République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique

UNIVERSITE DE MSILA

FACULTE DE TECHNOLOGIE

DEPARTEMENT D'ELECTRONIQUE

MEMOIRE DE MASTER

DOMAINE : SCIENCES ET TECHNOLOGIE

FILIERE : GENIE ELECTRIQUE

OPTION : INSTRUMENTATION ET MAINTENANCE INDUSTRIELLE

Thème

CONTRIBUTION DE L'ANALYSE MULTIVARIEE A L'ETUDE DE LA REGRESSION ET LA CLASSIFICATION SUPERVISEE DES DONNEES ENVIRONNEMENTALES

Présenté par :

SLIM Ameur

Proposé et dirigé par :

Mr. LADJAL Mohamed

N° d'ordre : 2012/ / 85/125/

Promotion : JUIN 2012

Dédicace

Après de longues années d'études et de travail, sachant l'importance de l'aide des êtres qui n'aiment, je voudrais humblement leurs, dédier ce modeste travail tout en avant qu'ils méritent le meilleur qui soit.

Je dédie ce travail :

À ma chère mère

qui a fait tant de sacrifice pour m'élever et m'instruire et qui ma encourage tout le long de mon parcours scolaire et académique.

À mon cher père

m'encourager et suer et à tant travail pour pouvoir m'instruire.

À mon grand père

À ma grand-mère

Tout en lui souhaitant bonne santé et longue vie.

À mes chers enseignants qui mont dirigé et aidé et surtout soutenu.

À tout mes amis et camarades d'études

« la chique de la pente de l'exprès promo de l'électronique »

SLIM AMEUR

REMERCIEMENTS

Je remercie tout d'abord Dieu de m'avoir prêté santé et volonté pour mener à terme ce mémoire.

Un remerciement particulier, à l'être le plus cher dans ma vie, à celle qui m'a donné la vie, celle qui s'est sacrifiée durant de longues années, celle qui a tant donnée… sans demander en revanche…. Les mots s'épuisent maman !! mais même en remplissant des pages entières, je demeurerai ingrate à ton égard. Je te dis tout simplement que tu es la perle qui orne ma vie et que ma réussite est la tienne !!!

Ce travail a été réalisé sous la direction de mon encadreur Monsieur **Mohamed LADJAL**, Je le remercie beaucoup, qui m'a guidé, dirigé et fournit la documentation et les ressources.

Je tiens aussi à remercier, **les membres du jury** qui ont accepté de juger ce travail.

Sommaire

***CHAPITRE II* : ANALYSE MULTIVARIEE**

***CHAPITRE III* : LES RESEAUX DE NEURONES ARTIFICIELS**

***CHAPITRE* IV : SIMULATION ET EVALUATION**

Liste des figures

CHAPITRE I : Processus de contrôle des eaux brutes

CHAPITRE II : ANALYSE MULTIVARIEE

CHAPITRE III : LES RESEAUX DE NEURONES ARTIFICIELS

CHAPITRE IV : SIMULATION ET EVALUATION

Liste des tableaux

CHAPITRE IV : SIMULATION ET EVALUATION

INTRODUCTION GENERALE

L’eau est la principale composante de notre corps, elle est à l’origine de la vie :

« و جعلنا من الماء كل شيء حي ، الآية »

Et Comme dirait le proverbe anglais « ***Nous ne connaissons la valeur de l'eau que lorsque le puits est à sec*** ».

L’eau fait partie de notre environnement naturel tout comme l’air que nous respirons. Elle constitue un des éléments familiers de notre vie quotidienne. Sans eau l’homme on ne peut survivre. Il en est de même pour tous les être vivants. Les aliments (ou toutes substances nutritives) déshydrates ne peuvent permettre sans apports complémentaires d’eau, ni le développement ni la reproduction des être vivants. La consommation en eau par habitant est désormais considérée comme un indicateur du développement économique d'un pays. Selon une étude des Nations Unies, l'eau pourrait même devenir, d'ici à 50 ans, un bien plus précieux que le pétrole. C’est dire toute l’importance de cette ressource que d’aucuns appellent déjà « l’or bleu » [1].

Avec toute cette importance de l’eau, il est donc nécessaire de réfléchir à des solutions pour la contrôler et la surveiller dont le but d’avoir une eau avec une plus grande qualité à un coût de production plus faible. Ces considérations motivent les efforts importants dans le développement de nouvelles méthodes et techniques de contrôle et de surveillance.

La surveillance de potabilité de l’eau peut être mise en pratique selon deux types de méthodes classique et moderne. Les méthodes classiques sont déterminées par une analyse chimique effectuée au laboratoire, cette méthode nécessite un temps d’analyse relativement important et peut donc être difficilement intégrée dans un système de surveillance et de diagnostic en temps réel de l’unité. En outre les méthodes modernes qui ont l’avantage de pouvoir effectuer un contrôle automatique permanent en temps réel, sont mieux placées pour être une alternative plus efficace. Une surveillance permanente des divers procédés de traitement et des paramètres relatifs à la qualité de l’eau est devenue nécessaire, où des systèmes de contrôle automatique infaillibles sont impératifs.

Durant ces dernières décennies, des efforts importants ont été réalisés dans le développement des méthodes de contrôle et de surveillance automatique des eaux potables ; Ces méthodes dites de haut niveau disposent d'outils qui sont plutôt orientés vers la communication avec un opérateur expert. Celles-ci représentent les techniques de l'intelligence artificielle (*IA*) qui servent comme outil de base pour l'aide à la décision. Leur réponse est plus élaborée et peut être obtenue soit à partir de données brutes venant directement des variables de surveillance ; soit, à partir de données traitées venant des sorties de traitements de bas niveau.

Il est logique de supposer que le problème de contrôle et de surveillance de l'eau brute peut être vu comme un problème de reconnaissance de formes, où les formes représentent l'ensemble des observations ou mesures liées aux caractéristiques de celle-ci. Parmi les techniques d'IA connues dans ce domaine, on trouve "les réseaux de neurones artificiels". Cette technique se démarque des autres outils par leur capacité d'apprentissage et de généralisation. Les différents paramètres physico-chimiques exploités dans le traitement de l'eau, tels que le pH, la température, l'oxygène dissous, la conductivité, et les matières en suspension,….etc, sont transformés en signaux électriques à partir d'une fusion de données multi- sensorielle et transmis vers une station de contrôle qui assure l'acquisition et le traitement des données. La technique devant être utilisée au niveau du système de décision doit pouvoir effectuer un contrôle quasi permanent de cette ressource précieuse. Cette technique peut aussi être utilisée dans la réalisation de capteurs logiciels ou les paramètres sont non mesurables en ligne (des essais de laboratoire tels que Ca, Mg, Cl,…). Pour des raisons diverses, tels que : le coût, non mesurable en continu,…. Ce type de capteurs peut être d'un grand intérêt aussi bien sur le plan économique que technique. De plus, le contrôle et la surveillance de eaux (*classification de données*), et par conséquent, le développement de capteurs (*régression de données*) a été précédé d'une analyse statistique (exemple : *Analyse en Composantes Principales*), permettant de déterminer les corrélations existantes entre les variables caractéristiques de l'eau brute puis de ne conserver que les caractéristiques apportant réellement une information pertinente. C'est qu'on peut utiliser ces sorties comme des variables d'entrée dans un autre système de surveillance[2].

Notre mémoire sera par conséquent réparti en quatre chapitres :

Dans le premier chapitre nous allons parler du processus de traitement des eaux brutes, en l'introduisant par des généralités sur leur traitement, leur cycle, les ressources naturelles,

les types de contrôles ainsi que les normes à appliquer. Suivent les différentes étapes de la chaîne de traitement des eaux potables. Nous terminerons par les paramètres descripteurs de l'eau ainsi que leurs capteurs correspondants ; les différentes méthodes et techniques de surveillance classiques et modernes y sont décrites.

Dans le deuxième chapitre nous allons aborder L'analyse multivarié des données d'une façon générale, son objectif ainsi que les différents types. Dans ce contexte, nous allons choisir l'une des méthodes de réduction de dimension « *Analyse en composantes principales,* ACP» , après un bref historique en partant du principe, en passant par les différentes étapes, l'algorithme, ainsi que l'application de l' ACP pour la régression et la classification de données.

Le troisième chapitre est dédié à une présentation des réseaux de neurones artificiels. On commencera ce chapitre par une présentation de la structure d'un neurone biologique et un neurone formel, puis les architectures et l'apprentissage des différents réseaux existants. Quelques modèles de réseaux de neurones et particulièrement le perceptron multicouches (PMCs), et le réseau de type RBF (*Radial Basis Functions).*

Dans le chapitre quatre nous présenterons la simulation de la technique d'analyse statistique des données ACP utilisée pour la réduction de dimension comme une étape préliminaire pour l'extraction des paramètres d'entrée des capteurs logiciels (*régression*) et classification (*binaire*) formé par apprentissage utilisant les réseaux de neurone de type RBF appliquée dans le domaine de l'eau. Ce dernier chapitre décrit la mise en œuvre de la méthode « *Analyse en composantes principales,* ACP» pour l'appliquée dans la reconnaissance de formes. L'application concerne le contrôle et la surveillance de l'eau. En effet, nous avons fourni des résultats relatifs à la méthode de réduction de dimension ACP. Une étude en simulation est effectuée pour valider et évaluer les performances de ces méthodes, permettant la validation de notre proposition. Une conclusion générale donnera une synthèse du travail effectué, et résumera les principaux résultats obtenus, ainsi que les perspectives envisagées pour d'éventuelles améliorations.

CHAPITRE I

Processus de contrôle des eaux brutes

Introduction

L'industrie de l'eau est sous une pression croissante pour produire une eau potable d'une plus grande qualité au plus faible coût. A l'heure actuelle, le contrôle de l'eau potable est effectué au quotidien par analyse hors ligne au laboratoire, à l'aide d'un essai expérimental appelé « Jar-test ». Cet essai consiste à mettre des doses croissantes de coagulant dans des récipients contenant la même eau brute. Après quelques instants, on procède sur l'eau décantée à toutes les mesures utiles de qualité de l'eau. La dose optimale est donc déterminée en fonction de la qualité des différentes eaux comparées. On voit ici tout l'intérêt de disposer d'un moyen de contrôle automatique pour effecteur cette détermination.

Objet de ce chapitre, consiste à l'introduction au domaine de contrôle de l'eau potable. Nous allons d'abord parler de l'eau d'une manière générale, plus particulièrement de l'eau potable et de ses sources, les différentes étapes de traitement des eaux de surface sont décrites, ainsi que les différents types de contrôle de qualité de l'eau avec une comparaison entre ces méthodes, et les méthodes de mesure des principaux paramètres de potabilité. Enfin, nous terminons ce chapitre par l'exposition de notre problématique d'application.

1.1. Composition de l'eau

L'eau est une molécule composée de trois atomes : deux hydrogène et un oxygène, son symbole chimique est H_2O on retrouve l'eu sur la terre sous forme : liquide, solide (glace) ou gaz[3].

1.2. Définition de l'eau potable

Plusieurs spécialistes ont définie l'eau par différents manières, mais en général l'eau potable est très malaisée. C'est en effet un terme générique qui ne peut s'appuyer sur un type unique, car toute eau que l'on peut consommer sans danger peut être considérée comme potable. A cette notion de danger potentiel peut se superposer une notion d'agrément vis-à-vis du goût et même de confort (aspect, température). Pour cela, plusieurs spécialistes ont définit l'eau comme suit :

- Une eau potable est une eau devant satisfaire à un certain nombre de caractéristiques la rendant propre à la consommation humaine.
- Eau propre à la consommation, signifiant qu'elle ne contient pas de micro-organismes ou autres substances nocives.
- On dit qu'une eau est potable lorsque sa consommation n'a pas de danger pour la santé humaine.
- Une eau potable est une eau que l'on peut boire sans risque pour la santé. Pour être consommable, l'eau doit être traitée afin d'éliminer les substances inertes ou vivantes qui peuvent être nocives pour l'organisme. Des normes sont d'ailleurs établies afin de fixer les teneurs limites.
- L'eau qui est fournie par le réseau de distribution doit être conforme aux normes de potabilités (limites), de qualité fixée par la réglementation. Lorsque la limite de qualité est dépassée, l'eau est déclarée **non potable**.[2]

1.3. Cycle de l'eau

La formation de la pluie résulte globalement de la condensation de l'eau contenue dans l'air, après évaporation de celle-ci à partir des mers des lacs et même des sols sous l'effet du rayonnement solaire ; En outre, l'air contient aussi des particules et des gaz d'origine naturelle et d'origine humaine qui se dispersent, circulent dans l'atmosphère et vont se redéposer au sol, soit par temps sec, soit par temps humide. Au sol, l'eau de pluie s'évapore, s'infiltre ou ruisselle et rejoint les cours d'eau. L'eau circule dans en perpétuel mouvement entre ciel et terre. C'est ainsi qu'au contact de l'eau, les gaz se transforment en acides, la pluie va donc naturellement se charger de particules et d'acides. Il y a donc un lien naturel entre pollution atmosphérique et pollution des eaux pluviales [3][4].

De toutes les manières, les eaux utilisées pour l'alimentation humaine sont rarement consommables telles quelles, et ce, quelque soit leur origines souterraine ou superficielle, il est souvent nécessaire de leur appliquer un traitement plus ou moins complexe, ne serait-ce qu'une désinfection dans le cas des eaux souterraines.

De la nature de cette pollution « naturelle » conjuguée à la pollution causée par des rejets de certaines activités humaines, dépendra le procédé de traitement de l'eau à mettre en œuvre afin de la rendre potable sans risque pour la santé humaine. Le traitement nécessaire dépend donc manifestement de la qualité de l'origine de l'eau brute à traiter pouvant évoluer dans le temps, il varie aussi avec le niveau d'exigence et les normes à appliquer[5].

Quant au cycle de l'eau destinée à la consommation, celle-ci est prélevée dans des cours d'eau, des nappes d'eaux souterraines ou pompée directement de la mer dans le cas du dessalement d'eau de mer. Elle est ensuite acheminée vers une usine de production d'eau potable où elle subit divers traitements physiques, chimiques et biologiques. Rendue potable, elle est distribuée aux consommateurs, recueillie ensuite après usage pour être conduite vers les usines de dépollution des eaux usées, avant d'être enfin rendue à la nature.

1.4. Les normes à appliquer

L'eau de consommation humaine est l'aliment le plus surveillé, le niveau d'exigence pour sa qualité est très élevé. Pour une eau potable, la notion de qualité distingue la qualité des eaux brutes puisées au niveau de leurs lieux de captage de la qualité de l'eau distribuée et livrée au robinet du consommateur, après traitement de potabilisation et parcours dans les canalisations.

A travers l'édition et la diffusion de directives, l'Organisation Mondiale de la Santé définit un cadre pour la sécurité sanitaire et la qualité requise qui garantit une eau saine et donc potable, et ce, en fixant des valeurs guides [6].

Sur cette base, un pays ou un groupe de pays transposent ces directives en droit propre et intègrent ces recommandations à leurs législations nationales qu'ils appliquent. Par conséquent, les fournisseurs d'eau potable doivent veiller à respecter trois niveaux d'exigences relatives à la qualité de l'eau potable :

- Respect des valeurs guides à ne pas dépasser (celles prescrites par l'OMS).

- Respect des limites de qualité de l'eau au robinet du consommateur.
- Respect des normes de qualité liées au fonctionnement propre de la station de traitement d'eau concernée.

1.4.1. Les normes bactériologiques

En Algérie, le décret exécutif n°11-125 du 22/03/2011 relatif à la qualité de l'eau de consommation humaine définie dans la loi N°05-12 du 04/08/2005 , fixe les **valeurs limites** des paramètres de qualité de l'eau à ne pas dépasser, ces valeurs sont maximales pour certains paramètres chimiques, radionucléides et microbiologiques et dont le dépassement constitue un danger potentiel pour la santé des personnes.

Paramètres avec valeurs limites :

- Les paramètres chimiques au nombre de 49 (constitués essentiellement de métaux, métaux lourds, d'hydrocarbures, de pesticides, de chlore, dégraissants chimiques).
- Les Radionucléides au nombre de 05, émettant un rayonnement ionisant
- Les paramètres microbiologiques au nombre de 03 (Escherichia coli, Entérocoques, Bactéries sulfito-réductrices) indicateurs de présence de contamination fécale, ne doivent pas être détectés dans l'eau de consommation humaine [7].

1.4.2. Les normes physico-chimiques

De même que pour les paramètres chimiques, les valeurs de référence sont données à titre indicatif pour certains paramètres organoleptiques et physico-chimiques par le même décret n°11-125 aux fins de contrôle du fonctionnement des installations de production de traitement et de distribution d'eau et d'évaluation des risques pour la santé des personnes.

Par conséquent, avant d'établir un programme de surveillance régulière, il est impératif que les responsables de la gestion de la qualité de l'eau examinent toute situation préoccupante existante et potentielle au site en regard de la qualité microbiologique de l'eau (le niveau de traitement, les résultats d'analyses antérieures, la vulnérabilité des eaux de surface et souterraines, les risques potentiels relatifs à la qualité de l'eau, les activités des usagers et les tendances locales).

Paramètres avec valeurs indicatives en Algérie :

- Les paramètres organoleptiques au nombre de 04 (couleur 15 mg/l, Turbidité 5 NTU, Odeur à 12°C, saveur à 25°C à des taux de dilution à 4) .
- Paramètres physico-chimiques au nombre de 11 (Alcalinité de 500 mg/l en CaCo3 ;

Calcium à 200 mg/l en CaCo3 ; Chlorures à 500mg/l ; 6,5 ≤ pH ≤ 9 ; Conductivité 2800 µS/cm à 20°C Dureté 200mg/l en CaCo3 ; Température à 25°C…).

En règle générale, les opérateurs d'installations de traitement doivent prélever au moins un échantillon aux fins d'analyses par mois. La fréquence d'échantillonnage, le nombre d'échantillons et les analyses à effectuer doivent être déterminés d'après les résultats de l'enquête sanitaire et de l'évaluation de la vulnérabilité. S'il ressort qu'il n'y a aucun problème, le nombre d'échantillons à prélever peut être moins élevé. Par contre, si le niveau de risque est plus élevé ou s'il y a des variables inconnues, il est recommandé de modifier le plan d'échantillonnage en conséquence de façon à recueillir suffisamment de données pour mettre en œuvre des mesures correctives. [7].

1.4.3. Les normes relatives au traitement

Les recommandations pour la qualité de l'eau potable contiennent une liste de nombreux paramètres chimiques et physiques dont la présence dans l'eau potable est préoccupante. C'est pourquoi, il est recommandé que les gestionnaires de l'eau et le personnel technique effectuent une analyse chimique de base de leurs sources d'approvisionnement afin de déterminer les substances qui devraient être surveillées dans le cadre du programme de surveillance dans une unité de traitement. Ils doivent adapter leur protocole de surveillance de qualité à l'interne, assorti d'un calendrier adapté aux conditions locales selon l'origine de l'eau brute à traiter (eaux superficielles constituées de lacs, oueds, sebkhas, eaux souterraines ou eaux de mer à dessaler) [1].

