduire les difficultés et de s'affranchir des limitations des modèles linéaires. Des modèles basés sur la technique des réseaux de neurones artificiels ont été mis au point (MAIER *et al.,* 2004). Un exemple de ce modèle a déjà été testé (VALENTIN *et al.,* 1999). Cette modélisation a été intégrée dans le cadre de la construction d'un capteur logiciel pour la détermination en ligne de la dose optimale de coagulant en fonction de différentes caractéristiques de la qualité de l'eau brute telles que la turbidité, le pH, la conductivité, etc. Le modèle de VALENTIN *et al.* (1999) est basé sur deux types de réseaux de neurones, un perceptron multicouche (MLP), d'une part, et un réseau basé principalement sur l'utilisation des cartes auto-organisatrices de Kohonen pour le prétraitement des données, d'autre part.

D'autres modèles ont été proposés (ADGAR *et al.,* 1995; ADGAR *et al.,* 2000; BÖHME *et al.,* 1999; GAGNON *et al.,* 1997; MIRSEPASSI *et al.,* 1997; NAHM *et al.,* 1996; YU *et al.,* 2000). Ils expriment tous la dose du coagulant à injecter en fonction des différentes variables descriptives caractérisant l'eau brute à l'entrée de la station de traitement des eaux. Certaines études (BAXTER *et al.,* 2001a; BAXTER *et al.,* 2001b; BAXTER *et al.,* 2002; COX *et al.,* 2003; HEDDAM *et al., 2011;* LAMRINI *et al.,* 2005) ont montré l'importance des réseaux de neurones comme outil pour l'élaboration des modèles mathématiques à des fins d'automatisation et de supervision des procédés impliqués dans les stations de traitement des eaux.

Dans cet article, on propose une autre méthode de prédiction de la dose du coagulant en fonction de six variables descriptives caractérisant l'eau brute à l'entrée de la station de traitement des eaux potables. Cette méthode est basée sur le modèle ANFIS (Adaptive Neuro Fuzzy Inference System), qui

combine la logique floue et les réseaux de neurones pour former un réseau hybride, utilisant la rétropropagation de l'erreur comme algorithme d'apprentissage. Les résultats obtenus sont comparés à ceux d'un modèle à base de réseaux de neurones artificiels, le perceptron multicouche (MLP) et d'un modèle à base de régression linéaire multiple (RLM).

2. MODÈLES UTILISÉS

2.1 La régression linéaire multiple

La régression linéaire multiple (RLM) est une généralisation du modèle de régression simple lorsque les variables explicatives sont en nombre fini. Elle consiste à rechercher une équation linéaire reliant la variable à modéliser Y = {y_i, i = 1...N} (variable à expliquer ou endogène) à la matrice d'entrées ou (variables explicatives ou exogènes), X = {x_{ip}, i = 1...N; p, nombre de variables explicatives}. N correspond au nombre d'individus ou d'observations.

L'équation linéaire recherchée est de la forme

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots \beta_p x_{ip}, i = 1, \ldots\ldots, N \quad (2)$$

Les paramètres β sont appelés coefficients de régression partielle. Ils mesurent l'influence de chacune des variables sur la grandeur étudiée. On remarque que le nombre de paramètres à déterminer pour un modèle à base de régression linéaire (RLM) est au nombre de (p+1).

2.2 Les réseaux de neurones artificiels

Les réseaux de neurones artificiels (RNA) sont des modèles mathématiques non linéaires, de type « boîte noire », capables de déterminer des relations entre données par la présentation (l'analyse) répétée d'exemples, à savoir de couples constitués par une information d'entrée (variables caractéristiques de l'eau brute) et une valeur de sortie que l'on voudrait approcher par le modèle (la dose de coagulant). Les RNA se composent d'un ensemble de processeurs élémentaires, les neurones qui sont largement connectés les uns aux autres et qui sont capables d'échanger des informations au moyen des connexions qui les relient. Les connexions sont directionnelles et à chacune d'elle est associé un réel appelé poids de la connexion. Cette représentation est le reflet de l'inspiration biologique qui a été à l'origine de la première vague d'intérêt pour les neurones formels, dans les années 1940 à 1970 (McCULLOCH et PITTS, 1943).

Dans cet article, nous considérerons une structure très particulière des réseaux de neurones, les perceptrons multicouches (MLP pour Multi Layer Perceptron) décrits dans la figure 1. Un perceptron multicouche consiste en une succession de couches constituées d'unités neuronales, lesquelles possèdent une fonction d'activation non linéaire. À l'intérieur d'une couche chaque neurone reçoit des signaux provenant de la couche précédente, effectue un calcul et transmet le résultat à la couche suivante. Il n'existe pas d'interconnexions entre les neurones situés à l'intérieur d'une même couche : les activations des différents neurones sont seulement propagées de la couche d'entrée vers la couche de sortie à travers tous les neurones constitutifs du réseau. La couche d'entrée collecte les variables d'entrée tandis que la couche de sortie produit les résultats.

La première couche du réseau est la couche d'entrée. Elle contient (n) neurones. La deuxième couche, appelée couche cachée, contient pour sa part (m) neurones. La dernière couche du réseau est sa couche de sortie qui contient (p) neurones. Les neurones d'entrée sont numérotés de 1 à n, les neurones cachés de 1 à m, et les neurones de sortie de 1 à p. Par convention, le paramètre w_{ij} est relatif à la connexion allant du neurone i (ou de l'entrée i) vers le neurone j. Ainsi le paramètre w_{jk} est relatif à la connexion allant du neurone caché j vers le neurone de sortie k.

Les états des neurones de la première couche seront fixés par le problème traité à travers un vecteur x = (x_1; x_2; …x_n). Les états de la première couche étant fixés, le réseau va pouvoir calculer les états des neurones des autres couches. Dans ce sens, chaque neurone de la couche cachée reçoit une somme pondérée par les paramètres (w_{ij}), qui sont alors souvent désignés sous le nom de « poids » ou, en raison de l'inspiration biologique des réseaux de neurones, « poids synaptiques », de toutes les entrées, à laquelle s'ajoute un terme constant w_0 ou « biais » :

$$\text{‘}_j = w_0 + \sum_{i=j}^{n} w_{ij} \times x_i \quad (3)$$

La sortie du neurone est une fonction non linéaire de son entrée (A_j) :

$$Y_j = f\left(A_j\right) \quad (4)$$

La fonction f est appelée fonction de transfert ou d'activation. On utilise le plus souvent une fonction d'activation sigmoïde, appliquée dans cette étude et donnée par la formule suivante :

$$S(y) = \frac{1}{1 + e^{-y}} \quad (5)$$

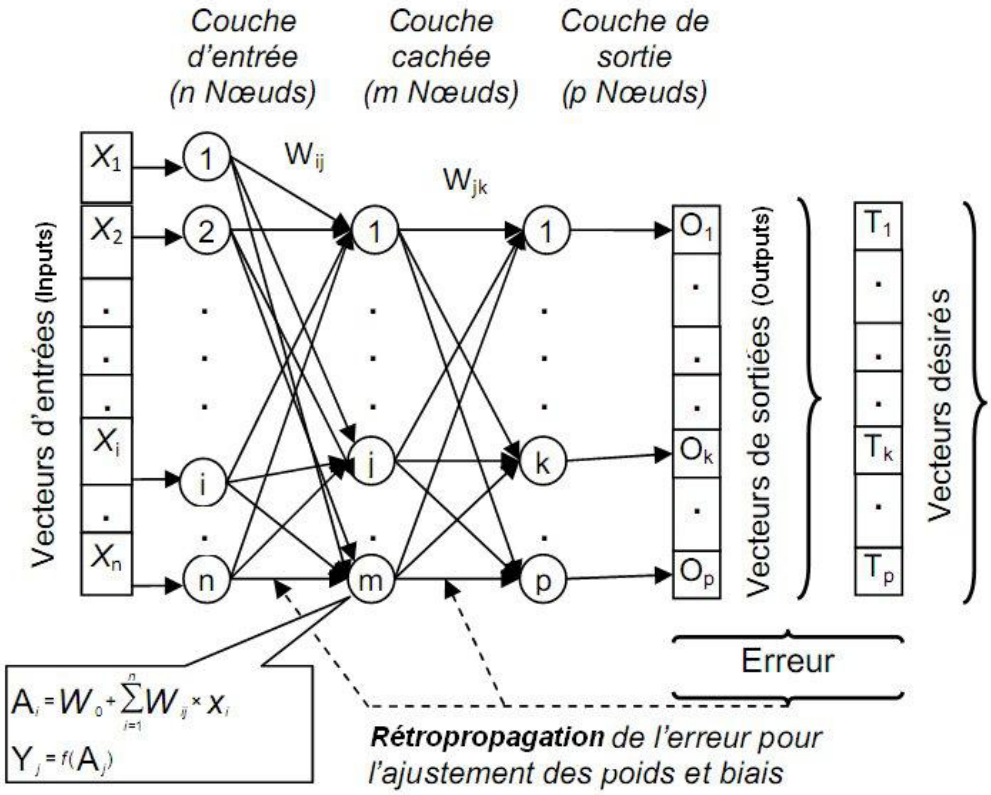

Figure 1. Architecture du perceptron multicouche modèle MLP.
Architecture of the Multilayer perceptron neural network MLP.