L'eau des réseaux de distribution qui alimentent les installations doit être salubre du point de vue microbiologique en plus du respect des paramètres physico-chimiques imposés par la législation.

Les systèmes de traitement doivent être conçus et adaptés en fonction de la qualité de l'eau brute et de sa quantité ainsi que des variations saisonnières. Les caractéristiques de l'eau traitée dépendront des procédés utilisés, des composantes du système de traitement, de la conception de l'équipement, des produits chimiques employés, de l'efficacité du traitement et des procédures de contrôle, etc.

La batterie d'analyses relatives au contrôle pour la surveillance de l'eau traitée doit être adaptée à toutes les étapes du processus de traitement de potabilisation (y compris l'eau de boisson transportée dans des camions citernes) [8]. La finalité étant le respect des valeurs de paramètres fixées par les normes en vigueur.

1.5. Importance de l'analyse et du traitement

Une analyse régulière de l'eau est importante pour les raisons suivantes :

- Elle permet de définir les problèmes existants.
- Elle garantit une eau qui convient à l'utilisation prévue.
- Elle garantit une eau potable sûre.
- Elle permet de vérifier l'efficacité du système de traitement.

La qualité d'une réserve d'eau peut changer au fil du temps et même subitement. Si l'apparence, l'odeur et le goût de l'eau restent les mêmes, le changement de qualité risque de passer inaperçu. La seule façon de connaître la salubrité de l'eau potable, est de la faire analyser. Comme les bactéries, les parasites et les virus nuisibles sont invisibles à l'œil nu, une eau au goût et à l'apparence agréables n'est pas forcément potable. Ces microbes, qui vivent parfois dans l'eau souterraine et de surface, risquent de causer rapidement des maladies chez les humains qui consomment l'eau sans la traiter adéquatement. Certains contaminants chimiques que l'on retrouve dans les réserves d'eau peuvent causer des problèmes de santé à long terme, qui n'apparaissent que des années après la consommation. Une analyse fréquente de l'eau permet de déterminer le niveau de salubrité de l'eau et de vérifier si le système de traitement a atteint un degré de purification satisfaisant. Plusieurs analyses disponibles sont utiles pour déterminer la salubrité et la sûreté des réserves d'eau. L'analyse de base de l'eau potable comprend plusieurs aspects d'analyse tels que celui des bactéries coliformes, des nitrates, du pH, du sodium, du chlorure, du fluorure, des sulfates, du fer, du manganèse, des matières totales dissoutes et celui de la dureté [8].

Si on soupçonne la présence d'un contaminant particulier dans l'eau, on peut procéder à d'autres analyses. On analyse parfois l'eau souterraine afin d'y détecter la présence d'arsenic, de sélénium ou d'uranium, par exemple. On peut aussi évaluer la contamination de l'eau de surface ou souterraine par les pesticides. Les réserves d'eau domestique doivent faire l'objet d'une analyse au moins une fois par an. L'eau potable provenant de puits peu profonds ou de réserves de surface, plus sujette à la contamination que l'eau souterraine ; doit être analysée plus souvent (chaque saison). Il est important d'analyser l'eau potable au robinet et à la source. Ces deux analyses permettent de vérifier l'efficacité du système de traitement et de détecter tout changement dans la qualité de l'eau à la source. Il est important de souligner que l'eau avant qu'elle parvienne au consommateur, subi des traitements plus ou moins poussés, elle est stockée, acheminée, puis distribuée. L'eau potable est donc une denrée rare et

précieuse qui a un coût, qu'il ne faut pas gaspiller. Par ailleurs, il faut garder à l'esprit qu'elle est produite à partir de ressources naturelles qu'il convient de protéger.

1.6. Processus de traitement des eaux

Le traitement d'une eau brute dépend de sa qualité, laquelle est fonction de son origine et peut varier dans le temps. L'eau à traiter doit donc être en permanence analysée car il est primordial d'ajuster le traitement d'une eau à sa composition et, si nécessaire, le moduler dans le temps en fonction de la variation observée de ses divers composants. Il peut arriver cependant qu'une pollution subite ou trop importante oblige l'usine à s'arrêter momentanément.

L'objectif principal d'une station de traitement d'eau brute est de fournir de l'eau potable qui réponde à des exigences édictées par les normes de qualité, et ce, d'une manière constante, à un prix raisonnable pour le consommateur. Le traitement classique est complet d'une eau s'effectue en plusieurs étapes dont certaines ne sont pas nécessaires aux eaux les plus propres. Pour cela, il est présenté dans un premier temps le procédé de prétraitement permettant d'éliminer l'ensemble des éléments de nature à perturber les traitements ultérieurs. Nous évoquerons le procédé de pré-oxydation constituant la première barrière face aux matières organiques en solution. Ensuite, nous aborderons la clarification qui représente l'étape la plus importante de la chaîne de traitement et que l'on va retrouver dans la majorité des usines de traitement d'eau potable.

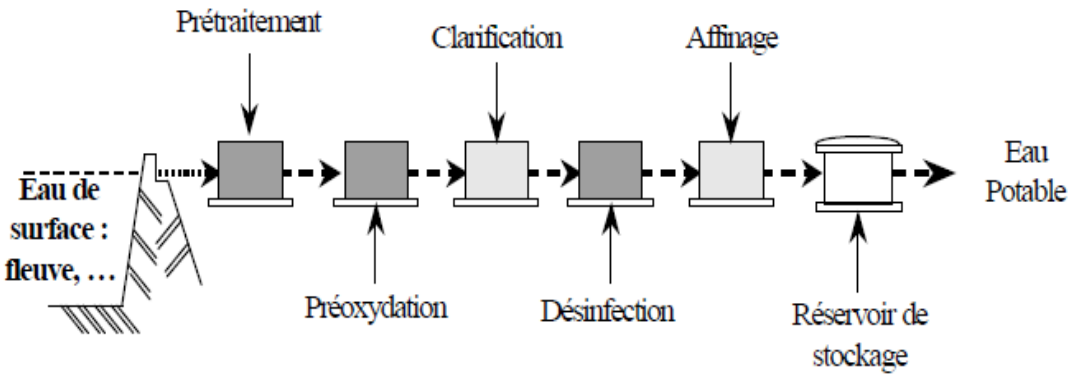


Fig. 1.1. Processus de Traitement de potabilisation de l'eau.

1.6.1. Prétraitement (dégrillage et tamisage)

Une eau, avant d'être traitée, doit être débarrassée de la plus grande quantité possible d'éléments dont la nature et les dimensions constitueraient une gêne pour les traitements ultérieurs. Pour cela, on effectue des prétraitements de l'eau de surface. Dans le cas d'une eau potable, les prétraitements sont principalement de deux types: le dégrillage et le tamisage.

Le dégrillage, premier poste de traitement, permet de protéger les ouvrages avals de l'arrivée en amont de gros objets susceptibles de provoquer des bouchages dans les différentes unités de traitement. Ceci permet également de séparer et d'évacuer facilement les matières volumineuses charriées par l'eau brute, qui pourraient nuire à l'efficacité des traitements suivants, ou en compliquer l'exécution. Le dégrillage est avant tout destiné à l'élimination de gros objets morceaux de bois, etc. Le tamisage, quant à lui, permet d'éliminer des objets plus fins que ceux éliminés par le dégrillage. Il s'agit de feuilles ou de morceaux de plastique par exemple [9] .

1.6.2. L'oxydation ou pré-oxydation

A l'issue de l'opération prétraitement précédemment décrite, l'eau est relativement propre, mais qui contient encore des particules colloïdales en suspension et des matières organiques en solution ; celles-ci n'ont en elles-mêmes rien de dangereux puisque consommées au quotidien dans les boisons chaudes qui sont chargées de matières organiques. Sachant qu'elles s'oxydent d'une manière spontanée au contact de l'air. Par conséquent leur élimination préalable est une opération nécessaire ; Elle permet d'éliminer plus facilement ces substances au cours de l'étape suivante dite de clarification. On utilise pour cela un oxydant comme le chlore ou pré-chloration, le dioxyde de chlore ou l'ozone ou pré-ozonation. La pré-chloration, effectuée avant le procédé de clarification, s'est surtout développée dans les années 60, elle tend à disparaître actuellement. Le chlore est le réactif le plus économique, mais il a comme inconvénient de former avec certains micro- polluants des composés organochlorés dont le goût et l'odeur sont désagréables [1].

Enfin, depuis quinze à vingt ans, on utilise comme pré-oxydant l'ozone, qui non seulement a l'avantage de détruire les matières organiques en cassant les chaînes moléculaires existantes, mais également a une propriété virulicide très intéressante, propriété que n'a pas le chlore.

1.6.3. La Clarification

La clarification est l'ensemble des opérations permettant de rendre l'eau limpide et d'éliminer les matières en suspension (MES) d'une eau brute ainsi que la majeure partie des matières organiques. La clarification comprend les opérations de coagulation, de floculation et de filtration.

La coagulation est l'une des étapes les plus importantes dans le traitement des eaux de surface. 90 % des usines de production d'eau potable sont concernées. La difficulté principale est de déterminer la quantité optimale de réactif à injecter en fonction des caractéristiques de l'eau brute.

L'opération de clarification (Floculation/coagulation et décantation) : est réalisée par ajout d'un produit chimique (chlorure de fer ou sulfate d'aluminium) à l'eau qui provoque le regroupement (agglomération) des particules encore présentes (poussières, particules de terre, etc.) en flocons. Ceux-ci s'agglomèrent et se déposent au fond du bassin par décantation. Ainsi 90 % des matières en suspension (MES) sont éliminées.

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. Cette opération a également une grande influence sur les opérations de décantation et de filtration ultérieures. En revanche, un contrôle efficace peut réduire les coûts de main d'œuvre et de réactifs et améliorer la conformité de la qualité de l'eau traitée.

En résumé, le contrôle de cette opération est donc essentiel pour trois raisons : la maîtrise de la qualité de l'eau traitée en sortie (abattement de la turbidité), le contrôle du coagulant résiduel en sortie (réglementation de plus en plus stricte relative à la présence de coagulant résiduel dans l'eau traitée) et la diminution des contraintes et des coûts de fonctionnement (coûts des réactifs et des interventions humaines) [9] .

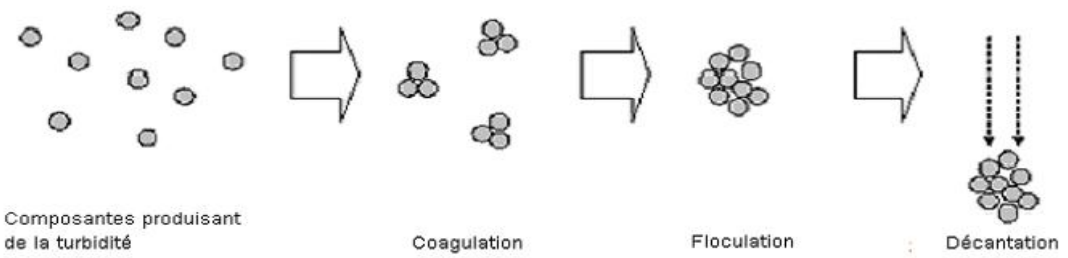


Fig.1.2 . Processus de coagulation, floculation et décantation.

1.6.4. Décantation

C'est grâce au coagulant qu'il y a agglomération (coagulation) des fines matières en suspension. L'eau est par la suite acheminée vers la cloche à vide du décanteur. C'est alors que tous les flocons coagulés sont poussés par le bas du décanteur, sont arrêtés par le lit de boue fluidisée.

Le phénomène de décantation est un procédé de séparation par gravité des matières solides sédiment sables aboutissant à l'élimination de solides en suspension de densité supérieure à celle de l'eau par l'action de la gravité et qui se déposent dans les décanteurs.

Ce sont des bassins dans lesquels les flocs (flocons) de particules qui décantent sont retenus et accumulés au fond des bacs. Environ 66% des matières vont se déposer dans ces bassins. Cette boue ainsi formée sera ensuite extraite périodiquement (flocons décantés) et acheminée vers l'usine d'épuration. Par la suite, pour finaliser sa clarification, l'eau est acheminée vers la chaîne de filtration.

1.6.5. Filtration

Pour éliminer les 10 % de MES restantes, l'eau traverse un filtre, lit de sable fin et/ou un filtre à charbon actif. La filtration sur sable élimine les matières encore visibles à l'œil nu. Les filtres à charbon actif retiennent en plus les micro- polluants, comme les pesticides et leurs sous- produits, les composés à l'origine des goûts et des odeurs (cette filtration peut avoir lieu après la désinfection également car ils retiennent également des sous-produits de désinfection) Cependant, il y a lieu de signaler qu'il existe des procédés de filtration encore plus poussés comme la filtration sur membranes.

1.6.6. Désinfection

Les eaux à traiter contiennent beaucoup de matières organiques ou encore de l'ammoniaque, du fer ou du manganèse. L'oxydation qui est la dernière étape, est donc nécessaire afin d'éliminer les micro-organismes qui pourraient être dangereux pour notre santé. On utilise pour cela des oxydants qui sont le chlore et l'ozone. Le chlore possède de remarquables propriétés désinfectantes et est également un oxydant puissant. Il est moins coûteux et d'un emploi plus simple que l'ozone. Son effet est, en outre, plus durable.

Le chlore est utilisé à trois reprises au cours du processus de traitement de l'eau. Il permet d'éliminer plus facilement les substances nuisibles au cours des étapes dites de clarification, filtration et distribution.

a). Le chlore (Chloration)

Du chlore est ajouté à la sortie de l'usine de production et sur différents points du réseau de distribution afin d'éviter le développement de bactéries et de maintenir la qualité de l'eau tout au long de son parcours dans les canalisations

b). L'ozone

L'eau est désinfectée grâce à l'ozone, qui a une action bactéricide et antivirus. Ce gaz, mélangé à l'eau, agit aussi sur les matières organiques en les cassant en morceaux. Il améliore également la couleur et la saveur de l'eau. L'ozone est de loin le meilleur oxydant pour le contrôle des goûts et des odeurs ou des couleurs provenant de la dissolution de substances humiques. Il peut s'avérer efficace pour oxyder partiellement la matière organique de l'eau, produisant ainsi des composés biodégradables que la bio-filtration est ensuite en mesure d'éliminer. De plus, son utilisation n'entraîne aucune nouvelle odeur.

Ses propriétés sont bactéricides et oxydantes. Ce gaz est fabriqué sur place avec un appareil nommé « ozonateur ». Par contre, l'ozone ne sert que de désinfectant primaire puisqu'il ne peut demeurer en concentration résiduelle suffisante dans le réseau de distribution, comme le chlore. Il faut donc associer au traitement à l'ozone une désinfection secondaire à l'aide de chlore.

L'ozonation ou la pré-ozonation de la plupart des eaux brutes permet de réduire la demande subséquente de l'eau filtrée en chlore. C'est à la sortie des filtres, dans une chambre de contact et à l'aide de diffuseurs installés au fond, que l'on mélange l'ozone à l'eau filtrée.

c). Le rayonnement UV

L'irradiation par une dose suffisante de rayonnement UV permet la destruction des bactéries, virus, germes, levures, champignons, algues, etc. Les rayonnements UV ont la propriété d'agir directement sur les chaînes d'ADN des cellules et d'interrompe le processus de vie et de reproduction des micro-organismes. Comme pour l'ozone, elle n'est pas caractérisée par un effet rémanent. Pour chaque type de traitement, il est nécessaire de

contrôler divers paramètres physico-chimiques. Ceux-ci doivent permettre d'évaluer l'efficacité de la désinfection [1].

1.7. Surveillance des eaux potables

La surveillance permanente des processus de traitement implique la mesure en continu d'un certain nombre de paramètres à l'aide des capteurs. Ceux-ci peuvent se classer en deux grandes familles : les paramètres usuels et spécifiques de l'eau [1].

1.7.1. Mesure des paramètres usuels

Les paramètres usuels sont principalement les débits, les niveaux de liquides ou de solides, les pressions, et les températures. Dans toute installation de traitement d'eau la connaissance du débit est impérative.

1.7.2. Mesure des paramètres spécifiques

Dans les appareils utilisés pour la mesure des paramètres spécifiques de l'eau, les différentes méthodes d'analyse sont mises en œuvre de façon automatique, en particulier : la néphélométrie (mesure de turbidité), la mesure de résistivité ou de conductivité, la potentiométrie (mesure de pH), l'ampérométrie (mesure de concentration en agent oxydant, chlore, ozone), la photocolorimétrie et la titrimétrie (mesure de la concentration de certaines substances dissoutes dans l'eau) [1]. On peut classer ces différents dispositifs en deux grandes catégories : celle des capteurs physiques et celle des analyseurs chimiques qui réalisent préalablement à toute mesure, une ou plusieurs réactions chimiques.

1.7.2.1. Capteurs physiques

- Mesure de la turbidité : La mesure de turbidité de l'eau correspond à une mesure optique des particules en suspension dans l'eau qui lui donnent un aspect trouble. L'unité employée est appelée unité néphélométrique de turbidité (NTU). Les particules sont d'origines diverses : Argiles, limons, organismes microscopiques, dépôts dans les canalisations, corrosion... Les risques sanitaires peuvent être liés à la présence de ces particules car elles permettent aux bactéries et aux virus de se fixer et d'être ainsi protégés de l'action des désinfectants. L'amélioration peut être obtenue par filtration ou coagulation. La prise d'eau s'effectuant dans une nappe peu profonde, donc sur une eau peu filtrée naturellement, la turbidité de l'eau peut

varier selon les pluies (et aussi selon les travaux effectués sur le réseau.). Elle est la plupart du temps comprise entre 0. 1 et 0.3, mais souvent dépasse ces valeurs limites. En France par exemple elle est égale à 2 ; la valeur maximale admissible européenne est de 4 [11].

Donc, le turbidimètre mesure la quantité de lumière diffusée par un échantillon d'eau brute du fait de la présence de particules dans l'eau. Cette valeur est directement proportionnelle à la turbidité de l'échantillon mesuré. Un faisceau lumineux vient toucher la surface sous une incidence telle que ni lui-même ni le faisceau réfléchi ne peut impressionner une cellule photorésistante placée sensiblement perpendiculairement au faisceau incident. Par contre, la lumière diffusée par les particules en suspension vient modifier d'autant plus l'éclairement de la cellule que leur nombre est élevé, ce qui permet d'obtenir la mesure de la turbidité de cette eau.

- Mesure de la conductivité : Le principe mis en œuvre pour la mesure de la conductivité, et de son inverse la résistivité, est simple puisqu'il consiste à mesurer l'intensité du courant électrique recueilli aux bornes de deux électrodes de géométries connues, plongées dans l'eau et soumises à une différence de potentiel alternatif, dont la fréquence doit être d'autant plus élevée que la concentration en acides, sels ou bases dissous est grande, pour éviter les phénomènes de polarisation. La résistivité d'une eau étant fonction du degré de dissociation des molécules dissoutes, la plupart des appareils comportent une compensation automatique de température pour ramener la valeur de la mesure à une température de référence donnée.

- Mesure de pH : Industriellement, la mesure du pH se fait toujours par potentiométrie à l'aide de deux électrodes : une électrode de référence et une électrode de mesure. L'électrode de référence est plongée dans une solution de concentration constante en ions hydrogène. Une cloison, laissant passer le courant électrique, sépare la solution de référence de celle dont on veut mesurer le pH et dans laquelle est plongée l'électrode de mesure. Une tension, fonction linéaire de la concentration en ions hydrogène de la solution, apparaît alors aux bornes des électrodes. Il suffit donc de relier ces bornes à un voltmètre pour connaître la valeur du pH. En pratique, les électrodes sont réunies pour former une sonde.

- Mesure d'Oxygène dissous : L'ampèremètre est utilisé industriellement en traitement des eaux pour la mesure en continu de la concentration en agents oxydants et met en œuvre une méthode simplifiée d'analyse par ampérométrie. La cellule de mesure, qui est alimentée à débit constant en eau à analyser, comporte une cathode inattaquable, par exemple en platine, et une anode qui peut-être en cuivre, en cadmium, en argent, etc. En l'absence d'agent

oxydant, la pile ainsi formée est polarisée et n'est traversée que par un courant très faible. Sa dépolarisation et, par conséquent, l'intensité du courant qu'elle débite sont sensiblement proportionnelles à la concentration de l'agent oxydant qui vient se réduire à la cathode. On mesure ainsi la concentration en chlore, ozone, oxygène d'une eau. L'inconvénient de ces appareils réside dans le fait qu'ils mesurent la somme des agents oxydants et qu'ils ne peuvent être vraiment utilisés que dans le cas où un seul corps se trouve en solution à concentration variable. L'effet d'un autre corps, éventuellement présent à concentration constante, peut être annulé par action sur le zéro de l'appareil.

1.7.3. Qualité des capteurs

Pour que le fonctionnement de l'ensemble du système de mesure soit correct, il est essentiel de s'assurer de la compatibilité de chacun des instruments mis en place en particulier les capteurs. L'information ainsi délivrée, surtout si elle est utilisée dans un système de surveillance, doit être la plus représentative possible de la valeur vraie du paramètre mesuré et être très fiable.

1.7.3.1. Précision, Sensibilité, gamme de mesure

De nombreux facteurs conditionnent l'écart entre la valeur du paramètre mesuré et l'information délivrée. Le premier facteur est la précision du capteur. Celle-ci, exprimée en pourcentage, est le quotient de l'incertitude de la valeur obtenue par l'étendue de mesure pour des conditions de mesure données. La précision du capteur est fonction du processus de mesure mais aussi des corrections annexes qui y sont apportées. Une bonne précision finale dépend d'une bonne corrélation entre une caractéristique et un phénomène étudié. Un autre facteur peut être l'existence d'erreurs systématiques dues à un étalonnage incorrect ou trop peu fréquent du capteur. Les erreurs accidentelles peuvent également être causées par des signaux parasites, ou des absences de correction de température, de pression, etc. La sensibilité initiale d'un appareil de mesure est un autre facteur à prendre en compte. Celle-ci est la valeur minimum du paramètre à mesurer en dessous duquel l'appareil ne réagit pas. La sensibilité en fonctionnement est la plus petite variation du paramètre mesuré décelable par la mesure. Elle n'est pas nécessairement constante dans toute la gamme de mesure. Il faut enfin tenir compte de la gamme de mesure du capteur, qui correspond aux valeurs de seuils au-delà desquels la précision et la sensibilité du capteur se dégradent.

1.7.3.2. Fiabilité et environnement

La fiabilité est définie comme la capacité du capteur à fonctionner correctement, c'est-à-dire à fournir des données avec la précision annoncée. Elle dépend naturellement de la qualité de conception du matériel qui doit être robuste. Mais elle dépend également de son adaptation à l'environnement dans lequel se trouve. Les contraintes des capteurs concernant la gestion de l'eau sont principalement l'humidité et la nature de l'eau. L'humidité peut provoquer de la condensation dans les boîtiers du matériel. Ceux-ci doivent être étanches, des submersions étant toujours possibles, et doivent comporter des dispositifs éliminant la condensation. Cette atmosphère humide peut également provoquer des courts circuits au niveau des câbles de jonction ou d'alimentation. La nature de l'eau, notamment celle des rivières, peut perturber les capteurs immergés avec des dépôts en modifiant les réactions. C'est en particulier le cas de nombreuses sondes dont le nettoyage doit être effectué très régulièrement car ces dépôts provoquent une dérive du capteur. C'est le principal défaut de ce type de capteur dont la surveillance doit être constante, les dispositifs de nettoyage automatique sous forme de brosses ou de rétro-lavage de la sonde n'étant pas toujours efficaces.

En conclusion, pour tirer pleinement parti des avantages des capteurs de mesure et de l'instrumentation associée, il est indispensable d'accepter certaines contraintes telles que le nettoyage des sondes de mesures, l'étalonnage régulier, etc. Malgré ces précautions, certains facteurs peuvent encore perturber l'information délivrée par les capteurs.

1.7.4. Les méthodes de surveillance des eaux potables

Quand on parle de surveillance des eaux potables, il s'agit en fait de connaître l'état de l'eau en continu (à chaque instant) à partir des différents paramètres ayant trait à sa qualité.

1.7.4.1. Méthode classique : essais de traitabilité en laboratoire

Cette technique a pour but de connaître les différents paramètres de l'eau brute pour décider après sur son état propre, et par suite chercher les techniques et méthodes pour la rendre potable. Ces méthodes sont traditionnelles, déterminées à l'aide d'un essai expérimental appelé « Jar-test » [1]. On procède généralement à un certain nombre de mesures utiles pour le test de qualité tels que : le contrôle bactériologique, le contrôle de désinfection, et le contrôle physico-chimique (pH, T°, Turbidité, conductivité, oxygène

dissous,……). La dose optimale recherchée est déterminée en fonction de la qualité des différentes eaux comparées. La fréquence de ces Jar-Test est souvent irrégulière. En général dans les usines importantes, un seul essai est effectué par jour [11]. L'opérateur fera un nouvel essai entre temps pour changer la dose du coagulant uniquement si la qualité de l'eau traitée se dégrade. L'inconvénient de cette technique est qu'elle nécessite de façon non stop des interventions et des déplacements sur site de l'opérateur. Ce type d'approche a également le désavantage d'avoir un temps de réponse relativement long. En effet, on ne modifie la dose du coagulant qu'une fois l'évènement apparu, vérifié puis analysé. De plus, elle ne permet pas de suivre finement l'évolution de la qualité de l'eau brute. Par exemple, si l'eau brute devient plus « facile à traiter » l'opérateur ne le verra pas forcément et donc ne modifiera pas la dose du coagulant, d'où un coût d'exploitation plus élevé que nécessaire et une économie non réalisée. En voici tout l'intérêt de disposer d'un contrôle automatique de ce procédé pour une meilleure efficacité de traitement et une réduction des coûts d'exploitation. La régulation de l'eau brute au niveau des usines de traitement doit se faire de façon immédiate en se basant sur une surveillance continue des paramètres descripteurs de la qualité de cette eau.

1.7.4.2. Surveillance moderne

La fonction surveillance moderne de l'exploitation d'un tel processus à travers des données quantifiables et qualifiables permet ainsi de détecter les états anormaux de l'objet à surveiller et prendre ainsi les décisions pour un meilleur état. L'importance de la mesure en continu des paramètres physiques et physico-chimiques à l'aide de capteurs dans un système de surveillance monté sur place vient du fait que :

- Ces paramètres ne sont pas conservatifs et changent instantanément.
- Les mesures sont relativement simples, rapides et peu coûteuses.

Ces mesures permettent de détecter immédiatement des anomalies de la composition de l'eau (élévation du pH par exemple) ce qui permet une intervention immédiate.

- **Les avantages des systèmes de surveillance**

La surveillance est un dispositif passif, informationnel qui analyse l'état du système. Elle consiste notamment à détecter et classer les anormalités en observant l'évolution du système puis à les diagnostiquer en localisant les éléments défaillants et en identifiant les causes premières, et prendre les décisions nécessaires et finales sur l'état de l'objet surveillé.

La surveillance se compose de deux fonctions principales, qui sont la détection et le diagnostic. Pour détecter toute anomalie du système ou au niveau de l'objet à surveiller, il faut être capable de classer les situations observables comme étant normales ou anormales. Cette classification n'est pas triviale, étant donné le manque d'information qui caractérise généralement les situations anormales. Une simplification communément adoptée consiste à considérer comme anormale toute situation qui n'est pas normale. Quant à l'objectif de la fonction diagnostic, c'est de rechercher les causes et de localiser les organes qui ont entraîné une observation particulière. Cette fonction se compose de deux fonctions élémentaires : localisation et identification. La localisation permet de déterminer le sous-ensemble fonctionnel défaillant. Alors que l'identification consiste à déterminer les causes qui ont mené à une situation anormale. Les avantages des systèmes de surveillance basés sur des méthodes modernes sont [1] :

- ✓ Amélioration des conditions d'exploitation et des performances d'une installation.
- ✓ Augmentation de la productivité.
- ✓ Fonctions temps réel et différé.
- ✓ Aide à la décision et à la maintenance.

La figure 1.3 montre l'exemple d'une boucle de supervision (surveillance + action) dans une usine moderne.

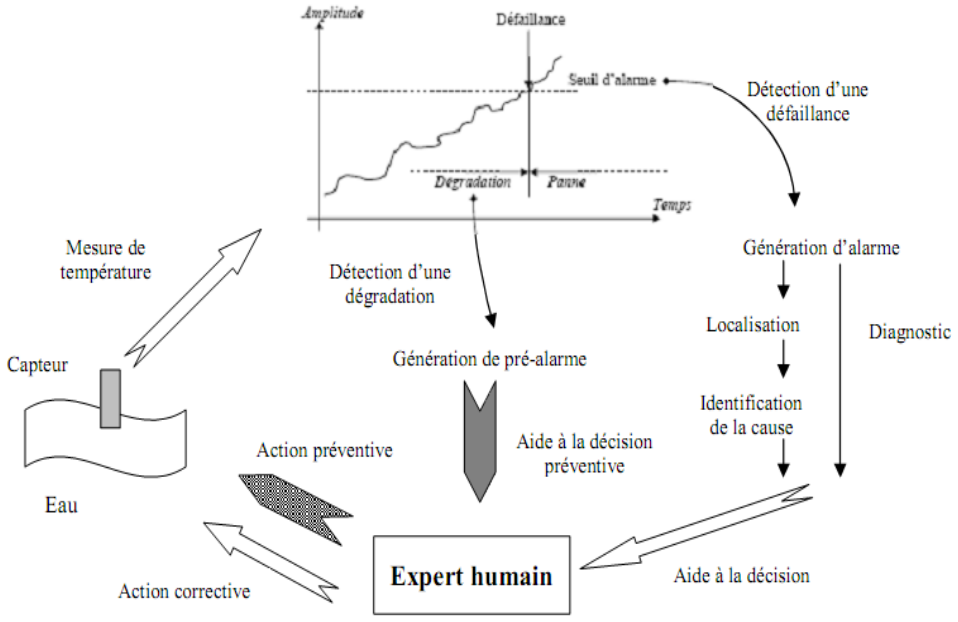


Fig. 1.3. Exemple d'une boucle de supervision d'une usine d'eau.

✓ Méthodes de surveillance

La surveillance avec modèle se base essentiellement sur deux techniques : les méthodes de redondance physique et analytique, et les méthodes d'estimation paramétrique. D'un autre coté, les méthodes qui ne se basent pas sur l'existence d'un modèle se divisent à leur tour en deux principales catégories : les méthodes utilisant des outils statistiques, et les méthodes de reconnaissance de formes [2]. Les outils statistiques établissent des tests sur les signaux d'acquisition. Des tests qui ne sont capables d'assurer que la fonction de détection de défauts. Par contre, les techniques de surveillance par reconnaissance de formes, sont plus élaborées par rapport aux simples tests statistiques et sont capables de détecter et de diagnostiquer toute anomalie de l'état d'un objet donné à surveiller. La figure 1.4 indique une classification des méthodologies de surveillance existantes actuellement.

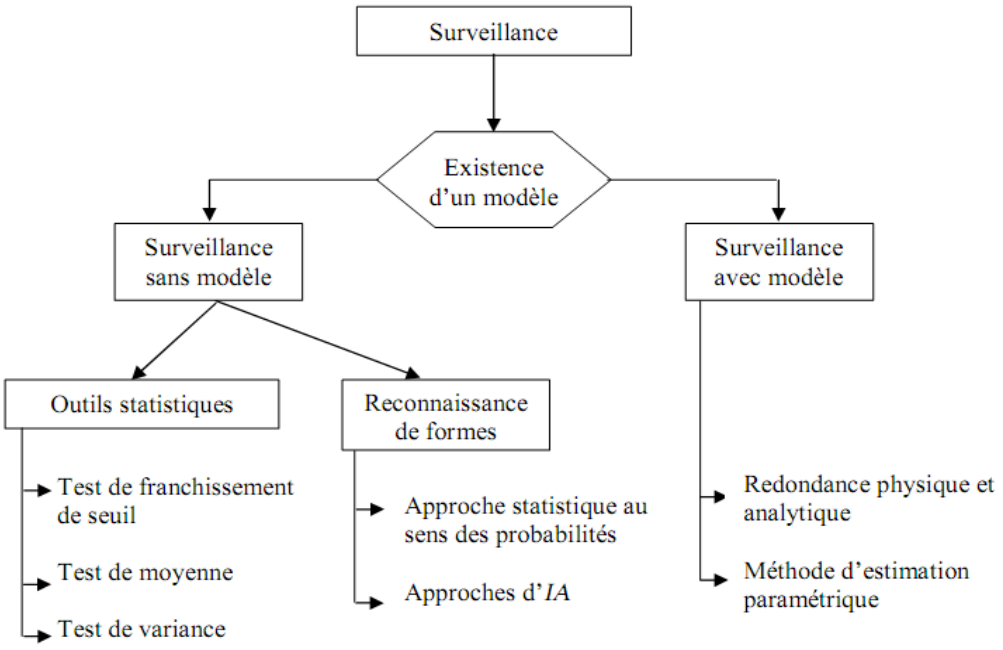


Fig. 1.4. Classification des méthodologies de surveillance.

✓ Méthodes de surveillance avec modèles

Les méthodes de surveillance avec modèle ont pour principe de comparer les mesures effectuées sur le système aux informations fournies par le modèle [2]. Tout écart est alors synonyme d'un état anormal. Les outils de la théorie de la décision sont ensuite utilisés pour déterminer si cet écart est dû à des états normaux, comme par exemple le bruit de mesure, ou s'il traduit un défaut du système, ou bien encore une dégradation de la qualité. Ces méthodes

peuvent être décomposées en deux catégories : les techniques de redondance physique et analytique, et les techniques d'estimation paramétrique. Ces deux techniques sont présentées brièvement.

a) Redondances physiques et analytiques

Redondances physiques : Afin de fiabiliser l'état anormal à partir des signaux mesurés, il faut un moyen pour distinguer les défaillances capteurs ou une dégradation. La méthode la plus simple consiste à utiliser la redondance physique. Il s'agit de doubler ou tripler des composantes de mesure du système [2]. Si ces composantes identiques placées dans le même environnement émettent des signaux identiques, on considère que ces composants sont dans un état de fonctionnement nominal et, dans le cas contraire, on considère qu'une défaillance capteur s'est produite dans au moins une des composantes. Cette méthode par redondance physique a l'avantage d'être conceptuellement simple mais est coûteuse à être mise en œuvre dans le cas de mesures de plusieurs paramètres du système, et conduit nécessairement à des installations encombrantes. Elle est, par conséquent, utilisée uniquement pour la surveillance des sous-ensembles critiques d'un système. Un autre inconvénient est que les composantes identiques fabriquées dans la même série peuvent se dégrader de la même façon et tomber en panne en même temps. Pour pallier ce dernier inconvénient, on peut utiliser des composantes différentes qui remplissent la même fonction.

Redondances analytiques : Les méthodes de redondance analytique nécessitent un modèle du système à surveiller. Ce modèle comprend un certain nombre de paramètres dont les valeurs sont supposées connues lors du fonctionnement nominal [10]. Dans la mesure où la surveillance est établie à partir des mesures échantillonnées des grandeurs observables du système, la modélisation de ce dernier sous forme discrète semble être raisonnable. Le but des méthodes de redondance analytique est d'estimer l'état du système afin de le comparer à son état réel. L'estimation de l'état du système peut être réalisée soit à l'aide de techniques d'estimation d'état, soit par l'obtention de relations de redondance analytique. La théorie de la décision est ensuite utilisée pour déterminer si l'écart observé est dû à des états normaux du fonctionnement, ou à des défaillances dans le système.

b) Méthodes d'estimation paramétrique

Les méthodes d'estimation paramétrique supposent l'existence d'un modèle paramétrique décrivant le comportement du système et que les valeurs de ces paramètres en fonctionnement nominal soient connues [11]. Elles consistent alors à identifier les paramètres caractérisant le fonctionnement réel, à partir de mesures des entrées et des sorties du système. On dispose ainsi d'une estimation des paramètres du modèle, effectuée à partir des mesures prises sur le système et de leurs valeurs réelles. Pour détecter l'apparition de défaillances dans le système, il faut effectuer la comparaison entre les paramètres estimés et les paramètres réels. Comme pour les méthodes de redondance analytique, la théorie de la décision sert alors à déterminer si l'écart observé est dû à des états normaux ou à des défaillances. La différence système. Les méthodes d'estimation paramétrique requièrent donc l'élaboration d'un modèle dynamique précis du système à surveiller. Ceci restreint leur utilisation à des procédés bien définis. Les valeurs estimées sont utilisées comme base pour la détection et le diagnostic d'untel système à surveiller. entre les méthodes de redondance analytique et les méthodes d'estimation paramétrique est qu'on effectue, pour les premières, la comparaison entre l'état estimé et l'état réel du système, alors que pour les secondes, on compare les paramètres estimés aux paramètres réels du système. Les méthodes d'estimation paramétrique requièrent donc l'élaboration d'un modèle dynamique précis du système à surveiller. Ceci restreint leur utilisation à des procédés bien définis. Les valeurs estimées sont utilisées comme base pour la détection et le diagnostic d'un tel système à surveiller.