Chaque neurone de la couche de sortie reçoit une somme pondérée par les paramètres (w_{jk}), à laquelle s'ajoute un terme constant B_0 ou « biais » :

$$O_K = B_0 + \sum_{j=1}^{m} w_{jk} \times Y_j \qquad (6)$$

La sortie du réseau O (pour Output) est une fonction linéaire des poids de la dernière couche de connexions (qui relient les m neurones cachés aux neurones de sortie), et elle est une fonction non linéaire des paramètres de la première couche de connexions (qui relient les n entrées du réseau aux m neurones cachés). Cette propriété a des conséquences très importantes (DREYFUS, 2004). Il a été démontré qu'un réseau de neurones comportant une couche de neurones cachés en nombre fini, possédant tous la même fonction d'activation, et un neurone de sortie linéaire est un approximateur universel (HORNIK *et al.,* 1989; HORNIK *et al.,* 1990; HORNIK, 1991).

Les valeurs des poids et du biais sont modifiées et mises à jour via un algorithme d'apprentissage supervisé. Ce dernier consiste à se procurer un ensemble d'exemples, c'est-à-dire un ensemble fini de couples entrée sortie connus (exemples qui constituent l'ensemble d'apprentissage). L'objectif de ce calcul est la minimisation d'une fonction d'erreur entre la réponse désirée et la réponse obtenue à la sortie du modèle. L'algorithme de rétropropagation de l'erreur est le plus utilisé. Ce dernier estime le gradient de la fonction d'erreur par rapport aux paramètres (poids et biais) du modèle et réalise l'adaptation de ces paramètres successivement de la couche de sortie vers la couche d'entrée (Figure 1). Cela consiste à effectuer une descente de gradient sur le critère d'erreur 'E' en minimisant une fonction coût, généralement l'erreur quadratique moyenne (RUMELHART *et al.,* 1986).

Les méthodes de gradient peuvent être réparties en deux catégories : les méthodes du premier ordre, qui n'utilisent que le gradient de la fonction (cas de l'algorithme de rétropropagation de l'erreur) et les méthodes du second ordre, qui généralisent la

descente du gradient au deuxième degré de la fonction d'erreur. Ce sont des méthodes itératives qui consistent à remplacer la fonction coût par son approximation quadratique. On peut citer par exemple les méthodes de Newton, de quasi-Newton et de Levenberg-Maquardt. Cette dernière est utilisée dans le cadre de notre étude.

2.3 Modèle neuroflou

2.3.1 La logique floue

La logique floue a été développée par ZADEH (1965) qui a proposé de modéliser un système complexe par un raisonnement « approximatif » basé sur des variables linguistiques et des sous-ensembles flous (ZADEH, 1971). Un sous-ensemble flou A est défini sur un domaine physique appelé univers de discours U, et par une fonction d'appartenance F(x) qui associe, à chaque élément x de U, le degré de vérité (d'appartenance) $f_{(A)}$ à X compris entre l'intervalle 0 et 1, soit, et cela contrairement à la logique classique où le degré d'appartenance ne peut prendre que deux valeurs (0 ou 1).

Les systèmes flous s'appuient sur une représentation de la connaissance sous forme de règles « Si…..Alors » qui permettent de représenter les relations entre les variables d'entrée et de sortie dont l'expression générique est de la forme :

$$\textbf{Si}\left(\text{Antécédent}\right) \quad \textbf{Alors}\left(\text{Conséquent}\right) \tag{7}$$

$$\textbf{Si}\left(\text{X est A}\right) \quad \textbf{Alors}\left(\text{Y est B}\right) \tag{8}$$

L'antécédent (prémisse) est une description linguistique qui indique les conditions de validité du phénomène représenté. Pour sa part, le conséquent (conclusion) représente le comportement associé aux conditions de validité décrites par l'antécédent, par exemple :

$$\begin{aligned}&\textbf{Si}\left(\text{la turbidité de l'eau est élevée}\right)\\&\textbf{Alors}\left(\text{dose de coagulant est élevée}\right)\end{aligned} \tag{9}$$

Aujourd'hui la logique floue a fait l'objet de plusieurs applications dans le domaine de l'ingénierie (BENKACI et DECHEMI, 2004; DECHEMI *et al.,* 2003; LEKFIR *et al.,* 2006).

2.3.2 Caractéristiques des sous ensembles flous

Une variable linguistique (ZADEH, 1971) est une variable dont les valeurs sont des mots ou des phrases exprimées dans une langue naturelle ou un langage artificiel (NAKOULA, 1997). Une variable linguistique est définie par : « x_nom, L(x), U, M_X » avec : (i) x_nom : le nom de la variable linguistique (ex : dose du coagulant), (ii) L(x) = {L_1; L_2;...; L_n} est l'ensemble des valeurs linguistiques (ou encore appelé symbole ou terme linguistique ou étiquette) que peut prendre la variable x_nom. Par exemple L(x) = {faible, moyenne, élevée} pour caractériser la dose de coagulant; (iii) U correspond à l'univers de discours associé à la variable x_nom (exemple : dose du coagulant varie entre 5 et 35 mg•L^{-1}). C'est l'ensemble de toutes les valeurs numériques que peut prendre la variable numérique associée à la variable linguistique x_nom; (iv) M_X est une fonction qui associe à tout symbole de L(x) une signification floue.

La modélisation d'un système entrée/sortie par la logique floue passe par trois étapes essentielles :

- La fuzzification des variables d'entrée, qui consiste à transformer les entrées numériques disponibles en parties floues. Il est alors possible d'associer à des variables des coefficients d'appartenance à des sous-ensembles flous prenant des valeurs dans l'intervalle [0,1].
- L'inférence floue, composée par la base de règles et par la base de données. La combinaison des entrées avec les règles floues permet de tirer des conclusions.
- La défuzzification qui est l'opération inverse de la fuzzification. Elle convertit les parties floues relatives aux sorties du mécanisme d'inférence en sorties numériques. Il existe plusieurs techniques de défuzzification (JANG *et al.,* 1997). Cependant la technique la plus utilisée est celle du centre de gravité (LEE, 1990).

2.3.3 Modèle flou utilisé : le modèle de Sugeno

Les systèmes flous sont répertoriés selon leur nature structurelle. On distingue les systèmes flous à conclusions symboliques (MAMDANI, 1977) ou modèles flous linguistiques (systèmes de Mamdani), dans lesquels l'antécédent et le conséquent sont tous les deux des propositions floues qui utilisent des variables linguistiques (Équation 10), et des systèmes flous à conclusions fonctionnelles ou modèles flous de Takagi-Sugeno-Kang (TS) (Équation 11) (TAKAGI et SUGENO, 1985).

$$\begin{aligned}&\textbf{Si}\left(\text{la turbidité est élevée et le pH est faible}\right)\\&\textbf{Alors}\left(\text{dose de coagulant est élevée}\right)\end{aligned} \tag{10}$$

$$\begin{aligned}&\textbf{Si}\left(\text{la turbidité est élevée et le pH est faible}\right)\\&\textbf{Alors}\left(\text{D} = 25\,\text{T} + 30\,\text{P} + 5\right)\end{aligned} \tag{11}$$

Étant donné que notre étude concerne un système d'entrée/sortie, nous nous sommes basés sur le modèle de Takagi_Sugeno (TS) de premier ordre. Dans ce cas la variable D du conséquent (dose du coagulant) est numérique sous la forme d'une fonction des variables associées à l'antécédent (Équation 11). Ici T et P représentent respectivement les valeurs numériques de la turbidité et du pH, données à titre d'exemple.

2.3.4 Le modèle ANFIS

L'utilisation conjointe des méthodes neuronales et floues dans des modèles hybrides permet de tirer des avantages, principalement, des capacités d'apprentissage des réseaux de neurones, et de la lisibilité et la souplesse de la logique floue. Le principal type d'association entre réseaux de neurones et systèmes flous est le cas où un système d'inférence flou est mis sous la forme d'un réseau multicouche (BUCKLEY et HAYASHI, 1994), dans lequel les poids correspondent aux paramètres du système, l'architecture du réseau dépendant du type de règles et des méthodes d'inférence, d'agrégation et de défuzzification choisies. Le plus utilisé dans ce domaine est le modèle ANFIS.

Le modèle ANFIS, connu sous le nom de réseau adaptatif à base de système d'inférence floue, développé par JANG (1993) est un approximateur universel (JANG *et al.,* 1997). ANFIS est une technique qui incorpore les concepts de la logique floue dans les réseaux de neurones. Il a été largement utilisé dans beaucoup d'applications (KISI ,2005; TUTMEZ *et al.,* 2006).

Ce modèle simule la relation entre l'entrée et la sortie d'un processus à travers un apprentissage hybride pour déterminer la distribution optimale des fonctions d'appartenances (Figure 2). Il est basé sur les règles floues « Si…..Alors » de Takagi et Sugeno (TAKAGI et SUGENO, 1985). L'architecture équivalente du modèle comporte cinq couches, chacune comportant plusieurs nœuds (Figure 2).Les nœuds carrés (adaptatifs) contiennent des paramètres, alors que les nœuds circulaires (fixes) n'ont pas de paramètres dans le système.