- ✓ **Méthodes de surveillance sans modèles**

Dans de nombreuses applications industrielles le modèle est difficile à construire, un modèle mathématique est quasiment impossible à cause de ses caractéristiques dynamiques et stochastiques. Pour cela, les seules méthodes de surveillance opérationnelles sont celles sans modèle. Deux solutions existent dans ce cas : la surveillance avec des tests statistiques, et la surveillance par reconnaissance de formes. La première technique est moins élaborée que la deuxième, dans le sens où elle ne remplit qu'une partie de la surveillance.

a) Surveillance avec outils statistiques

Les outils statistiques consistent à supposer que les signaux fournis par les capteurs possèdent certaines propriétés statistiques. On effectue alors quelques tests qui permettent de vérifier si ces propriétés sont présentes dans un échantillon des signaux mesurés [5].

Test de franchissement de seuils : Le test le plus simple est de comparer ponctuellement les signaux avec des seuils préétablis. Le franchissement de ce seuil par un des signaux capteurs génère une alarme. Ce type de méthode est très simple à mettre en œuvre mais ne permet pas d'établir un diagnostic des défaillances ou de dégradation. Cette méthode est aussi très sensible aux fausses alarmes (figure 1.5).

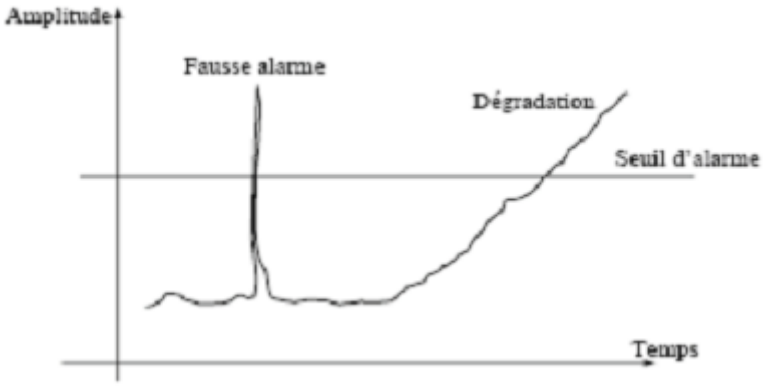


Fig. 1.5. Sensibilité de la méthode à franchissement de seuils aux fausses alarmes.

Test de moyenne : Contrairement à la méthode précédente, le test de comparaison est effectué sur la moyenne du signal contenu dans une fenêtre de n valeurs plutôt que sur une valeur ponctuelle.

Test de variance : On peut également calculer la variance d'un signal. Tant que cette variance se situe dans une bande située autour de sa valeur nominale, l'évolution du système est supposée normale.

b) Surveillance par reconnaissance de formes

L'approche de surveillance par reconnaissance des formes permet d'associer un ensemble de mesures (continues ou discrètes) effectuées sur le système à des états de fonctionnement connus. Cette fonction permet d'avoir une relation d'un espace caractéristique vers un espace de décision, de façon à minimiser le risque de mauvaise classification. Deux

techniques de reconnaissance des formes sont présentées. La première technique présentée est une technique classique de discrimination basée sur les outils de la probabilité. Cette technique peut se montrer insuffisante car elle suppose une connaissance a priori de tous les états de fonctionnement et ne prend pas en compte l'évolution du système. La deuxième technique de discrimination qui sera présentée repose sur la théorie de l'intelligence artificielle (IA). Ces techniques d'IA ont l'avantage de ne pas se baser sur les connaissances a priori des états de fonctionnement, mais plutôt sur une phase d'apprentissage. Deux techniques sont très utilisées dans plusieurs domaines d'application : la reconnaissance des formes par réseaux de neurones artificiels, et la reconnaissance des formes par les machines à vecteurs de support.

Les réseaux de neurones (RNAs), sont des outils de l'intelligence artificielle, capables d'effectuer des opérations de régression ou de classification. Leur principal avantage par rapport aux autres outils est leur capacité d'apprentissage et de généralisation de leurs connaissances à des entrées inconnues. Ils peuvent également être implémentés en circuits électroniques, offrant ainsi la possibilité d'un traitement en temps réel. Le processus d'apprentissage est donc une phase très importante pour la réussite d'une telle opération. Une des qualités de ce type de techniques, est leur adéquation pour la mise au point de systèmes de surveillance modernes, capables de s'adapter à d'éventuelles extensions et reconfigurations multiples. Nous détaillerons cette technique et sa mise en œuvre dans le chapitre trois de notre mémoire.

La figure 1.6, montre l'architecture générale qu'on peut imaginer pour une application de surveillance de l'eau potable par reconnaissance de formes. L'expert humain joue un rôle très important dans ce type d'application. Toute la phase d'apprentissage supervisé dépend de son analyse des variables du système, chaque décision est caractérisée par un ensemble de données (formes d'entrée) recueillies sur le système. L'association (entrées-sorties) sera apprise par la technique utilisée (RNAs). Après cette phase d'apprentissage, l'algorithme associera les décisions nécessaires représentant les sorties (capteurs logiciels) du système aux formes d'entrée par les données du système.

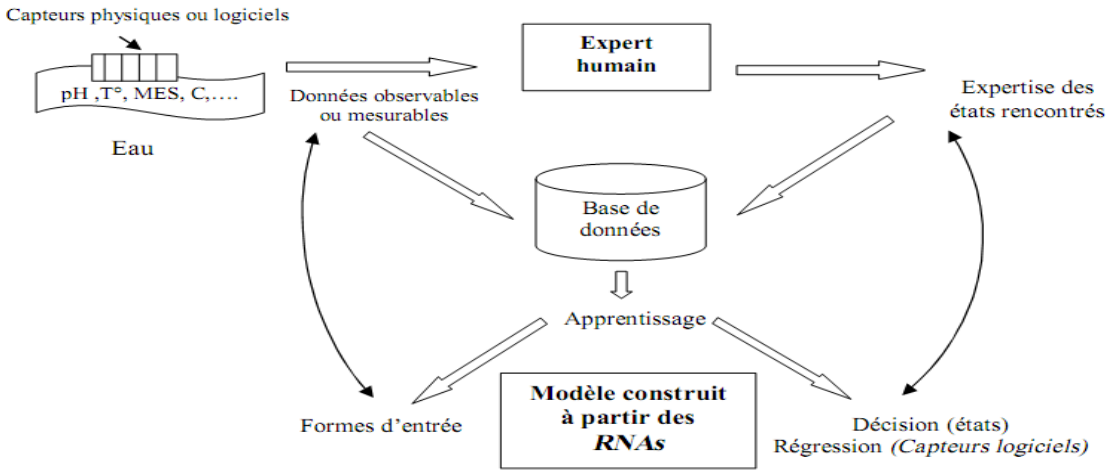


Fig.1.6. Schéma général du système de surveillance par reconnaissance de formes

1.8. Notre problématique d'application

Dans notre travail, la surveillance de l'eau brute peut être vue comme un problème de reconnaissance de formes (régression ou classification), où les formes représentent l'ensemble des paramètres relatifs à la qualité de l'eau, et les sorties correspondent aux différents états de l'eau (potable ou non potable) dans le cas d'une classification des données ou la prédiction de quelques paramètres physico-chimiques non mesurables en continu délivrés par des capteurs logiciels (soft sensor) dans le cas d'une régression des données (capteurs logiciels) ; cas étudié dans le présent mémoire. L'architecture modulaire du système de surveillance par reconnaissance de formes basée sur une approche multisensorielle, est présentée dans la figure 1.7.

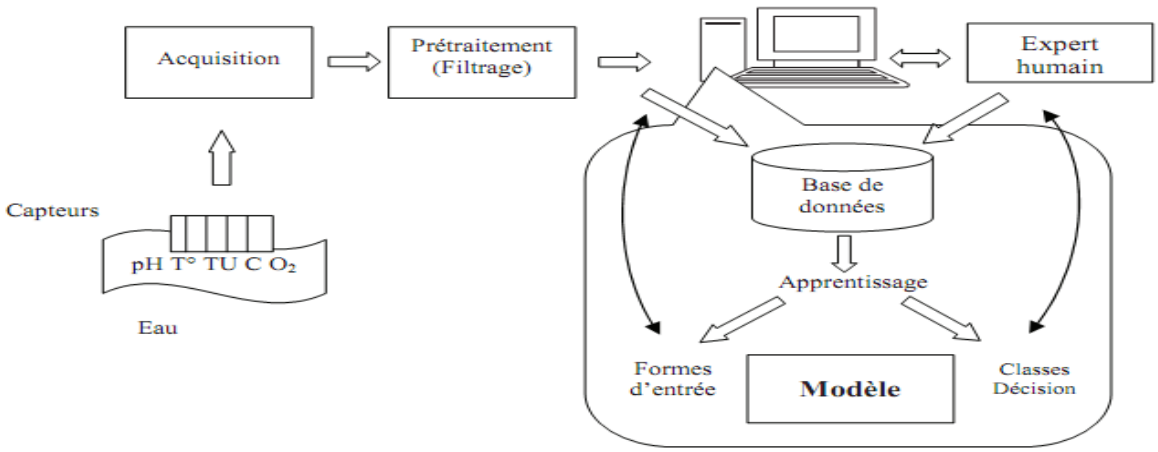


Fig. 1.7 Système général de surveillance

Les différents paramètres de l'eau sont transformés en signaux électriques à l'aide de capteurs, et transmis à travers un système d'acquisition vers une station de contrôle (environnement PC par exemple). A ce niveau, le traitement et l'analyse des signaux acquis sont effectués. Le système PC fournit le résultat de décision attendu (eau potable, eau non potable), et effectue éventuellement un apprentissage hors ligne exploitant les données ainsi reçues. La présence d'un expert humain est indispensable pour l'enrichissement de la base de données à constituer.

Il s'agit maintenant d'implanter au niveau du système la méthode d'apprentissage choisie qui représente notre objectif principal dans ce travail. Les chapitres suivants font l'objet de cette étude, et la simulation de la technique employée est par conséquent validée sur la base d'une évaluation des performances.

Conclusion

Ce premier chapitre a servi d'introduction au domaine de contrôle et de surveillance des eaux brutes. Les différentes étapes d'une chaîne de traitement sont présentées. Les paramètres ainsi que les capteurs physico-chimiques utilisés comme source d'information ayant trait à la qualité de l'eau ont été particulièrement décrits. De même, les différentes techniques existant actuellement dans le domaine du contrôle et de surveillance des eaux potables ont été aussi évoquées. Il est d'ores et déjà apparu qu'un contrôle automatique et permanent basé sur les paramètres descripteurs de l'eau, peut présenter une solution très intéressante. Le schéma de principe de surveillance proposé en fin de ce chapitre, illustre cette approche de fusion multisensorielle.

Dans le chapitre suivant, nous présentons la notion de la l'analyse de multivariée de données multidimensionnelle dans le sens de réduction de dimension, comme une solution à l'origine de l'extraction des paramètres caractéristique appliquée aux systèmes de contrôle et de surveillance à bases des techniques de reconnaissance de formes. Des systèmes capables de synthétiser l'information permettant de surveiller l'eau à traiter.

CHAPITRE II

ANALYSE MULTIVARIEE

Introduction

Les méthodes d'analyse multivariée de données ont commencées à être développées dans les années 50, poussées par le développement de l'informatique et du stockage des données depuis n'a cessé de croître. L'analyse de données a surtout été développée en France par J.P. Benzécri, qui a su par l'analyse des correspondances représenter les données de manière simple et interprétable. Parmi les méthodes de l'analyse multivarié en trouve la méthode d'analyse en composantes principales (ACP). Cette méthode permet de faire le traitement statistique de données, dont le but de représenter et d'expliquer les liaisons statistiques entre les phénomènes. Elle permet d'identifier des variables sous-jacentes, ou facteurs qui expliquent les corrélations à l'intérieur d'un ensemble de variables observées. Elle est souvent utilisée pour réduire un ensemble de données, et dans l'agrégation de l'information, en identifiant un petit nombre de facteurs qui expliquent la plupart des variances observées dans le plus grand nombre de variables manifestes. On peut également utiliser l'analyse factorielle pour résumer, synthétiser, et hiérarchiser l'information acquise pour l'étape de décision.

Ce chapitre, est consacré à l'analyse multivariée des données et leur objectif ainsi que de quelques méthodes d'analyse ; on doit insister sur la méthode d'analyse en composantes principales (ACP), qui est choisie dans notre approche. On doit introduire quelques méthodes utilisées dans le cadre de régression et classification de données telles que les réseaux de neurones, comme des fonctions de décision dans un système de traitement à fusion multisensorielle. Une étude détaillée des mécanismes de ces méthodes appliquées à l'analyse statistique des données pour la régression et de classification de données, est aussi étudié.

2.1. Etat de l'art

Il décrit l'analyse de données selon cinq principes, un peu désuets aujourd'hui :

- 1er principe : Statistique n'est pas probabilité.
- 2ème principe : Le modèle doit suivre les données et non l'inverse.
- 3ème principe : Il convient de traiter simultanément des informations concernant le plus grand nombre possible de dimensions.
- 4ème principe : Pour l'analyse des faits complexes et notamment des faits sociaux, l'ordinateur est indispensable.
- 5ème principe : Utiliser un ordinateur implique d'abandonner toutes techniques conçues avant l'avènement du calcul automatique[12].

Ces cinq principes montrent bien l'approche d'une part de la statistique à la différence des probabilités, les modèles doivent coller aux données et d'autre part de l'analyse de données, il faut traiter le plus grand nombre de données simultanément ce qui implique l'utilisation de l'ordinateur et ainsi l'utilisation de nouvelles techniques adaptées.

L'analyse de données fait toujours l'objet de recherche pour s'adapter à tout type de données et faire face à des considérations de traitements en temps réel en dépit de la quantité de données toujours plus importante. Les méthodes développées (l'analyse de données) sont maintenant souvent intégrées avec des méthodes issues de l'informatique et de l'intelligence artificielle (apprentissage numérique et symbolique) dans le data mining traduit en français par fouille de données ou encore extraction de connaissance à partir de données, la technique utilisée dans notre approche de l'Analyse en composantes principales (ACP) [13]

2.2. Analyse des données

2.2.1. Définition

Dans l'acception française, la terminologie « analyse des données » désigne un sous-ensemble de ce qui est appelé plus généralement la statistique multivariée. L'analyse des données est un ensemble de techniques descriptives, dont l'outil mathématique majeur est l'algèbre matriciel, et qui s'exprime sans supposer a priori un modèle probabiliste[14].

2.2.2. Quelques exemples de méthodes d'analyse multivariée des données

Analyse en composantes principales : La méthode d'Analyse en Composantes Principales permet d'étudier un tableau individus x variables dans le cas où toutes les variables sont quantitatives. La méthode permet d'obtenir une carte des individus en fonction de leurs proximités et une carte des variables en fonction de leurs corrélations [1].

L'analyse factorielle des correspondances : L'Analyse Factorielle des Correspondances (Analyse des Correspondances Simples ou Binaires) permet de représenter graphiquement un tableau de contingence créé par le ou les croisements (tris croisés) de deux ou plusieurs variables qualitatives. La méthode vise à rassembler sur un ou plusieurs graphiques (plan factoriel) la plus grande partie possible de l'information contenue dans le tableau en s'attachant non pas aux valeurs absolues mais aux correspondances entre les caractéristiques, c'est-à-dire aux valeurs relatives [15].

L'analyse des correspondances multiples : L'Analyse des Correspondances Multiples permet d'analyser un tableau individus x variables lorsque les variables sont qualitatives. Cette méthode effectue une analyse des correspondances sur le tableau disjonctif complet obtenu en remplaçant dans le tableau d'origine chaque variable qualitative par l'ensemble des variables indicatrices des différentes modalités de cette variable [16].

L'analyse sur tableau de distances ou de dissimilarités : La méthode d'Analyse sur Tableau de Distances ou de Dissimilarités permet d'étudier un tableau carré symétrique (à diagonale nulle) individus x individus contenant à chaque intersection ligne-colonne la distance ou la dissimilarité entre cette ligne et cette colonne. La méthode permet d'obtenir une carte des individus en fonction de leurs proximités ou dissimilarités [16].

La classification ascendante hiérarchique : La méthode de Classification Ascendante Hiérarchique permet de construire une typologie (ou partition) d'un ensemble d'individus en classes telles que les individus appartenant à une même classe sont proches alors que les individus appartenant à des classes différentes sont éloignés.

L'analyse factorielle multiple : L'Analyse factorielle multiple est spécialement conçue pour étudier une population d'individus caractérisés par un certain nombre de groupes de variables. Ces groupes de variables peuvent être constitués de variables mesurées à différents instants,

mais aussi de sous-tableaux issus d'un seul tableau : ces sous-tableaux correspondent alors à des regroupements de variables selon des critères [17].

L'analyse factorielle de données mixtes : L'Analyse Factorielle de Données Mixtes est une méthode spécialement conçue pour permettre l'étude simultanée de variables quantitatives et qualitatives (données dites mixtes) mesurées sur une population d'individus en tant qu'éléments actifs dans une même analyse. Cette analyse prend en compte les variables quantitatives comme une analyse en composantes principales normées (ACPN) et les variables qualitatives comme une analyse des correspondances multiples (ACM)[16].

L'analyse discriminante pas à pas : L'Analyse Discriminante Pas à Pas permet de sélectionner à partir d'un ensemble de variables quantitatives et d'une variable qualitative découpant la population en plusieurs groupes (2 ou plus), le sous-ensemble des variables quantitatives les plus explicatives des groupes qui seront alors utilisées pour définir des fonctions discriminantes robustes.

L'analyse factorielle discriminante : L'Analyse Factorielle Discriminante est une méthode géométrique permettant de construire à partir d'un ensemble de variables quantitatives et d'une variable qualitative découpant la population en plusieurs groupes (2 ou plus), des fonctions discriminantes qui les séparent au mieux dans l'échantillon d'apprentissage (population de base)[16].

L'analyse discriminante bayesienne : L'Analyse Discriminante Bayesienne (ou Stochastique) permet de construire à partir d'un ensemble de variables quantitatives et d'une variable qualitative découpant la population en plusieurs groupes (2 ou plus), des fonctions discriminantes qui définissent une règle de décision optimale à partir de laquelle on peut affecter des individus tests ou anonymes aux différents groupes. Cette technique suppose que l'on connaisse a priori les probabilités d'appartenance aux différents groupes et que les données suivent une loi multi-normale[13].

L'analyse discriminante qualitative : L'Analyse Discriminante Qualitative est une généralisation de l'Analyse Factorielle Discriminante dans le cas ou les variables explicatives sont qualitatives et non plus quantitatives. La première étape de l'analyse consiste à mettre en œuvre une Analyse des Correspondances Multiples des variables qualitatives. La deuxième étape remplace les variables qualitatives d'origine par les coordonnées sur les axes factoriels

issus de l'ACM et effectue sur ces données une Analyse Factorielle Discriminante. Les fonctions discriminantes sont ensuite exprimées en fonction des indicatrices desmodalités des variables qualitatives d'origine [16].

La régression sur composantes principales : La méthode de Régression sur Composantes Principales est une technique de régression utile lorsque de fortes colinéarités entre les variables explicatives sont présentes et que l'on ne désire pas utiliser les algorithmes de régression pas à pas pour éliminer les variables corrélées entre elles ou la régression Ridge. Cette technique utilise à la fois l'Analyse en Composantes Principales et la Régression Multiple pour élaborer un modèle dont les coefficients sont stables [18].

2.2.3. Objectif s de réduction des données

La réduction de dimension a pour objectif de créer un nombre réduit de variables qui décrivent les données de base presque aussi bien que le font les variables « brutes », habituellement en grand nombre. Ces nouvelles variables seront moins redondantes que les variables initiales.

Alors que le processus de réduction de données fera en général perdre de l'information, les objectifs essentiels de la réduction de la dimension des données sont citées ci-après :

La raison la plus évidente, est de réduire la quantité d'information que les algorithmes auront à traiter, réduisant ainsi, parfois fortement, les temps de calcul et l'encombrement mémoire.

Réduire le nombre de variables à 2 permet de donner de la base une représentation visuelle plane, ce qui permet de mettre en œuvre l'outil de d'Analyse des Données de loin le plus puissant qui existe : l'œil, et son fantastique système de détection de regroupements, d'alignements, de tendances, de dérives, de « niches » etc…

Mais la raison la plus importante a trait à la crédibilité à accorder à un modèle (prédictif ou descriptif). Pour des raisons fondamentales de compromis, à erreurs commises sur les données disponibles égales, un modèle ne prenant en entrée que peu de variables sera plus crédible qu'un modèle utilisant un grand nombre de variables d'entrée.

La réduction de dimensionnalité est un exercice à la fois indispensable et difficile. L'analyste se doit de lui consacrer le temps nécessaire sous peine de construire des modèles apparemment de bonne qualité, mais qui en fait ne représentent pas la réalité sous-jacente [1].

2.2.4. Les différentes méthodes de réduction

Les méthodes de réduction de dimension peuvent être supervisées ou non supervisées.

2.2.4.1. Les méthodes non supervisées

Dans les cas des méthodes non supervisées, nous exploitons les données sans connaissances préalables du modèle. Nous citerons quelques unes des ces techniques de prétraitement de données précédant l'analyse postérieure de celles-ci : l'analyse en composantes principales, la décomposition en valeur singulière ainsi que l'analyse en composantes indépendantes. Les deux différences majeures entre la sélection des caractéristiques, et les méthodes du prétraitement non supervisé destinées à projeter les données dans un nouvel espace de dimension inférieur (et donc dans un but de réduction) sont:

- Au lieu de choisir un sous espace de caractéristiques, la projection de données crée de nouvelles dimensions définies par une fonction de toutes les autres caractéristiques.
- Puis cette même réduction ne considère pas le label des classes, mais plutôt les données point par point [1].

2.2.4.2. Les méthodes supervisées

Contrairement aux méthodes non supervisées, celles-ci imposent des limites à l'analyse des données sous certaines contraintes du modèle. Nous énumérons quelques unes : l'analyse discriminante linéaire (LDA), l'algorithme de classification à erreurs minimales (MCE) et les séparateurs à vastes marges (SVM) [15] .

La méthode d'analyse en composantes principales sera détaillée, car c'est la méthode utilisée dans notre application [19].

2.3. Analyse en composantes principales

L'analyse en composantes principales est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui est utile pour la compression et la classification des données. Le problème consiste à réduire la dimensionnalité d'un ensemble des données (échantillon) en trouvant un nouvel ensemble de variables plus petit que l'ensemble original des variables, qui néanmoins contient la plupart de l'information de

l'échantillon. L'ACP consiste à transformer des variables liées entre elles (dites "corrélées" en statistique) en nouvelles variables indépendantes les unes des autres (donc "non corrélées").

Ces nouvelles variables, appelées composantes principales ou axes, sont ordonnées par fraction de l'information totale que chacune contient [1].

2.3.1. Bref historique de l'ACP

Conçue par Karl Pearson en 1901, intégrée à la statistique mathématique par Harold Hotelling en 1933, l'analyse en composantes principales (ACP) n'est vraiment utilisée que depuis la large diffusion des moyens de calcul informatique.

La technique d'analyse en composantes principales peut être présentée de divers points de vue. Pour le statisticien classique, il s'agit de la recherche des axes principaux de l'ellipsoïde d'une distribution normale multidimensionnelle, ces axes étant estimés à partir d'un échantillon. C'est la présentation initiale de Hotelling (l933), puis celle des manuels classiques d'analyse multivariée.

Pour le factorialiste classique, il s'agit d'un cas particulier de la méthode d'analyse factorielle des psychométriciens (cas de variances spécifiques nulles ou égales).

Enfin, du point de vue plus récent des analystes de données, il s'agit d'une technique de représentation des données, ayant un caractère optimal selon certains critères algébriques et géométriques, et que l'on utilise en général sans référence à des hypothèses de nature statistique ni à un modèle particulier.

Ce point de vue, fort répandu actuellement est peut-être le plus ancien. C'est celui qui avait été adopté par Pearson (1901). Bien entendu, il ne s'agissait pas de l'analyse en composantes principales telle que nous la présentons, mais les idées essentielles de la méthode étaient déjà présentées par cet auteur.

L'analyse en composantes principales présente de nombreuses variantes selon les transformations apportées au tableau de données : le nuage des points-individus peut être centré ou non, réduit ou non. Parmi ces variantes, l'analyse en composantes principales normée (nuage centré-réduit) est certainement la plus utilisée et c'est celle-ci que nous choisissons pour présenter les principes de l'analyse.

2.3.2. Quand utiliser l'ACP ?

L'ACP a l'avantage d'être une méthode pour laquelle il n'y a aucun paramètre à préciser au départ. Elle permet également d'analyser les associations entre un très grand

nombre de profils d'expression (plusieurs milliers) sans aucun problème et est implémentée dans de nombreux outils conviviaux comme J-Express (Dysvik and Jonassen 2001). Pour ces raisons, c'est sans doute la première méthode à appliquer à un jeu de données que l'on souhaite mieux comprendre. Le résultat de l'ACP donne une première idée de la structure des données et peut orienter les analyses ultérieures, comme par exemple choisir un nombre de groupes pertinent pour la classification[14].

2.3.3. Principe

L'idée principale de la décomposition orthogonale propre est de trouver un ensemble de vecteurs de base orthonormaux et ordonnés dans un sous espace sans perte conséquente d'information où l'on exprimera de manière optimale les vecteurs aléatoires de données en utilisant les premiers vecteurs de base obtenus. C'est une méthode qui souligne les similarités ainsi que les différences des données dans un espace à grandes dimensions où l'on n'a pas le luxe de la représentation graphique en identifiant la dépendance des structures des observations stochastiques multidimensionnelles dans le but d'avoir sa description compacte.

Elle est communément considérée comme un outil de visualisation des données, cependant c'est aussi un moyen :

- de décorréler ces données ; dans le nouvel espace, constitué des nouveaux axes.
- de débruiter ces données, en considérant que les axes que l'on oublie sont des axes bruités.

Cette technique peut être vue de deux points de vue : celui de maximiser la variance ou encore celui de minimiser l'erreur quadratique moyenne. Les composantes principales sont représentées dans le nouvel espace, Elles sont décorrélées et ordonnées dans le sens décroisant de la variance [15].

$$M = \begin{bmatrix} X_{1,1} \dots X_{1,N} \\ \vdots \quad \ddots \quad \vdots \\ X_{K,1} \dots X_{K,N} \end{bmatrix}$$

Il est usuel d'appliquer une analyse en composantes principales sur un ensemble de « *N* »

variables aléatoires $X_1, \ldots, X_N$ connues de dimension « K ». Ces « N » variables aléatoires peuvent être structurées dans une matrice M à K lignes et N colonnes. Chaque variable aléatoire $Xn = (Xn,1, \ldots, Xn, K)'$ a une moyenne $\overline{X}_n$ et un écart type X_n

Poids

Si les réalisations (les éléments de la matrice M) sont à probabilités égales alors chaque réalisation (un élément $X_{i,j}$ de la matrice) a la même importance $1/n$ dans le calcul des caractéristiques de l'échantillon. On peut aussi appliquer un poids p_i différent à chaque réalisation conjointes des variables (cas des échantillons redressés, des données regroupées, ...). Ces poids, qui sont des nombres positifs de somme 1 sont représentés par une matrice diagonale D de taille K:

$$D = \begin{bmatrix} p_1 & & 0 \\ & P_2 & \\ & \ddots & \\ 0 & & P_K \end{bmatrix}$$

Dans le cas le plus usuel de poids égaux, $D = \frac{1}{K} I$ où ***I*** est la matrice identité.

2.3.4. Traitement des données

On notera « g » le vecteur $\left(\overline{X}_1, \cdots, \overline{X}_N\right)$ qui représente le centre de gravité du nuage de points**.** Les données à traiter sont couramment centrées (normalisées) sur le centre de gravité :

La matrice M deviendra :

$$\overline{M} = \begin{bmatrix} X_{1,1} - \overline{X}_1 \cdots X_{1,N} - \overline{X}_N \\ \vdots \quad \ddots \quad \vdots \\ X_{K,1} - \overline{X}_1 \cdots X_{K,N} - \overline{X}_N \end{bmatrix} = M - \tilde{1} g^T$$

Elle peut être aussi **réduite** :

$$\widetilde{M} = \begin{bmatrix} \frac{X_{1,1} - \overline{X}_1}{\sigma(X_1)} & \cdots & \frac{X_{1,N} - \overline{X}_N}{\sigma(X_N)} \\ \vdots & \ddots & \vdots \\ \frac{X_{K,1} - \overline{X}_1}{\sigma(X_1)} & \cdots & \frac{X_{K,N} - \overline{X}_N}{\sigma(X_N)} \end{bmatrix}$$

Le fait de réduire ou non le nuage de points (X_1, …, X_N) dépend du modèle et du devenir des données:

· Lorsque les données ne sont pas réduites, toute variable à forte variance (un bruit par exemple) va majorer toutes les autres et faussera par la suite les résultats d'une PCA.

· Et lorsque les données sont réduites, toutes les variables vont se retrouver avec une variance apparente égale, ce qui rendra le bruit équivalent à une variable informative.

Dans ce cas, le choix de cette normalisation ou encore de ce blanchissement dépendra des données en premier lieu, et puis des applications postérieures des résultats de PCA[20].

2.3.5. Calcul des covariances et des corrélations

Le calcul des matrices de covariances et de corrélations est classique, du moment que les données sont déjà structurées en matrices et éventuellement normalisées ou réduites, il faut juste multiplier celle-ci par sa transposée :

· La matrice de variance-covariance des « X_1, …, X_N si *M »* est juste normalisée :

$$\text{Covariance} = \frac{1}{K . \overline{M}^T . \overline{M}}$$

- la matrice de corrélation des X_1, …, X_N si *M* est réduite :

$$\text{Corrèlations} = \frac{1}{K . \widetilde{M}^T . \widetilde{M}}$$

Ces deux matrices sont carrées (de taille *N*x*N*), symétriques et réelles. Elles sont donc diagonalisables dans une base orthonormée [1].

2.3.6. Projection

Comme cité ci-dessus, l'analyse en composantes principales vise à projeter les données d un espace initial vers un second espace, de dimension inférieure, tel que son premier axe « *u* » soit issu d'une combinaison linéaire des *Xn*, de manière à ce que la variance de la projection des points du nuage sur cet axe soit maximale. La projection de l'échantillon des *X* sur *u* « π » s'écrit :

$$\pi_u(M) = M.u$$

La variance de cette projection « $\pi_u(M)$ »vaudra donc :

$$\pi_u(M)^T.\frac{1}{K.\pi_u(M)} = u^T.\underbrace{M^T.\frac{1}{K.M}}_{c}.\frac{1}{u}$$

où *C* est la matrice de covariance.

Ayant vu précédemment que « *C* » était diagonalisable dans une base orthonormée, c'est cette diagonalisation qui va nous permettre en outre de voir que la variance exprimée par le $k^{ème}$ vecteur propre valait *k*. Et finalement, que la question de PCA nous ramène à un problème de diagonalisation de la matrice de corrélation. On notera « *P* » la matrice de changement de base associée et la matrice diagonale formée de son spectre :

$$\pi_u(M).\frac{1}{K.\pi_u(M)} = u^T p^T \Delta p u = (pu)^T \Delta \underbrace{(pu)}_{v}$$

A partir de cette reformulation, on cherchera désormais le vecteur « *v* » maximisant *v*'* *v*, où $\Delta = \mathrm{Diag}(\lambda_1, \cdots, \lambda_N)$ est une matrice diagonale dont les valeurs sont rangées dans un ordre décroissant. Là, il sera très facile de constater que le premier vecteur unitaire vérifiera *v*'* *v*= λ . Formellement et mathématiquement, on va utiliser les multiplicateurs de Lagrange « α » pour prouver ce résultat, et ce en maximisant la variance des données projetées sur *u*

sous la contrainte que *u* soit de norme 1:

$$L(u,\alpha) = u^T.C.u - \alpha(u^T u - 1)$$

En résolvant cette équation de Lagrange, deux résultats nous importent : en premier lieu « *u* » est un vecteur propre de *C* associé à la valeur propre λ_1 ; et en suite, il est de norme « 1 ». La valeur propre de la matrice de covariance *C*: « λ_1» étant la variance sur le premier axe de PCA.

On poursuit la recherche du deuxième axe de projection *w* sur le même principe en imposant qu'il soit orthogonal à *u*, ainsi de suite, jusqu'à l'obtention des « l » vecteurs propres recherchés [1].

2.3.7. Les étapes de l'ACP

Une analyse en composante principale se décompose selon les étapes suivantes :

• **Etape 1 :** calcul de la matrice des corrélations entre les variables.

Cette matrice fournit les premiers éléments de description des associations existant entre les variables. L'ACP permettra d'obtenir une synthèse de ces liaisons.

• **Etape 2 :** calcul des valeurs propres associées à la matrice des corrélations

Les valeurs propres (ou inertie liée à un facteur) sont les variances des coordonnées de points individus sur l'axe correspondant. Ce sont donc des indices de dispersion du nuage des individus dans la direction définie par l'axe. Il est souvent intéressant de regarder la décroissance des valeurs propres. En effet, si les données sont peu structurées, le nuage a une forme « régulière » qui s'observe par une décroissance régulière des valeurs propres. Dans ces conditions, l'analyse factorielle ne fournira pas de résultat intéressant.

• **Etape 3 :** calcul des vecteurs propres associés

Ces vecteurs propres représentent les axes factoriels du nouvel espace, combinaison linéaire des variables initiales. On procède alors axe par axe à partir des vecteurs propres associées aux plus grande valeurs propres, pour définir les composantes principales. L'examen du plan factoriel permet de visualiser les corrélations entre les variables et d'identifier des groupes d'individus ayant pris les mêmes valeurs pour les même variables.

Algorithme de l'ACP

· Calcul de $\overline{M}$

· A=cov $\overline{M}$.

· Calcul des valeurs et vecteurs propres.

· Calcul de « l »-le nombre de composantes à garder.

· Calcul de la matrice caractéristique = $[eig_1, eig_2, ..., eig_l]$.

· Les données réduites= (la matrice caractéristique)' * $\overline{M}$.

· (Les données originales - la moyenne)' =(la matrice caractéristique)$^{-1}$ *les données réduites

· (Les données originales)' = (la matrice caractéristique)^{T} *les données réduites + la moyenne[1].

2.3.8. Les domaines d'application

De part la nature des données que l'ACP peut traiter, les applications sont très nombreuses. Il y a en fait deux façons d'utiliser l'ACP :

- soit pour l'étude d'une population donnée en cherchant à déterminer la typologie des individus et des variables. Par exemple, dans la biométrie, l'étude des mensurations sur certains organes peut faire apparaître des caractéristiques liées à des pathologies, ou encore en économie, l'étude des dépenses des exploitations par l'ACP peut permettre des économies de gestion.
- soit pour réduire les dimensions des données sans perte importante d'information, par exemple en traitement du signal et des images, où l'ACP intervient souvent en prétraitement pour réduire la quantité de données issues de traitements analogiques[20].

2.4. Classification

La classification est une méthode qui a pour but de grouper les individus d'une population en des classes disjointes telles que dans une même classe, les membres sont similaires et dans les classes différentes, les individus sont dissimilaires.

Il faut distinguer la classification (la classification non supervisée) avec le classement (la classification supervisée) [20] :

- **Classification supervisée:** Etant donné un ensemble des classes déjà identifiées et un individu, il s'agit de trouver la meilleure classe à laquelle cet individu appartient.

- **Classification non supervisée** : Etant donné un ensemble des individus, il s'agit de structurer des classes pas encore identifiées qui groupent ces individus.

2.4.1. Classification automatique

La classification automatique est de découper l'ensemble des données étudiées en un ou plusieurs sous-ensembles nommés classes, chaque sous-ensemble devant être le plus homogène possible. Les membres d'une classe ressemblent plus aux autres membres de la même classe qu'aux membres d'une autre classe. Deux types de classification peuvent être relevés : d'une part la classification (partitionnement ou recouvrement) « à plat » et d'autre part le partitionnement hiérarchique. Dans les deux cas, classifier revient à choisir une mesure de (la similarité/dissimilarité), un critère d'homogénéité, un algorithme, et parfois un nombre de classes composant la partition[21]

La ressemblance (similarité ,dissimilarité) des individus est mesurée par un indice de similarité, un indice de dissimilarité ou une distance. Par exemple, pour des données binaires l'utilisation des indices de similarité tels que l'indice de Jaccard, l'indice de Dice, l'indice de concordance ou celui de Tanimoto est fréquente. Pour des données quantitatives, la distance euclidienne est la plus appropriée, mais la distance de Mahalanobis est parfois adoptée. Les données sont soit des matrices de p variables qualitatives ou quantitatives mesurées sur n individus, soit directement des données de distances ou des données de dissimilarité.

Le critère d'homogénéité des classes est en général exprimé par la diagonale d'une matrice de variances-covariances (l'inertie) inter-classes ou intra-classes. Ce critère permet de faire converger les algorithmes de ré-allocation dynamiques qui minimisent l'inertie intra-classe ou qui maximisent l'inertie inter-classes.

Les principaux algorithmes utilisent la ré-allocation dynamique en appliquant la méthode de B.W. Forgy des centres mobiles, ou une de ses variantes : la méthode des k-means, la méthode des nuées dynamiques, ou PAM (« Partitioning Around Medoids (PAM) »). Les méthodes basées sur la méthode de Condorcet, l'algorithme espérance-maximisation, les densités sont aussi utilisées pour bâtir une classification.