Pour deux variables d'entrée x1 (la température) et x2 (la conductivité) données à titre d'exemple avec la seule variable de sortie Y (la dose du coagulant), chaque variable d'entrée est décrite par deux termes linguistiques : M1 et M2 pour la variable x1, L1 et L2 pour la variable x2, respectivement, d'où une base de règle « Si…..Alors » décrite par deux règles floues R1 et R2 :

$$R1: \text{ Si } x_1 \text{est } M_1 \text{et } x_2 \text{est } L_1 \text{donc } y = f_1(x) \quad (12)$$

$$R2: \text{ Si } x_1 \text{est } M_2 \text{et } x_2 \text{est } L_2 \text{donc } y = f_2(x) \quad (13)$$

$$f_1(x) = p_1x_1 + q_1x_2 + r_1 \quad (14)$$

$$f_2(x) = p_2x_1 + q_2x_2 + r_2 \quad (15)$$

où p_i, q_i, r_i correspondent aux paramètres de la partie conclusion à ajuster durant l'apprentissage.

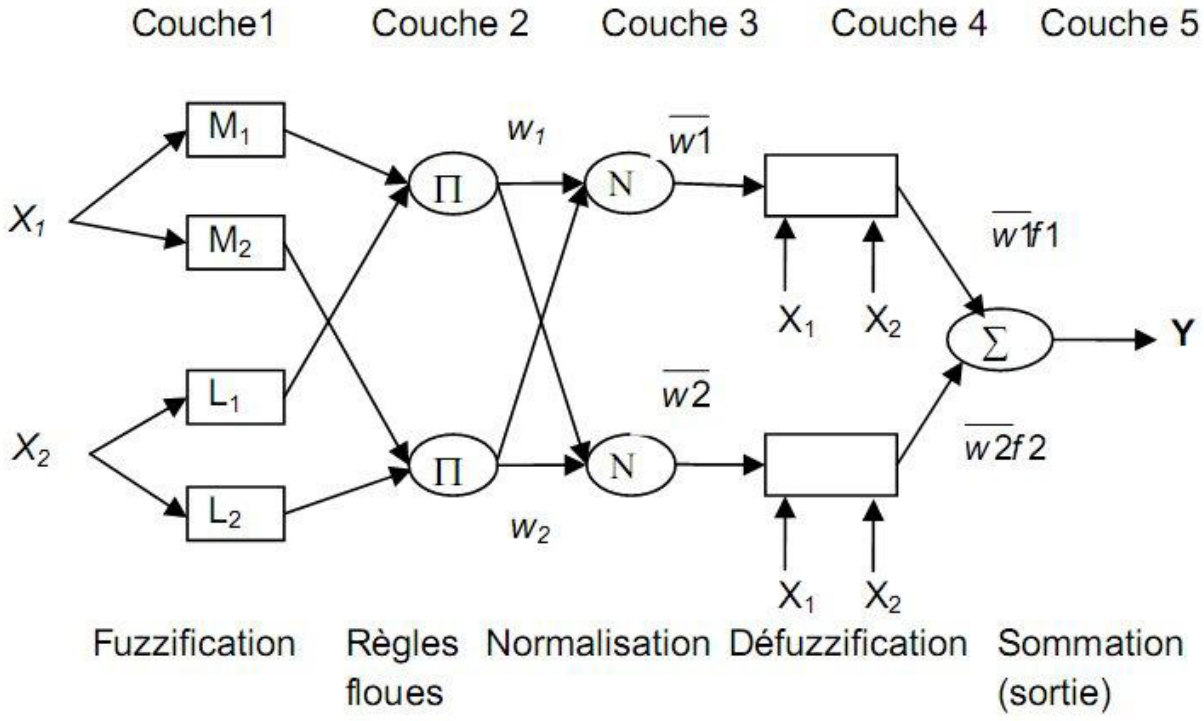

Figure 2. Architecture du modèle ANFIS.
Architecture of the ANFIS model.

- Couche 1 : chaque nœud de cette couche est un nœud carré adaptatif avec une fonction :

$$O_{1,i} = \mu_{Mi}(x_1) \text{ pour } i = 1,2,\ldots, m \quad (16)$$

$$O_{1,i} = \mu_{L_{i2}}(x_2) \text{ pour } i = 1,2,\ldots, m \quad (17)$$

où x_1 (ou x_2) est l'entrée du nœud i, M_i (ou L_i) est le terme linguistique associé à sa fonction.

Les nœuds de cette couche représentent le degré d'appartenance de x_1 (ou x_2) à M_i (ou L_i); c'est la phase de fuzzification.

- Couche 2 : chaque nœud de cette couche est un nœud circulaire fixe, appelé ($\prod$), qui reçoit les sorties des nœuds de fuzzification et calcule leur activation. Le nombre de nœuds dans cette couche est égal au nombre de règles « Si…..Alors » dans le système d'inférence flou.

$$O_{2,i} = w_i = \mu_{Mi}(x_1) \cdot \mu_{Li}(x_2),\ i = 1,2 \quad (18)$$

- Couche 3 : chaque nœud de cette couche est un nœud circulaire fixe, appelé (N).C'est la couche de normalisation dans laquelle chaque nœud calcule le degré d'appartenance normalisé à une règle floue donnée. Le résultat obtenu représente la participation de chaque règle floue au résultat final. Cette couche renvoie des sorties normalisées de défuzzification.

$$O_{3,i} = \bar{w}_i = \frac{w_i}{w_1 + w_2} \quad (19)$$

- Couche 4 : Chaque nœud i de cette couche est un nœud carré adaptatif qui correspond à l'entrée initiale pondérée par le degré d'appartenance normalisé de la règle floue.

$$O_{4,i} = \bar{w}_i f_i = \bar{w}_i (pi.x_1 + qi.x_2 + ri) \quad (20)$$

où $\bar{w}_i$ est la sortie normalisée de la couche 3, et {pi, qi, ri} est l'ensemble des paramètres de sortie de la règle i. C'est la phase de défuzzification.

- Couche 5 : composée d'un seul nœud fixe circulaire appelé (Σ) qui reçoit la somme des sorties de tous les nœuds de défuzzification, et fournit la sortie du modèle ANFIS.

$$O_{5,i} = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad (21)$$

2.3.5. Apprentissage du modèle ANFIS

L'ajustement des paramètres de l'ANFIS est réalisé lors de la phase d'apprentissage. Pour cela, un ensemble de données associant séquences d'entrées et de sorties est nécessaire. Pour la réalisation de cette phase, l'algorithme d'apprentissage hybride est utilisé. L'algorithme d'apprentissage hybride est une association de la méthode de descente de gradient de l'erreur et de la méthode d'estimation des moindres carrés. La méthode de descente de gradient de l'erreur permet d'ajuster les prémisses alors que la méthode LSM (Least Square Method) ajuste les paramètres linéaires (conséquents ou conclusions). L'apprentissage se fait de façon itérative jusqu'à ce que le nombre de cycles d'apprentissage soit atteint ou jusqu'à ce que l'erreur moyenne entre la valeur de sortie désirée et générée par l'ANFIS atteigne une valeur prédéterminée. Cette phase dépend donc de la qualité de l'ensemble des données au sens où cet ensemble doit représenter au mieux les différents comportements attendus (WANG et MENDEL, 1992a et 1992b).

Le modèle ANFIS permet de s'affranchir de l'effet « boîte noire » reproché aux réseaux de neurones classiques, d'associer la connaissance dysfonctionnelle disponible sous la forme de règles floues et de conserver une capacité d'apprentissage issue des réseaux de neurones. Une des plus importantes étapes pour la génération de la structure des réseaux neuro flous ANFIS est l'établissement des règles d'inférence floues. En utilisant un mécanisme d'inférence, les règles sont définies comme combinaisons des fonctions d'appartenance des différentes variables d'entrée. Les variables d'entrée sont divisées en un nombre limité de valeurs linguistiques (étiquette), chacune caractérisée par une fonction d'appartenance (et leurs combinaisons mènent à beaucoup de règles d'inférences floues).

2.4 Validation des modèles

La validation permet de juger l'aptitude du modèle à reproduire les variables modélisées. Plusieurs critères ont été choisis. Dans notre cas, nous nous sommes basés sur le coefficient de détermination (R^2), la racine de l'erreur *quadratique* moyenne (RMSE) et la moyenne biaisée (B).

2.4.1 Coefficient de détermination (R^2)

$$R^2 = \left[\frac{\frac{1}{N} \sum_{i=1}^{N} \left(Y_{iobs} - \bar{Y}_{obs}\right)\left(Y_{ical} - \bar{Y}_{cal}\right)}{\sigma_{obs}.\sigma_{cal}} \right]^2 \quad (22)$$

Avec : Y_{iobs} et Y_{ical} correspondent respectivement aux valeurs observées et calculées par le modèle de la dose du coagulant pour la journée i, $\overline{Y}_{obs}$ et $\overline{Y}_{cal}$ sont les moyennes des valeurs observées et calculées par le modèle, et σ_{obs} et σ_{cal} les écarts-types des valeurs observées et calculées.

2.4.2 Racine de l'erreur quadratique moyenne (RMSE)

$$RMSE = \sqrt{\frac{\sum \left(Y_{obs} - Y_{cal} \right)^2}{N}} \quad (23)$$

N représente le nombre de valeurs utilisées. Le modèle est bien optimisé si la valeur du RMSE est proche de zéro.

2.4.3 Moyenne biaisée (B)

C'est la différence entre la moyenne des doses de coagulant observées et celles calculées. Ce paramètre est défini par la relation suivante :

$$B = \overline{Y}_{obs} - \overline{Y}_{sim} \quad (24)$$

Lorsque B tend vers zéro, le modèle est sans biais.