Il n'y a pas de classification meilleure que les autres, en particulier lorsque le nombre de classes de la partition n'est pas prédéterminé. Il faut donc mesurer la qualité de la classification et faire des compromis. La qualité de la classification peut se mesurer à l'aide de l'indice R^2 qui est le rapport de l'inertie inter classe sur l'inertie totale, calculé pour plusieurs valeurs du nombre de classe total, le compromis étant obtenu par la méthode du coude

L'interprétation des classes, permettant de comprendre la partition, peut s'effectuer en analysant les individus qui composent chaque classe. Le statisticien peut compter les individus dans chaque classe, calculer le diamètre des classes la distance maximum entre individus de chaque classe.

Il peut identifier les individus proches du centre de gravité, établir la séparation entre deux classes - opération consistant à mesurer la distance minimum entre deux membres de ces classes. Il peut analyser aussi les variables, en calculant par exemple la fréquence de certaines valeurs de variables prises par les individus de chaque classe, ou en caractérisant les classes par certaines valeurs de variables prises par les individus de chaque classe

2.4.2. Classification hiérarchique

Les données en entrée d'une classification ascendante hiérarchique (CAH) sont présentées sous la forme d'un tableau de dissimilarités ou un tableau de distances entre individus.
Il a fallu au préalable choisir une distance (euclidienne, Manhattan, Tchebychev ou autre) ou un indice de similarité (Jacard, Sokal, Sorensen, coefficient de corrélation linéaire, ou autre).
La classification ascendante se propose de classer les individus à l'aide d'un algorithme itératif. À chaque étape, l'algorithme produit une partition en agrégeant deux classes de la partition obtenue à l'étape précédente.

Le critère permettant de choisir les deux classes dépend de la méthode d'agrégation. La plus utilisée est la méthode de Ward qui consiste à agréger les deux classes qui font baisser le moins l'inertie interclasse D'autres indices d'agrégation existent comme celui du saut minimum (« single linkage ») où sont agrégées deux partitions pour lesquelles deux éléments - le premier appartenant à la première classe, le second à la seconde - sont le plus proches selon la distance prédéfinie, ou bien celui du diamètre (« complete linkage ») pour lequel les deux classes à agréger sont celles qui possèdent le couple d'éléments le plus éloigné

L'algorithme ascendant se termine lorsqu'il ne reste qu'une seule classe.
La qualité de la classification est mesurée par le rapport inertie inter-classe sur inertie totale.

Des stratégies mixtes, alliant une classification « à plat » à une classification hiérarchique, offrent quelques avantages. Effectuer une CAH sur des classes homogènes obtenus par une classification par ré-allocation dynamique permet de traiter les gros tableaux de plusieurs milliers d'individus, ce qui n'est pas possible par une CAH seule. Effectuer une CAH après un échantillonnage et une analyse factorielle permet d'obtenir des classes homogènes par rapport à l'échantillonnage [22]

2.5. Régression

Il est intéressant de noter, en s'inspirant de ce qu'écrivent Henry Rouanet et ses coauteurs, que l'analyse des données descriptive et l'analyse prédictive peuvent être complémentaires, et parfois produire des résultats similaires [23].

Approche PLS (Partial Least Squares régression)

L'approche PLS est plus prédictive que descriptive, mais les liens avec certaines analyses que l'on vient de voir ont été clairement établis.

L'algorithme d'Herman Wold, nommé tout d'abord NILES (« Nonlinear Estimation by Iterative Least SquareS »), puis NIPALS (« Nonlinear Estimation by Iterative Partial Least SquareS ») a été conçu en premier lieu pour l'analyse en composantes principales.

En outre, PLS permet de retrouver l'analyse canonique à deux blocs de variables, l'analyse inter batteries de Tucker, l'analyse des redondances et l'analyse canonique généralisée au sens de Carroll. La pratique montre que l'algorithme PLS converge vers les premières valeurs propres dans le cas de l'analyse inter batteries de Tucker, l'analyse canonique à deux blocs de variables et l'analyse des redondances [24].

La régression sur composantes principales (PCR) utilise l'ACP pour réduire le nombre de variables explicatives en les remplaçant par les composantes principales qui ont l'avantage de ne pas être corrélées. PLS et PCR sont souvent comparées l'une à l'autre dans la littérature.

2.5.1. Description du capteur logiciel

La technique de mise en œuvre de capteur logiciel est basée sur deux opérations, la première faire une réduction de dimension de la base de données réelle qui déjà élaborée, donc on à acquérir une nouvelle base de donner de dimension réduire. Et la deuxième on à exploit le réseau de neurones artificiel pour faire l'apprentissage et les tests après la partage de la base de données nouvelle, et la figure présent les défirent étapes [ ]:

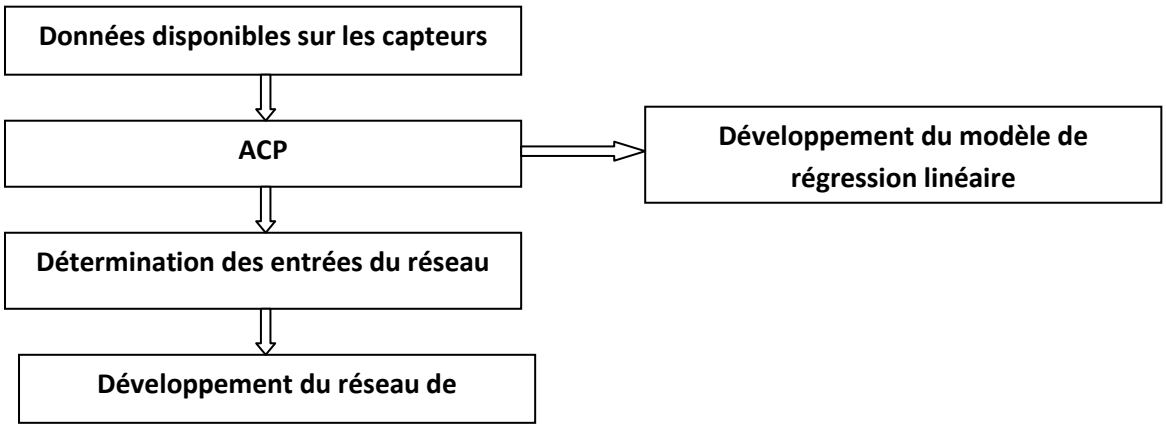


Fig. 2. 1 : Méthode pour le capteur logiciel basé sur l'ACP et les RNA.

Conclusion

Dans ce chapitre, nous avons parlé sur l'analyse multivariée des données et leur objectif ainsi quelques méthodes d'analyse ; on a insisté sur la méthode d'analyse en composantes principales qui est choisie dans notre approche. L'ACP est une méthode de traitement statistique de données dont le but est de représenter et d'expliquer les liaisons statistiques entre les phénomènes. Elle permet d'identifier des variables sous-jacentes, ou facteurs qui expliquent les corrélations à l'intérieur d'un ensemble de variables observées. Elle est souvent utilisée pour réduire un ensemble de données, et dans l'agrégation de l'information, en identifiant un petit nombre de facteurs qui expliquent la plupart des variances observées dans le plus grand nombre de variables manifestes. On peut également utiliser l'analyse factorielle pour résumer, synthétiser, et hiérarchiser l'information acquise pour l'étape de décision. Nous avons introduit quelques méthodes utilisées dans le cadre de régression et classification de données telles que les réseaux de neurones, comme des fonctions de décision dans un système de traitement à fusion multisensorielle. Une étude détaillée des mécanismes de ces méthodes appliquées à la régression et de classification de données fera l'objet du chapitre suivant.

CHAPITRE III

LES RESEAUX DE NEURONES ARTIFICIELS

Introduction

Le cerveau humain est capable de s'adapter, d'apprendre et de décider, et c'est sur ce fait que des chercheurs se sont intéressés à comprendre son principe de fonctionnement et de pouvoir l'appliquer au domaine de l'informatique. C'est ainsi que dans les années cinquante on formalisa le neurone en un modèle mathématique à partir du modèle biologique.

Nous allons présentés dans ce chapitre, une étude générale sur les réseaux de neurones, le passage du modèle biologique au modèle artificiel, le protocole d'apprentissage, ainsi les différentes classes et architectures de ces réseaux. Puis nous consacrerons une étude sur les réseaux de neurones de type RBF (Radial Basis Functions), qui représente le modèle adéquat pour notre travail.

3.1. Etat de l'art

Avec l'apparition des ordinateurs l'homme a découvert un moyen d'effectuer diverses tâches avec deux capacités non négligeables que lui ne possède pas : la rapidité et la précision.

Cependant l'exécution d'une tâche pour l'ordinateur nécessite sa programmation préalable par l'homme. Cette caractéristique fait apparaître les ordinateurs comme des machines exécutant des ordres « aveuglement » et l'homme n'a pas désespéré de voir un jour construire une machine à son image, c'est à dire intelligente, capable d'apprendre, de raisonner, de réfléchir sans son intervention. Ce sont des recherches basées sur le fonctionnement du cerveau qui ont constituées le point de départ des différentes recherches. Des travaux de neurobiologistes ont, en effet, révélé que le cerveau est constitué d'un nombre extrêmement élevé d'unités de traitement élémentaire de l'information (les neurones biologiques) fortement interconnectées. L'information contenue dans le cerveau est stockée

dans les connexions entre les neurones et c'est la coopération entre les neurones, qui effectuent un traitement fortement parallèle et distribué, qui donne sa puissance au cerveau [1].

3.2. Systèmes Nerveux

Le cerveau humain, est le meilleur modèle de machine, polyvalente incroyablement rapide et surtout douée d'une incomparable capacité d'auto-organisation. Son comportement est beaucoup plus mystérieux que le comportement de ses cellules de base. Il est constitué d'un grand nombre d'unités biologiques élémentaires (environ 10^{12} neurones), chacune reçoit et envoie des informations (1000 à 10000 synapse par neurone).

Les cellules nerveuses appelées " neurones ", sont les éléments de base du système nerveux centrale. Elles sont constituées de trois parties essentielles : le corps cellulaire, les dendrites et l'axone (figure 3.1**)**.

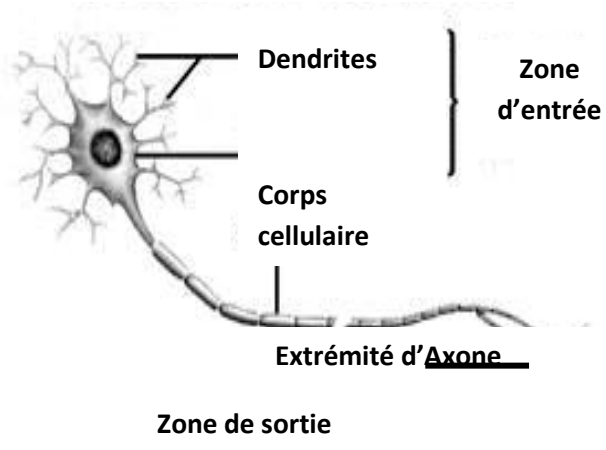


Fig. 3.1 : Le neurone biologique.

3.3. Le neurone formel

Le neurone formel est le modèle mathématique du neurone biologique. Son travail consiste a faire une sommation pondérée de ses entrées provenant de l'extérieur ou de la sortie d'un autre neurone artificiel, le résultat obtenu est ensuite calculé en utilisant une fonction non linéaire, appelée fonction de seuil. La figure 2.3, montre le modèle d'un neurone artificiel.

L'élément de base d'un réseau de neurones est, bien entendu, le neurone artificiel. Un neurone contient deux éléments principaux:

• un ensemble de poids associés aux connexions du neurone, et

• une fonction d'activation[25].

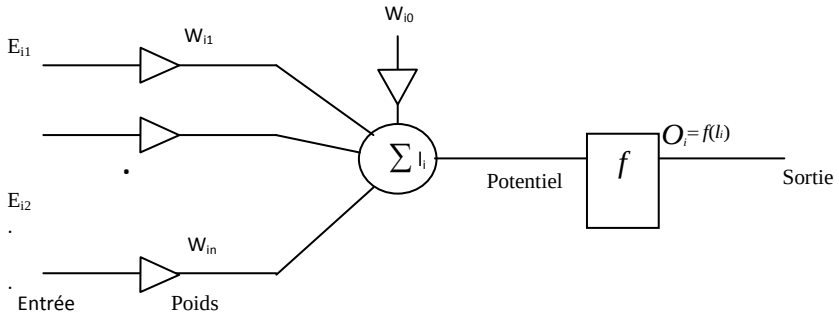


Fig. 3.2. Le neurone formel.

Le choix de la fonction d'activation dépend de l'application. S'il faut avoir des sorties binaires, c'est la première fonction que l'on choisit habituellement [25].

3.3.2. Principe de fonctionnement

L'équation de sortie (O_i) du neurone i est donnée par : $O_i = f(I_i)$

Où :

$I_i = \sum W_{ij}^K E_{ij} - W_{i0}^K E_{i0}$ W_{ij} sont les coefficients de pondération appelés aussi coefficients synaptiques (ou poids), de la $j^{éme}$ entrée (E_j), du neurone i, dans la couche *k*. La somme pondérée est appelée potentiel somatique. L'entrée E_0 pondérée par le poids W_{i0} est considérée comme la valeur de seuil interne du neurone *i*.

3.3.3. Fonction d'activation ou de seuillage

C'est une fonction non linéaire appelée aussi fonction de seuil, elle présente la relation qui lie la fréquence moyenne des potentiels d'action (Y_j) limités en amplitude, au potentiel somatique (S_j). Le potentiel d'action (Y_j) sert ensuite à exciter les autres neurones qui lui sont connectés.

La transformation non linéaire appliquée au potentiel somatique, traduit deux caractéristiques importantes du neurone :

- ✓ l'effet du seuil.

- ✓ la saturation de la réponse au de là d'une certaine valeur du potentiel somatique.

La fonction d'activation *F* est de nature très variante, elle peut être déterministe, continue, discontinue ou aléatoire.

La Fig.3.3 donne quelques modèles des fonctions d'activation utilisées.

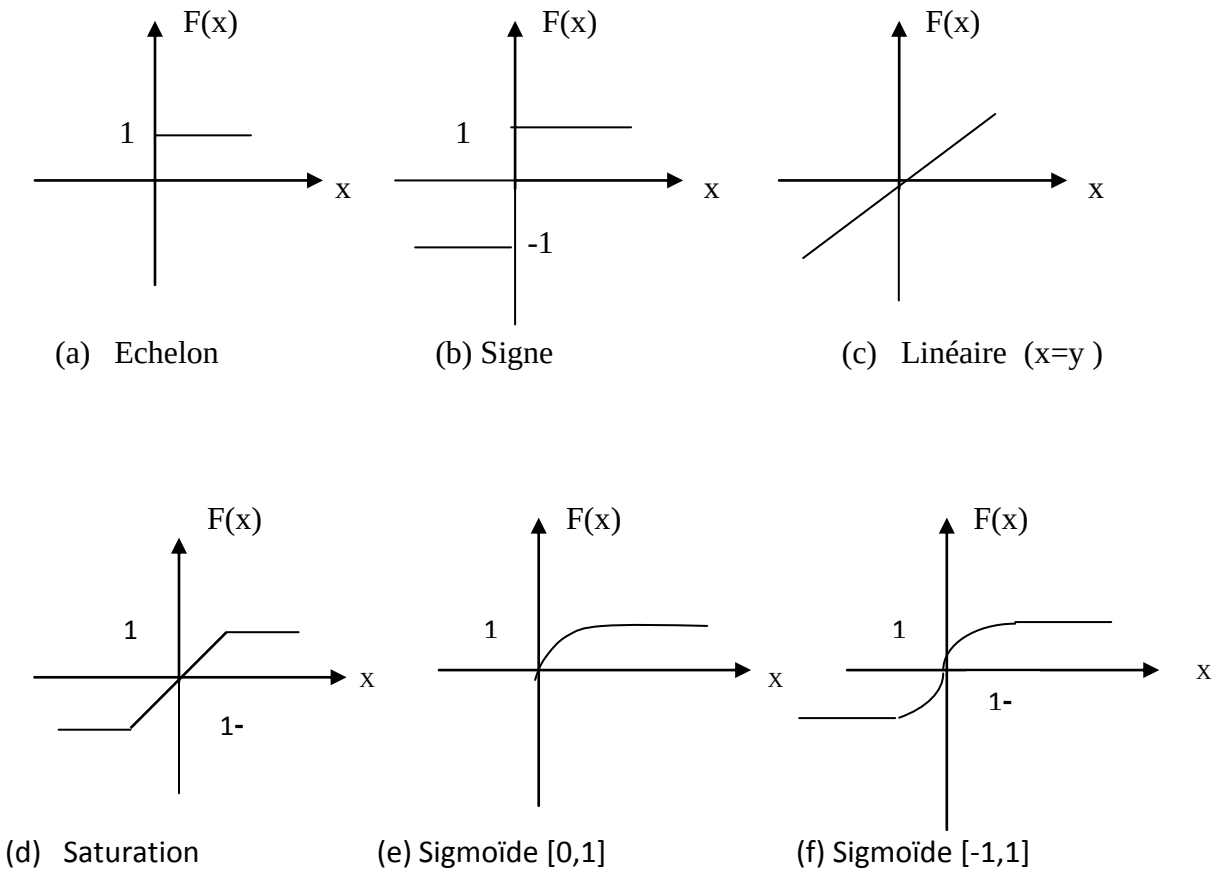


Fig. 3.3. Les fonctions d'activation les plus utilisées.

Toutes les fonctions d'activation utilisées doivent être différentiables, car l'architectures du réseaux de neurones l'impose pour que l'apprentissage soit possible.

3.4. Structure des connexions

Les connexions entre les neurones qui composent le réseau décrivent la " topologie " du modèle. Elles sont très variées, le nombre de connexions étant énorme. Cette topologie fait apparaître une certaine régularité de l'arrangement des neurones.

Il existe de nombreuses topologies, nous citerons quelques-unes dans ce qui suit :

3.4.1. Réseau multicouches (classique)

Les neurones sont arrangés par couche, les entrées des neurones de la deuxième couche sont en fait les sorties des neurones de la couche amont. Les neurones de la première couche sont reliés au monde extérieur et reçoivent le vecteur d'entrée. Il peut y avoir une ou plusieurs sorties à un réseau de neurone.

Dans un réseau multicouche classique, il n'y a pas de connexion entre neurones d'une même couche et les connexions ne se font qu'avec les neurones de la couche aval, et tous les neurones de la couche amont sont connectés à tous les neurones de la couche aval. On appelle :

-**couche d'entrée :** contient l'ensemble des neurones d'entrées, cette couche est une couche passive, ses neurones n'effectuent aucun traitement

-**couche de sortie :** contient l'ensemble des neurones de sorties

-couches cachées : les couches intermédiaires n'ayant aucun contact avec l'extérieur.

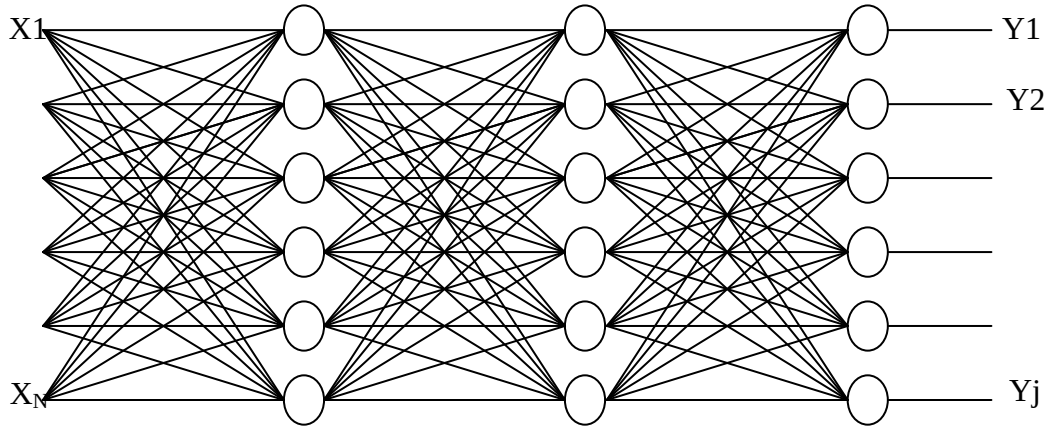


Fig. 3.4. Réseau multicouches.

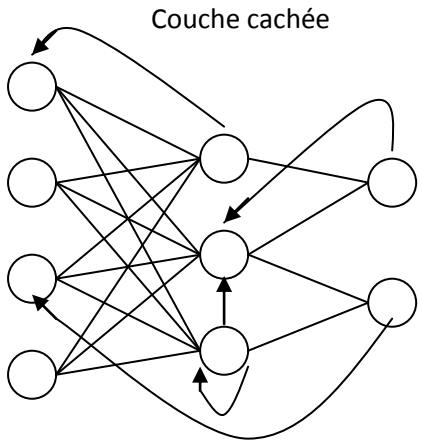
3.4.2. Réseau à connexions locales

C'est aussi un réseau multicouche, mais tous les neurones d'une couche amont ne sont pas connectés à tous les neurones de la couche aval. Nous avons donc dans ce type de réseau de neurones un nombre de connexions moins important que dans le cas du réseau de neurones multicouche classique.

3.4.3. Réseau à connexions récurrentes

Un réseau de ce type signifie qu'une ou plusieurs sorties de neurones d'une couche aval sont connectées aux entrées des neurones de la couche amont ou de la même couche. Ces connexions récurrentes ramènent l'information en arrière par rapport au sens de propagation défini dans un réseau multicouches[26].

La rétroaction de la sortie vers l'entrée permet à un réseau de ce type de présenter un comportement temporel.



F**ig. 3.5.** Réseau a connexion récurrente.

3.4.4. Les réseaux entièrement connectés (complexe)

Chaque neurone est connecté à tous les neurones du réseau y compris lui-même, c'est la structure d'interconnexion la plus générale. L'exemple de cette structure est le réseau de " kohonen " (Fig 3.6).

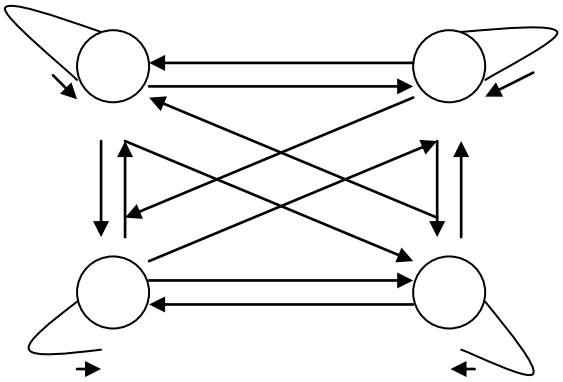


Fig. 3.6. Réseau de kohonen.

3.5. L'apprentissage dans les réseaux de neurones

On peut considéré les réseaux de neurones comme une boite noire contenant l'information qu'elle doit apprendre et mémoriser. Mais au démarrage lorsque on choisit notre réseau, la boite noire est vide et ne contient aucune information, ni aucune connaissance sur son sujet, c'est pourquoi un apprentissage est nécessaire. L'enseignement que doit subir le réseau de neurones est un apprentissage qui est une phase du développement d'un réseau de neurones durant laquelle le comportement du réseau est modifié jusqu'à l'obtention du comportement désiré. L'apprentissage neuronal fait appel à des exemples de comportements .

L'apprentissage des réseaux de neurones consiste à adapter ses différents paramètres (poids) d'après un algorithme itératif d'ajustement ou d'adaptation lui permettant de prendre en considération toutes les données (exemples) qui lui sont fournies à son entrée et ainsi ajuster ses paramètres pour trouver le juste milieu permettant de prendre en charge n'importe quel exemple ou donnée apparaissant à son entrée provenant de son environnement .

Le mécanisme d'apprentissage d'un réseau comprend la récurrence des phases suivantes [27] :

· Le réseau est stimulé par l'environnement,

· En réponse à cette stimulation, le réseau adapte son comportement,

· Le réseau réagit alors différemment à l'environnement en fonction de la nouvelle expérience acquise consécutivement à la stimulation.

L'apprentissage des réseaux dépend des informations qui lui sont fournies à son entrée et à sa sortie. En tenant compte de la façon dont on peut lui faire fournir les informations et lui faire apprendre à les assimiler, c'est à dire le guider ou non durant son apprentissage. De la, il apparaît deux types d'apprentissages : l'apprentissage supervisé et l'apprentissage non supervisé [28].

Les algorithmes d'apprentissages donnent de meilleurs résultats lorsqu'on leur fournit des exemples multiples et variés ; ainsi le réseau peut assimiler toutes les connaissances. Ils existent différente règles d'apprentissages parmi lesquelles on peut distinguer [28]:

- la règle de Widrow-Hoff,
- la règle de Hebb,
- la règle du Perceptron,
- la règle de Grossberg, etc...

3.5.1. Apprentissage supervisé

Les réseaux multicouches avaient déjà été définis par Rosenblatt, mais on ne savait pas comment faire l'apprentissage. Avec la découverte de l'algorithme de rétropropagation de l'erreur (RP) par Rumelhart et al, on a commencé à faire de l'apprentissage des réseaux de neurones multicouches à partir d'exemples. Cette méthode de détermination des poids est appelée apprentissage supervisé

L'apprentissage supervisé "Supervised Learning", repose sur le fait que les exemples sont des couples (Entrée, Sortie associée), Fig3.7. C'est à dire que l'on suppose l'existence d'un expert qui prend en charge la sortie de notre réseau en lui fournissant une sortie désirée et les associe aux sorties réelles fournies par le réseau d'après les données à l'entrée. Le réseau adapte ses paramètres en fonction de la différence qui existe entre la sortie réelle et la sortie désirée en prenant compte de tous les exemples de l'environnement

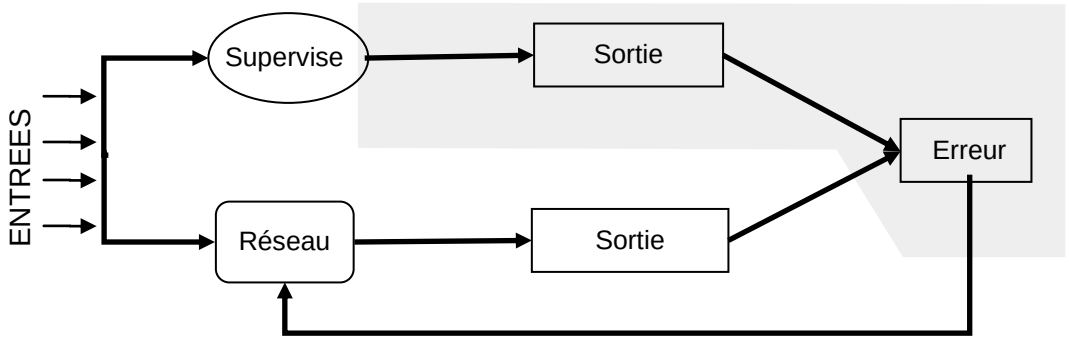


Fig.3.7. Illustration de l'apprentissage supervisé.

3.5.2. Apprentissage non supervisé

La différence majeure entre l'apprentissage supervisé et non supervisé peut être résumée dans le fait que le deuxième type d'apprentissage est autodidacte qui n'a pas besoin d'expert pour le guider à adapter ses paramètres mais qu'il s'auto adapte alors qu'il ne dispose que des valeurs (Entrée), Fig.3.8. Remarquons cependant que les modèles d'apprentissage non supervisé nécessite avant la phase d'utilisation une étape de labellisation effectuée par l'opérateur, qui n'est pas autre chose qu'une part de supervision

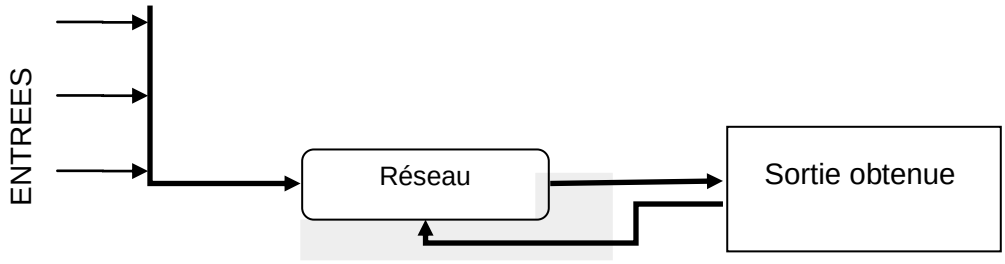


Fig.3.8. Illustration de l'apprentissage non supervisé

3.6. Le Perceptron Multicouches

Comme nous l'avons déjà dit le cerveau humain est composée de milliers de neurones, alors il est évident qu'un simple neurone, ne peut rien faire à lui tout seul, il lui faut la coopération d'autres neurones. En suivant ce résonnement, il est évident qu'il vaut trouver

une architecture qui relie les neurones entre eux, qui crée une liaison entre les neurones pour créer un réseau de neurones [29].

En s'inspirant du perceptron monocouche, une architecture plus complexe englobant plusieurs neurones a été mise au point. Cette nouvelle architecture est le perceptron multi-couches (PMC ou MLP pour Multi Layer Perceptron en anglais). L'apparition de cette architecture a permis de résoudre les problèmes de classification non linéaire du perceptron et de dépasser les limites principales de celui-ci.

3.6.1. Architecture du PMC

La sortie désirée de notre réseau est notée d_k. Le réseau possède N_0 entrée, L-1 couches cachées contenant chacune N_i neurones ($1<i<L-1$) et une couche de sortie avec N_L neurones. Chaque neurone est caractérisé par un couple d'indice (j,k), où j désigne le nombre de couches et k le nombre de neurones, (Fig. 3.9) Le coefficient synaptique est désigné par w_{jki} où le troisième indice i indique le numéro des neurones transmetteurs. Le signal y_{jk} est la somme pondérée de toutes les entrées du neurone jk. x_{jk} est la sortie non linéaire du neurone jk et f est une fonction d'activation choisit comme suit : $f(y)=\dfrac{1-\exp(-a.y)}{1+\exp(-a.y)}$ où 'a' représente le seuil de la fonction d'activation.

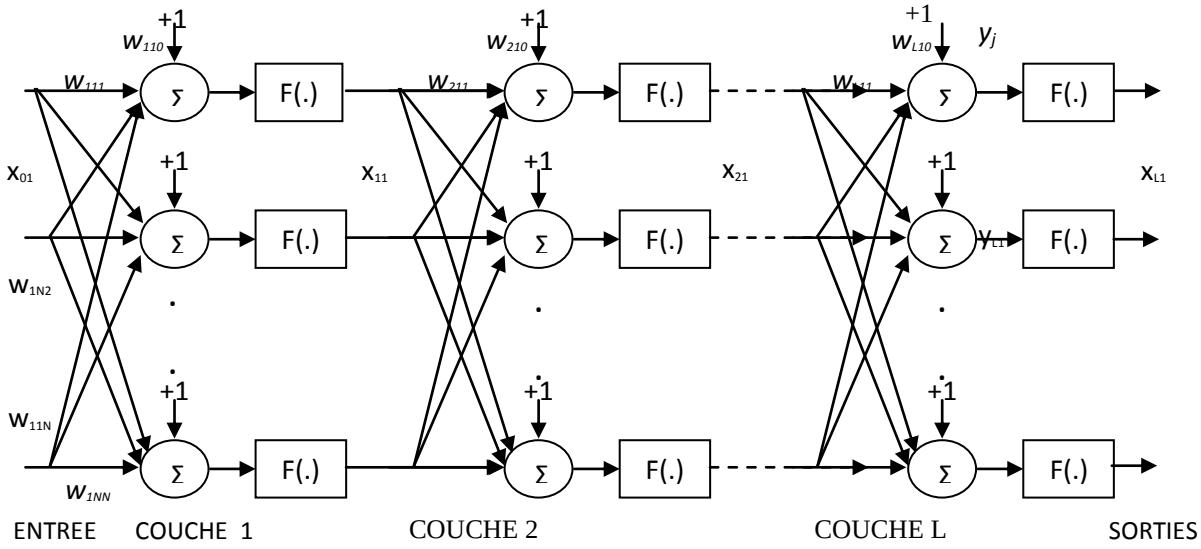


Fig. 3.9. Architecture MLP.

3.6.2. Réseau de neurones de type RBF (Radial Basis Functions)

Les réseaux à fonctions radiales de base (RBF) sont des modèles connexionnistes simples à mettre et œuvre et assez intelligible, sont très utilisés pour la régression et la discrimination.

Leur propriétés théoriques ont été étudiées en détail depuis la fin des années 80 ; il s'agit certainement, avec le perceptron multicouche, du modèle connexionniste le mieux connu.

3.6.2. 1. Architecture

Introduit par Powell et Broomhead , le réseau RBF (Radial Basis Functions) fait partie des réseaux de neurones supervisés. Il est constitué de trois couches (Fig.3.10): une couche d'entrée qui retransmet les entrées sans distorsion, une seule couche cachée qui contient les neurones RBF qui sont généralement des gaussiennes et une couche de sortie dont les neurones sont généralement animés par une fonction d'activation linéaire. Chaque couche est complètement connectée à la suivante et il n'y a pas de connexions à l'intérieur d'une même couche.

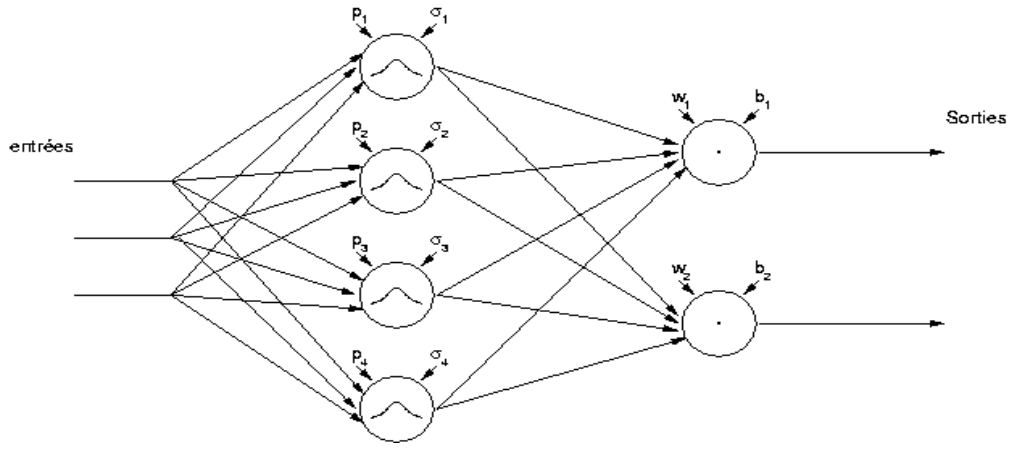


Fig.3.10. Présentation schématique d'un réseau RBF.

3.7. Algorithme de rétropropagation du gradient

L'algorithme de rétropropagation (Backpropagation / BP) est le plus utilisé dans l'apprentissage [29] des réseaux. Il consiste :

1- Initialiser les coefficients synaptiques avec des valeurs petites et aléatoires

2- Présenter au réseau les exemples à apprendre. Choisir un exemple et procéder à la propagation et la rétropropagation (étapes 3et4)

3- La propagation : l'activation des neurones se propage de la couche d'entrée à la couche de sortie.

Les activités internes du neurone k dans la couche j sont :

$$y_{jk} = \sum_{i=0}^{N_{j-1}} x_{j-1,i} w_{jki}$$

la sortie non linéaire est calculée d'après :

$$x_{jk} = f(y_{jk}) = \frac{1-\exp(-a.y_{jk})}{1+\exp(-a.y_{jk})}$$

4- rétropropager l'erreur : On calcule la dérivée de $f(y_{jk})$ en utilisant :

$$f'(y_{jk}) = \frac{2a[\exp(-a.y_{jk})]}{[1+\exp(-a.y_{jk})]^2}$$

on calcule l'erreur à la couche de sortie (j=L) en évaluant la valeur suivante :

$e_{Lk} = f'(y_{Lk}).(O_k - x_{Lk})$ pour chaque neurone k, où O_k est la $k^{éme}$ sortie désirée.

Après cela, on calcule le signal d'erreur dans les couches cachées pour toute valeur de j=L-1 jusqu'à 1 par :

$$e_{jk} = f'(y_{jk}) \sum_{i=1}^{N_{j+1}} e_{j+1,k} w_{j+1,k,i}$$

5- adapter les coefficients synaptiques par :

$w_{jki}(n+1)=w_{jki}(n)+\mu e_{jk}(n)x_{j-1,i})+\alpha(w_{jki}(n)-w_{jki}(n-1))$

μ : est le pas d'apprentissage

α : le momentum

6- présenter les paramètres pour une nouvelle itération jusqu'à que les coefficients synaptiques se stabilisent autour d'une valeur et l'erreur quadratique totale du réseau $E=\sum_{p=1}^{m}\sum_{k=1}^{N_L}(O_k^p-x_{Lk}^p)^2$ soit inférieure à un certain seuil, m est le nombre totale des paramètres d'apprentissage. O_k^p et x_{Lk}^p représente respectivement la k^{éme} sortie désirée et la k^{éme} sortie du réseau pour la p^{éme} paramètre.

En plus, il est possible d'arrêter l'apprentissage en fixant une limite au nombre d'itérations. Noter que l'ordre de présentation des exemples doit être aléatoire.
Généralement le pas d'apprentissage et le momentum doivent être adaptés quand le nombre d'itération augmente.