3. PRÉSENTATION DE LA STATION ÉTUDIÉE

La station de traitement des eaux potables a été mise en service en 1987. Cette station se situe à environ 7 km du barrage de Keddara, entre les villes de Boudouaou et d'Ouled Moussa (Figure 3). Elle occupe une superficie de 17 hectares et fait partie du Système de Production Isser Keddara (SPIK). Elle traite les eaux des barrages de Béni Amrane, de Keddara et du Hamiz et alimente la population de la capitale (Alger), estimée à 4 000 000 d'habitants, avec une capacité de traitement de 540 000 $m^3 \bullet j^{-1}$ (SEAAL, 2008).

Cette station de traitement compte : (i) un ouvrage d'arrivée et de mélange, (ii) une étape de clarification assurée par le procédé de coagulation-floculation grâce à des décanteurs de type « PULSATOR » lamellaires à lit de boue, utilisant le sulfate d'aluminium comme coagulant, (iii) des filtres type « AQUAZUR V ». Après ce traitement, l'eau est stockée dans deux réservoirs de capacité totale 2 x 50 000 m^3, avant qu'elle ne soit pompée vers la ville d'Alger (SEAAL, 2008).

L'objectif de notre travail est la modélisation de la dose du coagulant (DC) en fonction des variables descriptives caractérisant l'eau brute à l'entrée de la station. Nous disposons

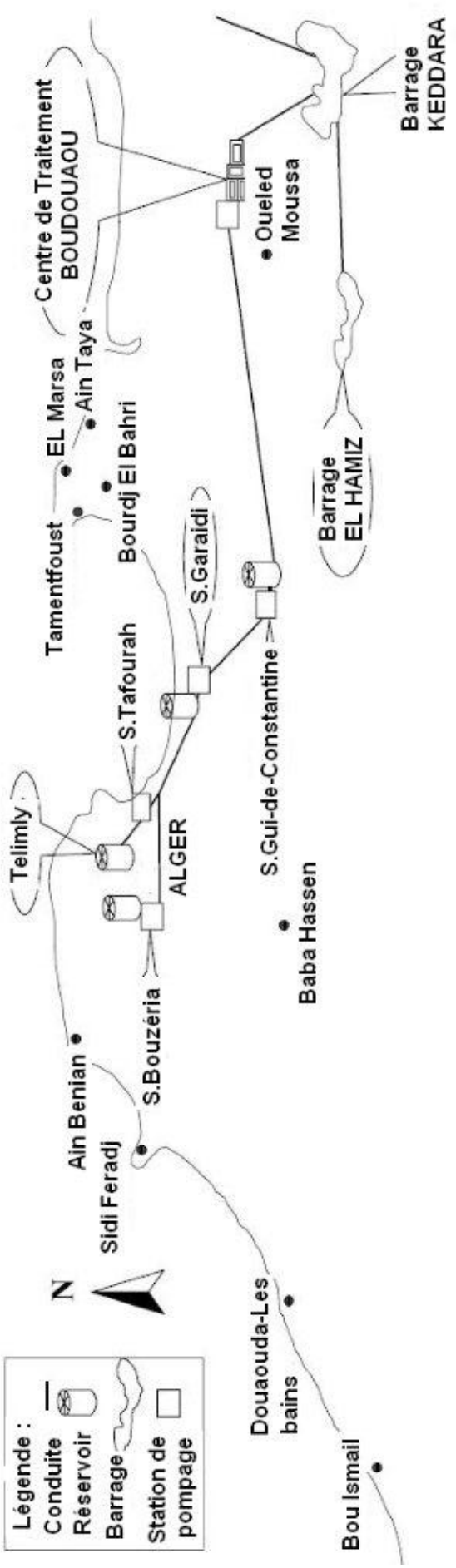

Figure 3. Présentation de l'usine de production.
Presentation of the production factory.

pour cela de six variables : la température (TE), le pH, la turbidité (TU), la conductivité électrique de l'eau (CE), l'oxygène dissous (OD) et l'absorbance à 254 nm (UV_{254}). Ces six variables sont mesurées à raison de deux fois par jour. Parallèlement, la dose de coagulant est déterminée par les essais Jar-Test effectués en laboratoire. Les caractéristiques statistiques des variables retenues sont présentées dans le tableau 1.

4. MÉTHODOLOGIE

La base de données utilisée a été scindée aléatoirement en deux parties, l'une pour le calage des modèles (MLP, RLM et ANFIS) et l'identification des paramètres qui représentent 80 % de la taille totale de la base de données et l'autre pour la validation (20 %). Les critères de performance sont calculés aussi bien en mode de calage qu'en mode de validation.

Dans le cas de notre étude, les variables sont de nature physique différente, caractérisées par des unités différentes, ce qui nous amène à les normaliser afin de ramener la plage d'évolution des valeurs prises par les variables à l'intérieur d'un intervalle standardisé, fixé *a priori*. Elle est souhaitable car elle évite au système de se paramétrer sur une plage de valeurs particulières, ignorant ainsi les valeurs extrêmes. Pour notre cas, nous avons normalisé les données en utilisant la formule suivante :

$$x_{n\,i,k} = \frac{x_{i,k} - m_k}{\sigma_k} \qquad (25)$$

avec :

$x_{n,\,i,\,k}$: la valeur normalisée de la variable k pour l'individu i
m_k : la moyenne de la variable k
σ_k : l'écart type de la variable k.

Une analyse en composantes principales (ACP) a été appliquée afin de déceler l'apport de chaque variable d'entrée (variable descriptive caractérisant l'eau brute) dans l'explication du phénomène étudié afin d'optimiser le nombre pertinent d'entrées pour les modèles appliqués, et de mettre en évidence d'éventuelles influences d'une variable descriptive sur le phénomène.

4.1 L'analyse en composante principale ACP

L'analyse en composantes principales (ACP) est une technique descriptive permettant d'étudier les relations qui existent entre les variables, sans tenir compte, *a priori*, d'une quelconque structure (JOLLIFE, 1986). L'objectif de l'ACP est de fournir des résumés linéaires des variables d'origine, c'est-à-dire de remplacer les variables initiales par des combinaisons linéaires de celles-ci. Ces nouvelles variables sont appelées composantes principales. Les résultats intéressants issus de l'application d'une ACP sont les coefficients de corrélation des variables initiales, associés à chaque composante principale, la matrice des vecteurs propres ainsi que les valeurs propres associés. Notons que chaque composante principale est représentative d'une portion de la variance des mesures du processus étudié. Les valeurs propres sont les mesures de cette variance et peuvent donc être utilisées dans la sélection du nombre de composantes principales à retenir. De nombreux travaux de recherche ont proposé d'utiliser l'analyse en composantes principales comme outil de modélisation des processus complexes à partir de laquelle un modèle peut être obtenu. Récemment SOUAG *et al.* (2007) ont proposé un modèle de simulation des débits mensuels en zone semi-aride basé sur l'analyse en composantes principales.

L'analyse en composantes principales (ACP) nous a permis d'obtenir une vue d'ensemble sur les données, à savoir de déterminer s'il existe des sous-populations d'individus et comment sont reliées les variables prises simultanément. Nous conservons, pour la suite de l'analyse, les composantes principales qui représentent 90 % de la variance totale. Chaque composante principale est représentée par un axe factoriel; les variables fortement corrélées avec un de ces axes contribuent

Tableau 1. **Résumé statistique des variables retenues.**
Table 1. ***Statistical summary of raw water data.***

Variables	Moyenne	Écart-type	Min	Max
Température (°C)	16,53	3,49	10,2	26,2
pH	7,76	0,25	7,23	8,6
Conductivité ($\mu s \bullet cm^{-1}$)	1 009	122	668	1 432
Turbidité (NTU)	7,58	4,54	0,44	32,4
Oxygène dissous ($mg \bullet L^{-1}$)	4,73	2,98	0,14	13,2
UV_{254} ($do \bullet m^{-1}$)	0,11	0,05	0,01	0,98
Coagulant ($mg \bullet L^{-1}$)	22,79	7,48	10	40

à la définition de cet axe. Cette corrélation correspond à la coordonnée de la variable sur l'axe factoriel correspondant. Pour l'interprétation, les variables qui nous intéressent sont celles présentant les plus fortes coordonnées en valeurs absolues (SAPORTA, 1990).

Le processus d'extraction des composantes principales se poursuit jusqu'à ce qu'il y ait autant de composantes principales que de variables. Les statistiques intéressantes issues d'une ACP sont les vecteurs de pondération des variables, associés à chaque composante principale (Tableau 2), et leur variance, λi (Tableau 2). Le portrait des pondérations des variables originelles sert à interpréter chaque composante principale alors que la variance associée indique quel pourcentage de la variance totale de l'ensemble des variables originelles chaque composante principale représente. À la lumière des résultats obtenus, on remarque qu'il est indispensable de tenir compte des cinq premières composantes principales, pour avoir 90 % de la variance totale.

5. RÉSULTATS ET DISCUSSION

5.1 Description des résultats obtenus avec l'ACP

La première valeur propre (λ1 = 2,627) représente la variance expliquée par la première composante principale (CP1). Elle correspond à 37,52 % de la variance totale (Tableau 2), et se trouve donc être l'axe prédominant. Il est expliqué par les variables pH, TU et TE. Il existe une forte corrélation entre ces variables et la première composante principale. Nous tiendrons compte par la suite de cette observation pour éviter la redondance d'information. La composante principale CP2 a la deuxième plus grande valeur propre (λ2 = 1,427). Elle représente 20,38 % de la variance totale et est construite autour de la variable OD avec un coefficient de corrélation de 0,53 en valeur absolue. Les deux composantes CP1 et CP2 expliquent 57,90 % de la variance totale. La troisième composante principale représente 15,58 % de la variance totale avec une valeur propre de (λ3 = 1,091) et est construite autour de la variable CE avec un coefficient de corrélation de 0,65 en valeur absolue. Les trois composantes CP1, CP2 et CP3 expliquent 73,48 % de la variance totale.

Étant donné que l'analyse en composantes principales n'a pas permis de réduire significativement le nombre de variables d'origine, par un plus petit nombre, une tentative de comparaison entre différentes combinaisons des variables d'entrée a été conduite. Plusieurs modèles ont été testés. Nous avons élaboré cinq variantes de modèles (Tableau 3), à savoir la variante V2 à deux variables d'entrée (15 modèles), V3 à trois variables d'entrée (18 modèles), V4 à quatre variables d'entrée (13 modèles), V5 à cinq variables d'entrée (6 modèles) et la variante V6 à six variables d'entrée (1 modèle). En totalité, 53 modèles représentant cinq variantes ont été testés et le meilleur modèle de chaque variante a été retenu (Tableau 3), à savoir le modèle M2 utilisant TE et CE; le modèle M3 avec TE , CE et pH; le modèle M4 avec TE , CE, pH et TU; le modèle M5 avec TE , CE, pH , TU et OD et le modèle M6 avec TE , CE, pH , TU, OD et l'absorbance à 254 nm (UV_{254}) comme entrées. Pour les cinq modèles cités, la dose de coagulant représente la sortie du modèle.

5.2 Description des résultats obtenus avec le modèle ANFIS

Dans le but de mettre en évidence les avantages de l'approche de modélisation neuro floue proposée, une étude

Tableau 2. Résultats de l'application de l'analyse en composantes principales.
Table 2. Results of the principal components analysis.

Matrice des valeurs propres							
λ1	λ2	λ3	λ4	λ5	λ6	λ7	Somme
2,62	1,42	1,09	0,69	0,52	0,36	0,27	7
Contribution des composantes principales							
CP1	CP2	CP3	CP4	CP5	CP6	CP7	Somme
37,52	20,38	15,58	9,91	7,42	5,27	3,9	100
Cumul des contributions des composantes principales							
37,52	57,90	73,48	83,89	90,81	96,08	99,98	100
Matrice de corrélations composantes principales-variables d'origine							
	CP1	CP2	CP3	CP4	CP5	CP6	CP7
DC	-0,348	-0,657	0,482	0,348	-0,219	-0,017	-0,216
TE	-0,749	-0,401	0,179	0,091	0,370	0,028	0,316
pH	0,777	0,009	0,020	0,381	0,396	-0,296	-0,076
CE	-0,428	-0,490	-0,654	-0,153	-0,079	-0,342	-0,043
TU	0,806	-0,295	0,121	0,047	-0,354	-0,120	0,328
OD	0,523	-0,535	-0,506	0,160	0,093	0,386	-0,030
UV_{254}	0,485	-0,470	0,357	-0,607	0,196	-0,013	-0,096

Tableau 3. **Structures des modèles testés.**
Table 3. ***Structures of the tested models***

Variantes	Nombre de Modèles testés	Modèles retenus	Variables d'entrées						Variable de sortie
			TE	CE	pH	TU	OD	UV_{254}	
V2	15	M2	I	I	O	O	O	O	DC
V3	18	M3	I	I	I	O	O	O	DC
V4	13	M4	I	I	I	I	O	O	DC
V5	6	M5	I	I	I	I	I	O	DC
V6	1	M6	I	I	I	I	I	I	DC

I : variable inclus, O: variable exclus.

comparative a été effectuée en comparant les performances obtenues avec le modèle neuro flou ANFIS et celles obtenues en utilisant un modèle à base de réseaux de neurones artificiels, le perceptron multicouches (MLP) et un modèle à base de régression linéaire multiple (RLM), respectivement. Dans le cas des modèles neuro flous de type ANFIS, utilisés dans le présent travail, le nombre total de règles floues (Tableau 4) à optimiser sera déterminé par la règle suivante :

$$NRF = NSF^{NVE} \tag{26}$$

avec : NRF représente le nombre de règles floues établies; NSF représente le nombre de valeurs linguistiques (étiquette) pour chaque variable d'entrée et NVE représente le nombre de variables d'entrée. Nous avons choisi trois valeurs linguistiques pour chaque variable d'entrée, chacune représentée par une fonction d'appartenance de type Gaussienne, et donnée par la formule suivante :

$$f(X,\sigma,c) = e^{\frac{-(x-c)^2}{2\sigma^2}} \tag{27}$$

Une fonction d'appartenance Gaussienne peut être définie par deux paramètres : σ et c. Ces deux derniers constituent les paramètres des parties prémisses à optimiser pendant la phase d'apprentissage. On remarque immédiatement que le nombre de paramètres des parties prémisses à optimiser (Tableau 4) sera déterminé par la règle suivante :

$$NPP = NSF \times NVE \times 2 \tag{28}$$

avec : NPP représente le nombre de paramètres des parties prémisses.

Les paramètres des parties conclusions (conséquents) à optimiser de leur part sont déterminés par la règle suivante :

$$NPC = NRF \times (NVE + NVS) \tag{29}$$

avec : NPC représente le nombre de paramètres des parties conclusions; NVS représente le nombre de variables de sortie (la dose de coagulant). Par ailleurs, il est à noter que plus le nombre de partitions en valeurs linguistiques augmente, plus le nombre de paramètres à optimiser augmente. Ainsi, le nombre total de paramètres à optimiser (NTP) est égal à la somme des paramètres des parties conclusions (NPC) et des parties prémisses (NPP).

$$NTP = NPP + NPC \tag{30}$$

5.3 Description des résultats obtenus avec le modèle MLP

Le deuxième type de modèle utilisé est à base de réseaux de neurones artificiels : Il s'agit du perceptron multicouche (MLP). Dans cette étude, nous avons utilisé une seule couche cachée avec une fonction d'activation sigmoïde, avec un nombre variable de neurones. Pour chaque modèle testé nous avons varié le nombre de neurones de 1 à 20, et la meilleure topologie pour chaque type de modèle a été retenue (Tableau 5). La couche de sortie contient un seul neurone avec une fonction de transfert linéaire. Mathématiquement, pour un MLP à trois couches, avec E le nombre de nœuds d'entrées, C le nombre de nœuds cachés et S le nombre de nœuds de sortie. Le nombre total de paramètres à optimiser (NTP) est déterminé par la règle suivante :

$$NTP = [E \times C] + [C] + [C \times S] + S \tag{31}$$

5.4 Description des résultats obtenus avec la régression linéaire multiple

Pour le modèle à base de régression linéaire multiple, la formule de prédiction prend la forme générale représentée par l'équation 2. La construction du modèle dans ce cas se résume à la détermination des coefficients de régression partielle (Tableau 6).

Tableau 4. **Nombre total de paramètres pour chaque modèle ANFIS testé.**
Table 4. ***Total number of parameters for each ANFIS model tested.***

Modèle	Nombre de règles floue (NRF)	Nombre de paramètres prémisses (NPP)	Nombre de paramètres conséquents (NPC)	Nombre total de paramètres (NTP)
M2	9	12	63	75
M3	27	18	189	207
M4	81	24	567	591
M5	243	30	1 701	1 731
M6	729	36	5 103	5 139

Tableau 5. **Nombre total de paramètres pour chaque modèle MLP testé.**
Table 5. ***Total number of parameters for each MLP model tested.***

Caractéristiques	Modèles				
	M2	M3	M4	M5	M6
Nombre de neurones cachés	7	19	15	7	13
Nombre de paramètres optimisés	29	96	91	50	105

Tableau 6. **Coefficients de régression pour les différents modèles testés.**
Table 6. ***Regression coefficients for each RLM model tested.***

Modèle	Coefficients de Régression						
	β_0	β_1	β_2	β_3	β_4	β_5	β_6
M2	-0,0297	0,4954	-0,0084				
M3	-0,0296	0,5001	-0,0141	-0,0219			
M4	-0,0235	0,6049	-0,1454	-0,0224	0,3380		
M5	-0,0238	0,6036	-0,1405	-0,0159	0,3435	-0,0162	
M6	-0,0239	0,6020	-0,1408	-0,0150	0,3392	-0,0167	0,0087

5.5 Comparaisons et discussions des résultats obtenus par les différents modèles

Comme nous l'avons souligné dans le paragraphe 1, le processus de coagulation met en œuvre des réactions fort complexes et non linéaires. L'objectif de notre travail est de concevoir un modèle de détermination de la dose de coagulant en tenant compte d'un nombre important de paramètres. Dans cette perspective, les réseaux de neurones et les systèmes neuro flous semble constituer une voie de recherche intéressante. Nous avons essayé de trouver un rapport adéquat entre toutes les (ou quelques-unes) variables d'entrée du modèle. Pendant toutes les phases de calcul, nous nous sommes intéressés à la comparaison des graphiques issus de la validation et du calage des différents modèles testés, ainsi qu'à la comparaison des critères numériques calculés. Pour le modèle ANFIS, nous avons utilisé trois valeurs linguistiques (étiquette) pour chaque variable d'entrée, alors que pour le modèle MLP nous avons varié le nombre de neurones dans l'unique couche cachée de 1 à 20 comme nous l'avons souligné dans le paragraphe 5.3. Après chaque essai, on compare la sortie obtenue et la sortie désirée, on corrige les poids de façon à minimiser l'erreur commise.

Nous avons procédé à une comparaison entre les résultats obtenus par les cinq modèles retenus (Tableaux 7 et 8) en mode de calage autant qu'en mode de validation, à savoir les modèles M2, M3, M4, M5 ainsi que le modèle M6 à six entrées qui incluent les six variables descriptives caractérisant l'eau brute. On remarque d'après les tableaux 7 et 8 que les résultats obtenus par la régression linéaire multiple (RLM) sont très médiocres, que ce soit en mode de calage ou en mode de validation; quel que soit le nombre de variables d'entrée utilisées, le coefficient de détermination ne dépasse pas 0,36, tandis que le RMSE avoisine les 8,15 en mode de validation pour le modèle M6 à six entrées, qui représente le meilleur modèle à base de régression linéaire multiple (RLM) (Figure 4). On remarque, d'autre part, que pour les deux modèles M2 et M3, le modèle à base de régression linéaire multiple (RLM) présente des résultats meilleurs par rapport à ceux obtenus par

Tableau 7. **Résultats des modèles en période de calage.**
Table 7. ***Model results during the calibration phase.***

Modèle	ANFIS			MLP			MLR		
	R^2	B ($mg \bullet L^{-1}$)	RMSE ($mg \bullet L^{-1}$)	R^2	B ($mg \bullet L^{-1}$)	RMSE ($mg \bullet L^{-1}$)	R^2	B ($mg \bullet L^{-1}$)	RMSE ($mg \bullet L^{-1}$)
M2	0,40	-0,04	5,35	0,53	-0,05	4,28	0,27	-0,15	6,20
M3	0,50	-0,04	4,91	0,58	-0,05	4,24	0,27	-0,15	6,21
M4	0,72	-0,06	3,64	0,64	-0,06	3,69	0,34	-0,19	5,88
M5	0,85	-0,07	2,67	0,72	-0,06	2,65	0,36	-0,20	5,80
M6	0,95	-0,04	1,89	0,80	-0,07	2,52	0,37	-0,21	5,70

Tableau 8. **Résultats des modèles en période de validation.**
Table 8. ***Model results during the validation phase.***

Modèle	ANFIS			MLP			MLR		
	R^2	B ($mg \bullet L^{-1}$)	RMSE ($mg \bullet L^{-1}$)	R^2	B ($mg \bullet L^{-1}$)	RMSE ($mg \bullet L^{-1}$)	R^2	B ($mg \bullet L^{-1}$)	RMSE ($mg \bullet L^{-1}$)
M2	0,26	-0,02	8,26	0,15	-0,01	10,31	0,25	-0,14	8,26
M3	0,40	-0,04	7,38	0,17	-0,01	9,76	0,25	-0,14	8,26
M4	0,72	-0,06	5,07	0,60	-0,05	7,82	0,32	-0,18	8,20
M5	0,90	-0,08	2,94	0,62	-0,05	7,65	0,33	-0,19	8,19
M6	0,92	-0,08	2,11	0,75	-0,07	7,34	0,35	-0,20	8,15

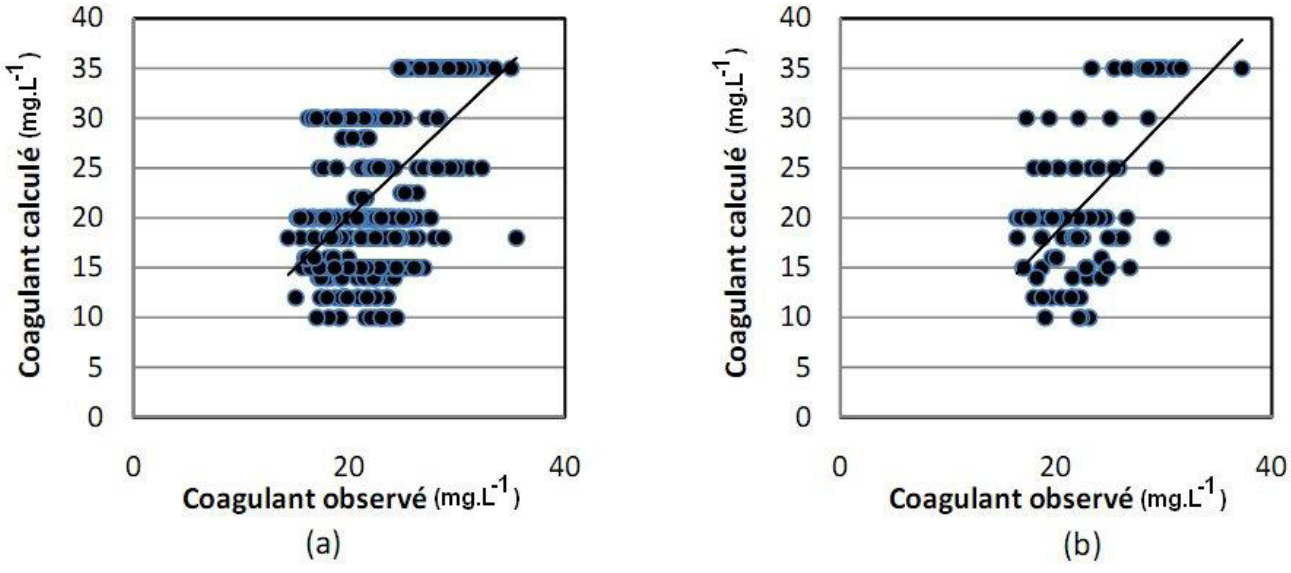

Figure 4. **Comparaison des valeurs observées et calculées pour le modèle RLM, (a) calage, (b) validation.**
Scatterplots for calculated versus observed values for the RLM model for (a) training, (b) validation.

le modèle MLP en mode de validation, avec un coefficient de détermination de l'ordre de 0,25 et une RMSE de 8,26 pour le modèle M3, alors que pour le modèle MLP, on enregistre un coefficient de détermination de l'ordre de 0,17 et un RMSE de 9,76 pour le même modèle M3. Cela reflète clairement la complexité du phénomène étudié, d'une part, et, d'autre part, ces deux modèles ne reflètent pas la réalité physique du processus de coagulation étudié. Nous verrons par la suite que ces deux modèles seront exclus et qu'il est indispensable d'intégrer plus de variables en entrée des modèles pour bien démontrer la forte non-linéarité de la relation dose de coagulant en fonction des variables descriptives de l'eau brute.

Pour les modèles MLP et ANFIS (Tableaux 7 et 8), les résultats obtenus sont nettement meilleurs par rapport à ceux obtenus par la régression linéaire multiple (RLM), pour les modèles M4, M5 et M6. Nous remarquons que le coefficient de détermination R^2 ne dépasse pas 0,35, que ce soit en mode de calage ou en mode de validation, pour les modèles à base de régression linéaire multiple. À partir du modèle M4 qui fait appel à quatre variables descriptives, nous remarquons une nette amélioration des performances. Cependant le modèle ANFIS donne des résultats meilleurs que le modèle MLP. R^2 atteint 0,72 aussi bien en mode de calage qu'en mode de validation, tandis que pour le modèle MLP, il est de l'ordre de 0,64 en mode de calage et 0,60 en mode de validation.

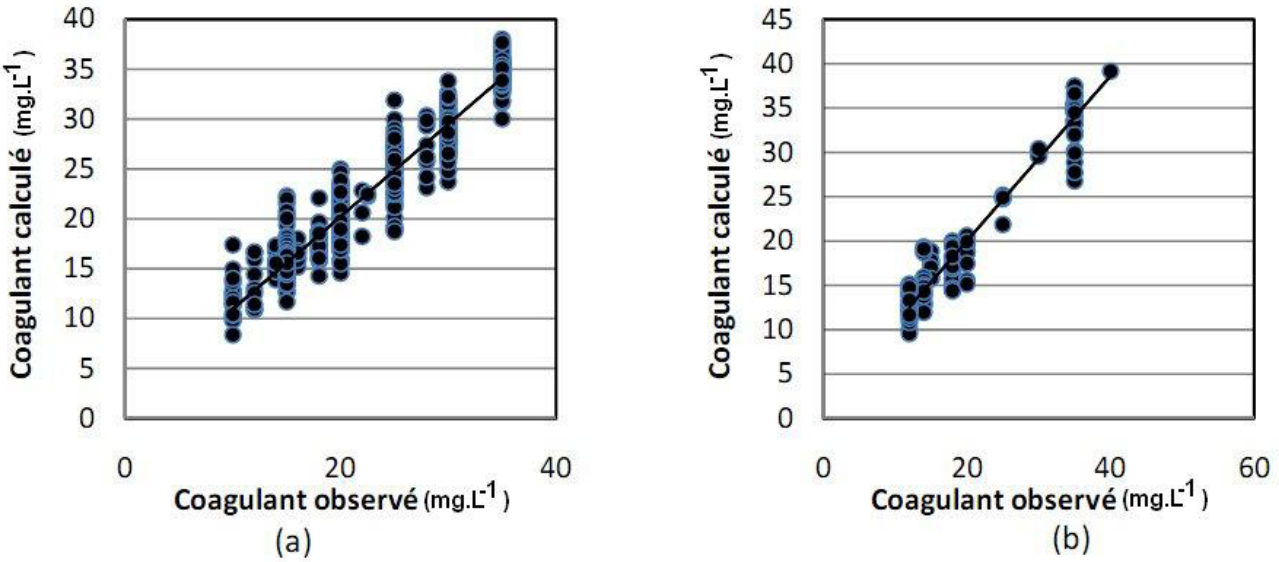

Figure 5. Comparaison des valeurs observées et calculées pour le modèle ANFIS, (a) calage, (b) validation.
Scatterplots for calculated versus observed values for the ANFIS model for (a) training, (b) validation.

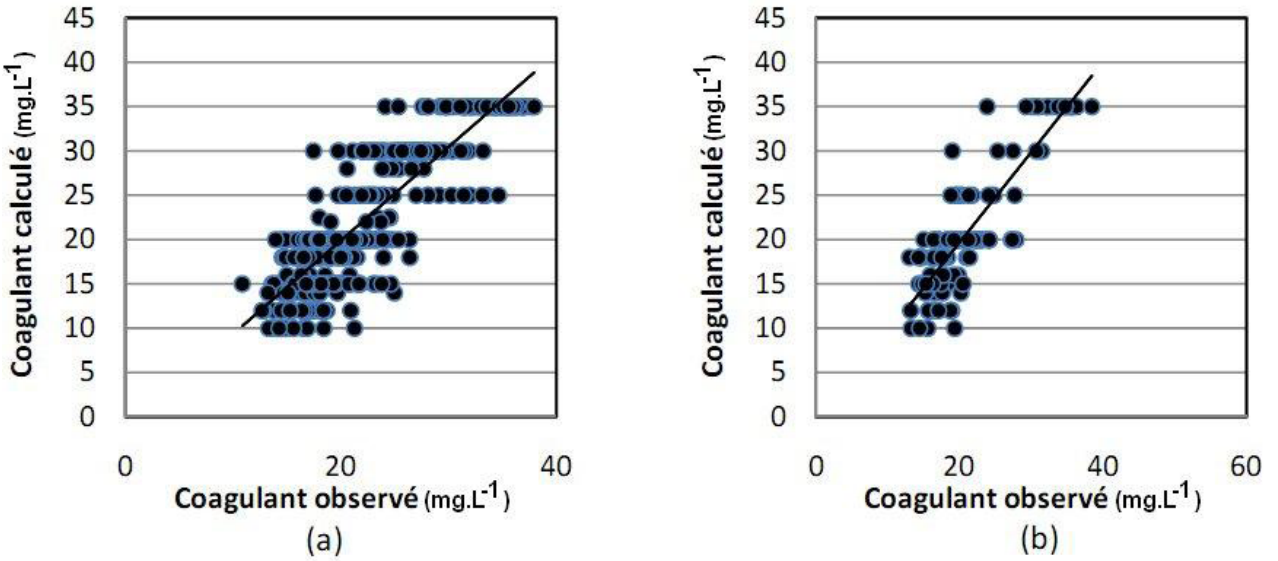

Figure 6. Comparaison des valeurs observées et calculées pour le modèle MLP, (a) calage, (b) validation.
Scatterplots for calculated versus observed values for the MLP model for (a) training, (b) validation.

Les meilleurs résultats de notre de recherche sont obtenus par le modèle M6 qui inclut les six variables descriptives. Le modèle ANFIS (Figure 5) est plus performant que le modèle MLP (Figure 6). Cela est surtout dû à la capacité des modèles flous à simuler les phénomènes fort complexes et non linéaires. Nous remarquons que le coefficient de détermination R^2 atteint 0,95 pour une RMSE de 1,89 en mode de calage alors qu'il est de l'ordre de 0,92 pour une RMSE de 2,11 en mode de validation (Tableaux 7 et 8), alors que le MLP donne un coefficient de détermination égal à 0,8 en mode de calage et 0,75 en mode de validation. Le réseau de neurones dans ce cas se compose de 13 neurones cachés avec un nombre de paramètres égal à 105 (Tableau 5**).**

À la lumière des résultats obtenus, on peut conclure que le modèle ANFIS qui inclut les six variables descriptives (M6), à savoir (la température, le PH, la conductivité, l'oxygène dissous, l'absorbance à 254 et la turbidité), est le modèle final retenu dans le cadre de ce travail. L'importance du modèle neuro flou ANFIS réside dans sa capacité à simuler des processus complexes et non linéaires en tenant compte d'un nombre important de paramètres. Il est important de rappeler que le modèle retenu se compose de plus de 5 139 paramètres avec plus de 729 règles floues (Tableau 6**).**

6. CONCLUSION

Le fruit du présent article s'est concrétisé par une contribution à la modélisation neuro floue que nous introduisons pour la première fois dans la gestion de la station de traitement des eaux de Boudouaou, considérée comme la plus importante station en Algérie. La connaissance de la variation de la qualité des eaux au niveau de cette station est importante pour comprendre et mieux interpréter le comportement des différentes composantes du processus mis en jeu.

Afin d'établir un modèle mathématique de prédiction de la dose du coagulant, nous avons proposé une comparaison entre deux modèles basés sur le concept neuronal, l'un utilisant une structure neuronale propre qui est le perceptron multicouche (MLP), et le deuxième un modèle neuro flou qui combine un système d'inférence flou dans un réseau de neurones (ANFIS), et un troisième modèle à base de régression linéaire multiple (RLM).

Les résultats obtenus par la régression linéaire multiple sont loin d'être acceptables et il est exclu d'aborder ce type de problème par une approche linéaire. Les résultats obtenus par le modèle ANFIS sont plus performants par rapport à ceux trouvés par le réseau de neurones. Les performances numériques sont plus appréciables pour le modèle utilisant six variables descriptives. Cela confirme la complexité du processus et la forte non-linéarité de la relation entre la dose du coagulant et les différentes variables descriptives.

REMERCIEMENTS

Nous remercions l'organisme qui nous a aimablement communiqué les données relatives à la qualité des eaux : *SEAAL Société Eau et Assainissement Alger*, tout particulièrement *M. ADIM Nabil* ainsi que les deux relecteurs anonymes qui, par leurs commentaires, critiques et suggestions, ont permis d'améliorer cet article.

RÉFÉRENCES BIBLIOGRAPHIQUES

ADGAR A., C.S. COX, P.R. DANIEL, A.J. BILLINGTON et A. LOWDON (1995). Experiences in the application of the artificial neural networks to water treatment plant management. Dans : *Proceedings of the International COMDEM'95*, Vol. 1, Canada, pp. 33-38.

ADGAR A., C.S. COX et T.J. BÖHME (2000). Performance improvements at surface water treatment works using ANN-based automation schemes. *Transactions Inst. Chem. Eng.*, 78, Part A, 1026-1039.

AMIRTHARAJAH A. et C.R. O'MELIA (1990). *Coagulation processes: Destabilization, mixing, and flocculation. Water quality and treatment*. PONTIUS F.W. (Éditeur), McGraw-Hill, New York, NY, États-Unis, pp.269-365.

BAXTER C.W. (1998). *Full-scale artificial neural network modelling of enhanced coagulation*. Thèse de maîtrise, University of Alberta, Edmonton, Canada, 151 p.

BAXTER C.W., S.J. STANLEY et Q. ZHANG (1999). Development of a full-scale artificial neural network model for the removal of natural organic matter by enhanced coagulation. *J. Water Supp. Res. Technol.* AQUA, 48, 129-136.

BAXTER C.W., Q. ZHANG, S.J. STANLEY, R. SHARIFF, R.R.T. TUPAS et L. STARK (2001a). Drinking water quality and treatment: the use of artificial neural networks. *Can. J. Civ. Eng.,* 28 (Suppl. S1), 26-35.

BAXTER C.W., R.R.T. TUPAS, Q. ZHANG, R. SHARIFF, S.J. STANLEY, B.M. COFFEY et K.G. GRAFF (2001b). Artificial intelligence systems for water treatment plant optimization. *Am. Water Works Ass. Res. Found*. et *Am. Water Works Ass*., Denver, CO, États-Unis, 141 p.

BAXTER C.W., R. SHARIFF, S.J. STANLEY, D.W. SMITH, Q. ZHANG et E.D. SAUMER (2002). Model-based advanced process control of coagulation. *Water Sci. Technol.*, 45, 9-17.

BAZER-BACHI A., E. PUECH-COSTE, R. BEN AIM et J.L. PROBST (1990). Mathematical modelling of optimum coagulant dose in water treatment plant. *Rev. Sci. Eau*, 3, 377-397.

BENKACI A.T. et N. DECHEMI (2004). Modélisation pluie débit journalière par des modèles conceptuels et « boîte noire »; test d'un modèle neuroflou. *J. Sci. Hydrol.*, 49, 919-930.

BÖHME T.J., C.S. COX et A. LOWDON (1999). Performance assessment of a neuro self-tuning PI controller to be used at a water treatment plant. Dans : *Proceedings of the American Control Conference*, San Diego, CA, États-Unis, pp. 3216-3220.

BUCKLEY J.J. et Y. HAYASHI (1994). Fuzzy neural networks: a survey. *Fuzzy Sets Sys.*, 66, 1-13.

CARDOT C. (1999). *Les traitements de l'eau. Procédés physico-chimiques et biologiques*. Ellipses Edition Marketing S.A., 247 p.

COX C.S., L. RIETVELD, A. ADGAR et A. VAN DER HELM (2003). A new modelling and control system simulation environment for the rapid design of potable water systems. Dans : *International Symposium on Advanced Control of Chemical Processes-Hong Kong ADCHEM*. Hong Kong, http://www.ust.hk/adchem2003/.

CRITCHLEY R.F., E.O. SMITH et P. PETTIT (1990). Automatic coagulation control at water treatment plants in the North-West region of England. Water Environ. J., 4, 535-543.

DECHEMI N., T. BENKACI et A. ISSOLAH (2003). Modélisation des débits mensuels par les modèles conceptuels et les systèmes neuro-flous. *Rev. Sci. Eau*, 16, 407-424.

DEMPSEY B.A., H. SHEU, T.M. TANZEER AHMED et J. MENTINK (1985). Polyaluminum chloride and alum coagulation of clay-fulvic acid suspensions. *J. AWWA,* 77, 74-80.

DREYFUS G., M. SAMUELIDES, J. MARTINEZ, M. GORDON, F. BADRAN, S.THIRIA et L. HERAULT (2004). *Réseaux de neurones - Méthodologies et applications. ÉDITIONS EYROLLES, Paris, France,* 200 p. ISBN : 2-212-11464-8, www.editions-eyrolles.com.

EDZWALD J.K. et J.E. TOBIASON (1999). Enhanced coagulation: US requirements and a broad review. Water Sci. Technol., 40, 67-70.

EDWARDS G.A. et A. AMIRTHARAJAH (1985). Removing color caused by humic acids. *J. AWWA*, 77, 50-57.

ELLIS G., A.G. COLLINS, X. GE et C. FORD (1991). Chemical dosing of small water utilities using regression analysis. J. Environ. Eng., 117, 308-319.

GAGNON C., B.P.A. GRANDJEAN et J.THIBAULT (1997). Modelling of coagulant dosage in a water treatment plant. *Artif. Intel. Eng.*, 11, 401-404.

GIROU A., M. FRANCESCHI, E. PUECH-COSTES et L. HUMBERT (1992). Modélisation des phénomènes de coagulation et étude de la morphologie des flocs : optimisation du taux de coagulant. *Récents Progrès en Génie des Procédés*, SFGP (Société Française de Génie des Procédés) (Éditeur), 6, 373-385, www.sfgp.asso.fr.

HEDDAM S., A. BERMAD et N. DECHEMI (2011). Applications of radial basis function and generalized regression neural networks for modelling of coagulant dosage in a drinking water treatment: A comparative study. *J. Environ. Eng.*, Doi: 10.1061/ (ASCE) EE.1943-7870.0000435.

HORNIK K., M. STINCHCOMBE et H. WHITE (1989). Multilayer feedforward networks are universal approximators. *Neural Networks,* 2, 359-366.

HORNIK K., M. STINCHCOMBE et H. WHITE (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3, 551-560.

HORNIK K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks* ,4, 251-257.

JANG J.R.S. (1993). ANFIS: adaptive-network-based fuzzy inference system. I*EEE Trans. Syst Man. Cybern.*, 23, 665-685.

JANG J.R.S., C.T. SUN et E. MIZUTANI (1997). *Neuro-fuzzy and soft computing.* Prentice-Hall, Upper Saddle River, New Jersey, NJ, États-Unis, 614 p. www.prenticehall.com. ISBN-13: 978-0132610667.

JOLLIFFE I.T. (1986). *Principal component analysis*. Springer-Verlag, New York, NY, États-Unis, 519 pp.

KISI O. (2005). Suspended sediment estimation using neuro-fuzzy and neural network approaches. *Hydrol Sci. J.*, 50, 683-696.

KRASNER S.W. et G. AMY (1995). Jar-test evaluations of enhanced coagulation. *J AWWA,* 87, 93-107.

LAMRINI B., M.V. LE LANN, A. BENHAMMOU et K. LAKHAL (2005). Detection of functional states by the "LAMDA" classification technique: application to a coagulation process in drinking water treatment. *Elsevier C.R. Phys*., 6, 1161-1168.

LEE C.C. (1990). Fuzzy logic in control systems: Fuzzy logic controller – part I and II. *IEEE Trans. Sys., Man. Cybern*., 20, 404-435.

LEFEBVRE E. et B. LEGUBE (1993). Coagulation floculation par le chlorure ferrique de quelques acides et phénols en solution aqueuse. *Water Res*., 27, 433-447.

LEKFIR A., T. BENKACI et N. DECHEMI (2006). Quantification du transport solide par la technique floue, application au barrage de Beni Amrane (Algérie). *Rev. Sci. Eau*, 19, 247-257.

LIND C. (1994a).Coagulation control and optimization: Part one. *Pub. Works*, pp. 56-57, octobre.

LIND C. (1994b).Coagulation control and optimization: Part two. *Pub. Works*, pp. 32-33, novembre.

MAIER H.R., N. MORGAN et W.K. CHRISTOPHER (2004). Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters. *Environ. Model. Software,* 19, 485-494.

MAMDANI E. (1977). Application of fuzzy logic to approximate reasoning using linguistic systems. *Fuzzy Sets Sys*., 26, 1182-1191.

McCULLOCH W. S. et W. PITTS (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5, 115-133.

MIRSEPASSI A., B. CATHERS et H. DHARMAPPA (1997). Predicted of chemical dosage in water treatment plants using ANN and Box-Jenkings models. Dans : *Preprints of 6th IAWQ Asia-Pacific Regional Conference*. Korea, 16, pp. 561-1568.

MOHTADI M.F. et P.N. RAO (1973). Effect of temperature on flocculation of aqueous dispersions. *Water Res.*, 7, 747-767.

NAHM E., S. LEE, K. WOO, B. LEE et S.SHIN (1996). Development of an optimum control software package for coagulant dosing process in water purification system. Dans : *Proceedings of the Society of Instrument and Control Engineers Annual Conference*, Tottori, Japon, 35, 1157-1161.

NAKOULA Y. (1997). *Apprentissage des modèles linguistiques flous par jeu de règles pondérées*. Thèse de Doctorat, Université de Savoie, France, 155 p.

RATNAWEERA H. et H. BLOM (1995).Optimisation of coagulant dosing control using real-time models selective to instrument errors. *Water Supp*., 13, 285-289.

RUMELHART D.E., E. HINTON et J.WILLIAMS (1986). Learning internal representation by error propagation. Dans : *Parallel Distributed Processing*. Vol.1, MIT Press, Cambridge, Massachusetts, États-Unis, pp. 318-362.

SAPORTA G. (1990). *Probabilités, analyse des données et statistique*. Éditions Technip, Paris, France, 493 p.

SEAAL. (2008). Société eau et assainissement d'Alger. Rapport interne.

SOUAG G.D., N. DECHEMI et A. BERMAD (2007). Simulation des débits mensuels en zone semi-aride par l'analyse en composantes principales et les modèles conceptuels. *Sécheresse*, 18, 97-105.

STUM W., et J.J. MORGAN (1962). Chemical aspect of coagulation. *J. AWWA*, 54, 971- 992.

TAKAGI T. et M. SUGENO (1985). Fuzzy identification of systems and its application to modeling and control. *IEEE Trans. Sys. Man Cyber*., 15, 16-132.

TUTMEZ B., Z. HATIPOGLU et U. KAYMAK (2006). Modelling electrical conductivity of groundwater using an

adaptive neuro-fuzzy inference system. *Comput. Geosci.*, 32, 421-433.

VALENTIN N. (2000). *Construction d'un capteur logiciel pour le contrôle automatique du procédé de coagulation en traitement d'eau potable*. Thèse de doctorat, UTC/Lyonnaise des Eaux/CNRS, 153 p.

VALENTIN N., T. DENOEUX et F. FOTOOHI (1999). An hybrid neural network based system for optimization of coagulation doing in a water treatment plant. Dans : *Proceedings of IJCNN99*, Washington DC., IEEE, pp. 3380-3385.

VAN LEEUWEN J., C.W. CHOW, D. BURSILL et M. DRIKAS (1999). Empirical mathematical models and artificial neural networks for determination of alum doses of southern Australian surface waters. *J. Water Sci. Res. Technol. Aqua*, 48, 115-127.

WANG L. et J.M. MENDEL (1992A). Fuzzy basis functions, universal approximation, and orthogonal least squares. *IEEE Trans. Neural Networks*, 3, 807-814.

WANG L., et J.M. MENDEL (1992B). Back-propagation fuzzy system as nonlinear dynamic system identifiers. Dans : *Proc. of the IEEE Int. Conf. on Fuzzy Systems*, San Diego, CA, États-Unis, pp. 1409-1416.

WEISHAAR J.L., G.R. AIKEN, B.A. BERGAMASCHI, M.S. FRAM, R. FUJII et K. MOPPER (2003). Evaluation of specific ultraviolet absorbance as an indicator of the chemical composition and reactivity of dissolved organic carbon. *Environ. Sci. Technol.*, 37, 4702-4708.

YU R., S. KANG, S. LIAW et M. CHEN (2000). Application of artificial neural network to control the coagulant dosing in water treatment plant. Water Sci. Technol., 42, 403-408.

ZADEH L. (1965). Fuzzy sets. *Inform. Control*, 8, 338-353.

ZADEH L. (1971). Quantitative fuzzy semantics. *Inf. Sci.*, 3, 159-176.