3.8. Mise en ouvre d'algorithme RNAs

Un réseau de neurones artificiels définit une famille de fonctions. L'apprentissage consiste à déterminer la solution du problème posé par cette famille, ces fonctions pourraient avoir des capacités limitées. Le principe de l'apprentissage est l'optimisation d'une fonction de coût qui représente le but d'apprentissage. Les méthodes numériques utilisées sont le plus souvent des méthodes approchées basées sur des techniques de gradient (parce qu'on ne sait pas résoudre analytiquement un système d'équations non linéaires).

3.8.1. Apprentissage

Nous avons posé précédemment le problème de l'apprentissage par réseau de neurones artificiels comme un problème d'optimisation d'une fonction de coût. L'algorithme de rétropropagation de l'erreur est parmi le plus utilisé dans le cas d'un problème de classification supervisée.

Les paramètres d'entrée du programme d'apprentissage sont les suivantes :
- Base de données. Vecteurs d'entrée et la classe correspondante ;

-Les poids, les biais initiaux ;

- La fonction d'activation ;

- Le nombre d'itérations.

La structure générale du programme d'apprentissage de l'algorithme de rétropropagation de l'erreur suit les étapes suivantes :

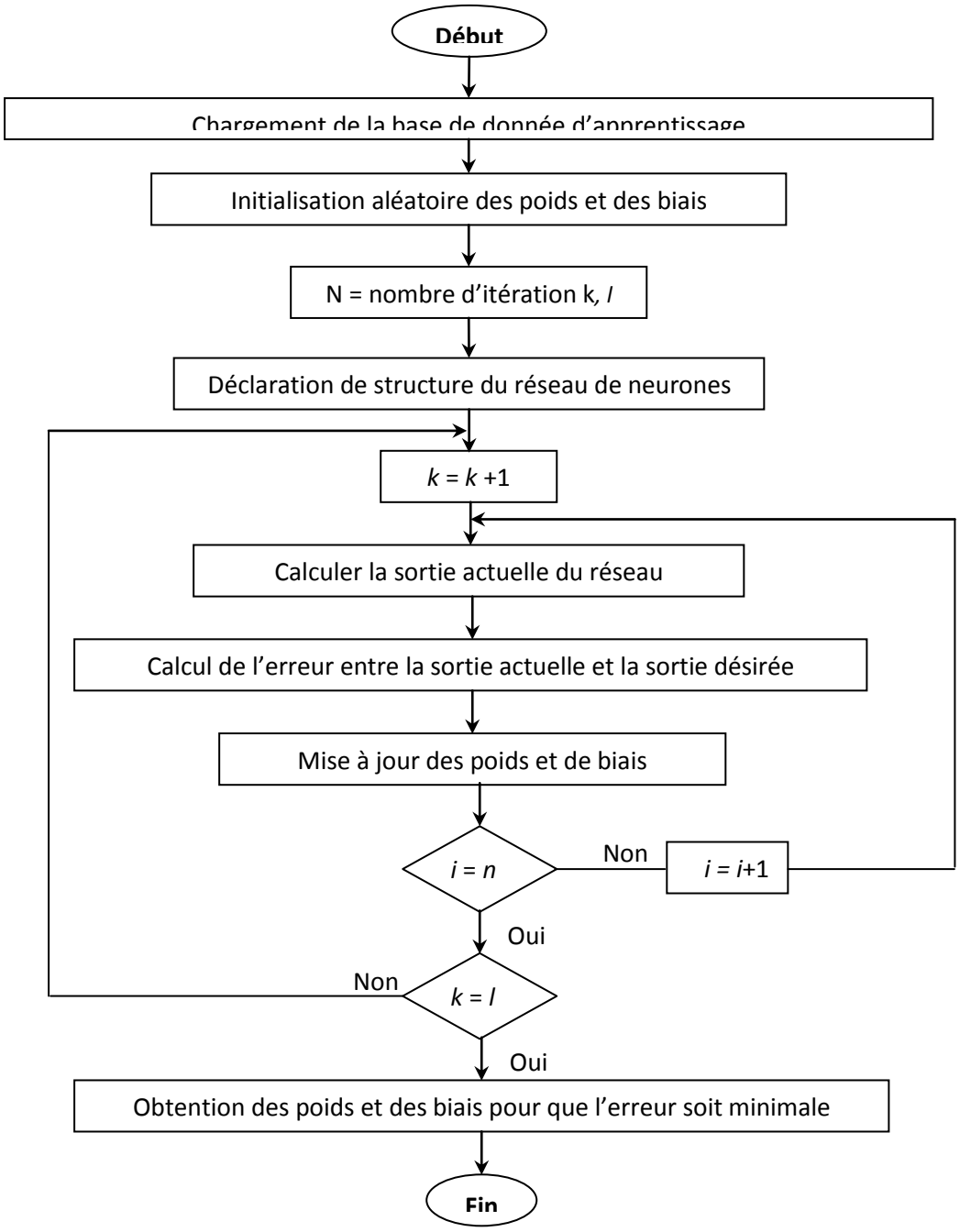


Fig. 3.11 : Structure générale du programme d'apprentissage par réseau de neurones.

Les sorties du programme sont : Les poids finals ,Les biais finals , Les sorties du réseau . Et La structure finale du réseau (nombre de couche et le nombre de neurone pour chaque couche). Le temps de calcul, ainsi que l'information générale comme l'erreur d'apprentissage.

3.8.2. Généralisation

La validation de l'algorithme de généralisation s'appuie sur la programmation de la première étape de l'apprentissage qui est la propagation des vecteurs de test comme une entrée au réseau adopté. En fixant la structure du réseau et leurs paramètres (poids, biais, fonction d'activation, nombre de couches cachées, le nombre de neurones correspondant), une fois son apprentissage est achevé. Et puis en testant le réseau sur des données qui n'ont pas servi à l'apprentissage.

Nous avons donc, pour le programme de généralisation, les paramètres suivants :
- La base d'exemples à classifier ;
-Les poids, les biais, la fonction d'activation à celle obtenu par apprentissage ;
-La structure finale du réseau après l'apprentissage.

La structure générale du programme de généralisation (test) suit les étapes suivantes :

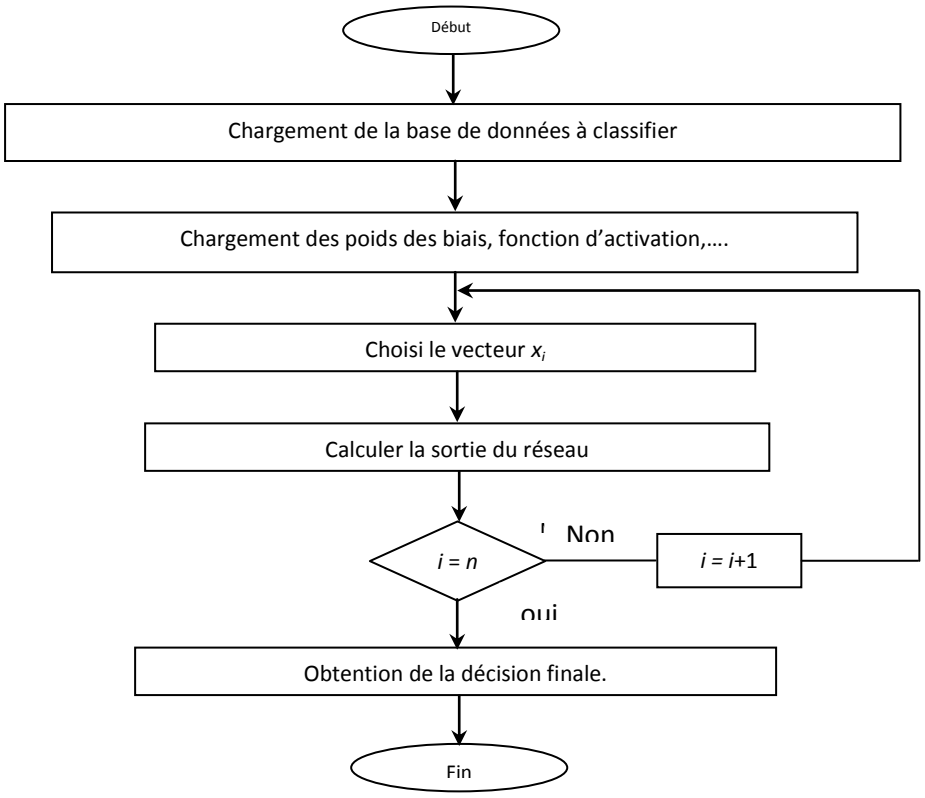


Fig.3.12. Structure générale du programme de généralisation par réseau de neurones.

Les sorties de généralisation représentent les classes des exemples évaluées à partir de la fonction de décision. Comme dans l'apprentissage, le temps de calcul approximatif est obtenu lors de l'exécution [1].

Dans la phase d'apprentissage par les RNAs, on découple la recherche de l'architecture de la détermination de ses paramètres, il faut chercher les paramètres de plusieurs structures afin de choisir celle qui garantit le meilleur pouvoir de généralisation. Ceci implique la partition de la base de données en une base d'apprentissage et une base de test.

3.9. Application Des Réseaux De Neurones

Les réseaux de neurones servent dans aujourd'hui à toutes sortes d'applications dans divers domaines. On peut cité par exemples :

- Autopilotage des avions.
- Système de guidage des automobiles.
- Lecture automatique des cheques bancaires et d'adresses postales.
- Production des systèmes de traitement signal et pour la synthèse de la parole.
- Les réseaux de neurones ils sont utilisés aussi pour les systèmes de vision par ordinateur.
- Ils sont utilisés en robotique et en télécommunication.
- Ils sont aussi utilisés dans les domaines de finance.
- Ils sont utilisés pour les diagnostiques médical [30].

Conclusion

Dans ce chapitre, nous avons essayé de donner en bref une description sur les réseaux de neurones artificiels, et aussi les différents types d'architectures et de modèle qui existent particulièrement l'architecture de type RBF. Nous avons aussi présenté une définition de l'apprentissage des réseaux de neurones, puis une description générale de l'algorithme de rétropropagation, aussi mise en ouvre d'algorithme RNAs et aussi les domaines d'utilisation des réseaux de neurones.

L'étude et l'analyse des performances des cette architecture RBF choisie après une analyse multivariée de données appliquée au domaine de contrôle et de surveillance des eaux brutes, constituent notre principal objectif. Une étude en simulation ayant pour but d'évaluation des performances de cette technique hybride appliquée à la régression et classification fera l'objet du chapitre suivant. L'évaluation des résultats, reflétant les performances obtenues, nous conduira à la validation de notre choix de la méthode choisie pour notre application. L'architecture du système de contrôle et de surveillance proposé est aussi présentée.

CHAPITRE IV

SIMULATION ET EVALUATION

Introduction

Le présent chapitre est dédié spécifiquement à la mise en œuvre de la technique de réseaux de neurones artificiels (RNAs) de type RBF en temps que technique de reconnaissance de formes appliquée au domaine de contrôle et de surveillance des eaux brutes; cette étude a été précédée d'une analyse statistique multivariée par la méthode de l'Analyse en Composante Principale « ACP » qui nous permettra de faire la réduction de dimension par la détermination des corrélations existantes entre les variables physico-chimiques de l'eau.

Une étude en simulation permettra de valider et d'évaluer les performances de méthode présentée, les impératifs principaux d'efficacité sont formulés sur deux points essentiels : les tests de spécification qui vérifient que le programme réalise bien la tâche pour laquelle il a été conçu concernant la réduction de dimension et la reconnaissance, ainsi que les tests de performances qui vont servir à mesurer l'efficacité avec laquelle ces tâches ont rempliées. Après cette vérification on évaluera les paramètres liés au taux de reconnaissance, au temps d'apprentissage et à l'erreur d'entraînement. Dans ce contexte, nous allons montrer et discuter les différentes simulations effectuées. Les résultats obtenus grâce à ces simulations nous permettront de valider les concepts théoriques des chapitres 2 et 3.

4.1. Problématique

Divers procédés de contrôle et de surveillance automatique modernes ont été développés ces dernières années ; parmi ces procédés, on retrouve les méthodes basées sur

l'intelligence artificielle (*IA*) et qui servent comme un outil de base pour l'aide à la décision. La démarche de contrôle et de surveillance par reconnaissance de formes est considérée comme une technique moderne. Cette démarche repose sur les méthodes de régression ou de classification. La technique la plus utilisée dans plusieurs domaines d'application est la reconnaissance des formes par réseaux de neurones artificiels (*RNAs*).

La technique citée auparavant permet de résoudre le problème de reconnaissance de formes posé, tel que dans notre application, entre autre le contrôle et la surveillance des eaux brutes. Dans cette application, l'approche de contrôle et de surveillance ne s'applique en fait que si on se trouve dans le cas d'un apprentissage supervisé. Nous procédons alors lors d'une étape préliminaire d'apprentissage, à paramétrer le modèle pour la reconnaissance. Ce module d'apprentissage permet de collecter de manière continue les paramètres relatifs aux différents états de l'eau. La reconnaissance des formes par réseaux de neurones artificiels possèdent l'avantage de couvrir plusieurs applications et de décrire des relations non linéaires entre les variables d'entrée et celles de sortie ; dans notre application, la non-linéarité existe bien entre les paramètres physico-chimiques de l'eau. Il s'agit d'utiliser une base de données constituée de vecteurs descripteurs de la qualité de l'eau brute. Pour chaque vecteur on possède les résultats de mesure de la quantité des substances chimiques choisies en sortie, et l'analyse physico-chimique correspondante (capteurs logiciels - régression), ainsi que la discision désirée en cas de classification (différents états de l'eau bien définis), le tout constitue un ensemble de paramètres descripteurs de la qualité de l'eau brute.

Dans un but de simulation, nous utilisons une méthode statistique dite *Analyse en Composantes principales* ACP dans une étape préliminaire pour réduire la dimension dans le sens d'éliminer la redondance compris entre les données de la base complète, et de conserver seulement les données non corrélés. Cette base est constituée de 10 paramètres physico-chimiques de l'eau qui sont : pH, Conductivité (C), Oxygène dissous (OD), Demande biologique (DBO5), Calcium (Ca), Hydrogénocarbonates (B), Chlorure (Cl), Magnésium ($Mg_)$, Matière en suspension (MES). Cette étape préliminaire permet de conservé que l'information essentielle et qui doit être plus facile à analyser que l'ensemble des données d'origine. Dans le but de construire des capteurs logiciels dans le sens de régression de données à partir de réseaux de neurones artificiels; ceux-ci nous fournissent la quantité de quelques paramètres non mesurables en continu ou à un prix important. L'objectif recherché,

consiste à valider les techniques employées (ACP et RNAs), dans le système proposé permettant à la fois l'extraction des caractéristiques, le contrôle et l'apprentissage[1].

4.2. Présentation du système de contrôle et de surveillance

L'architecture du système de contrôle et de surveillance proposée basée sur une approche multisensorielle et présentée dans la Fig. 4.1 , elle propose une solution au problème de contrôle vu comme un problème de reconnaissance de formes, où les formes représentent l'ensemble des observations ou mesures multisensorielles des paramètres liés à ses caractéristiques.

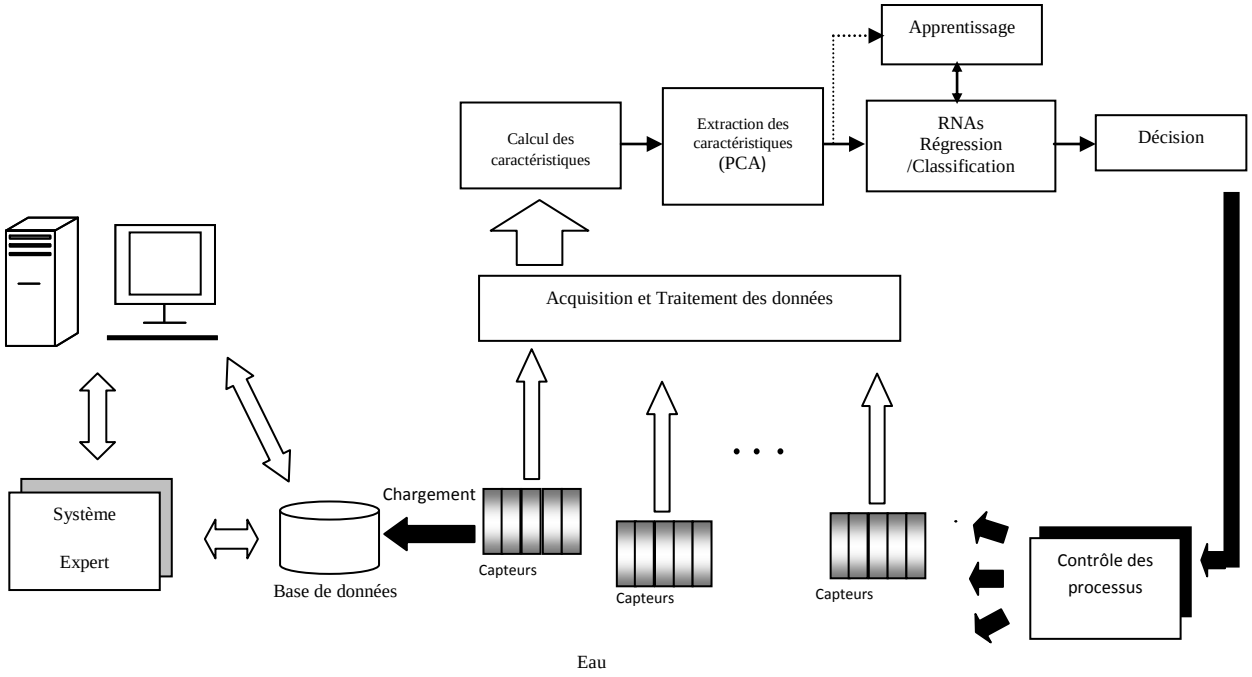


Fig. 4.1. Système de contrôle et de surveillance des eaux brutes.

Au niveau du système proposé, les différents paramètres physico-chimiques utilisés dans l'analyse de l'eau, tels que le pH, l'oxygène dissous (OD), la conductivité (*C*), les matières en suspension (MES),……etc, sont transformés en signaux électriques à partir des capteurs physiques, et transmis vers une station de contrôle qui assure l'acquisition et le traitement des données. A partir de la réponse d'un capteur on peut appliquer un ensemble d'opérations telles que les filtres destinés à annuler les effets de taille ou de positionnement, amplification, transmission dans le but de construire une base de données complète entre autre d'obtenir une nouvelle représentation meilleure et moderne de l'application. Il existe d'autres

modules de traitement complémentaire pouvant élaborer des représentations successives de l'objet; ces différentes représentations ont généralement pour objectif d'extraction des caractéristiques dans le sens de réduction de la dimension, c'est-à-dire de diminuer le nombre de descripteurs corrélés de l'ensemble et d'élaborer des descripteurs de plus en plus pertinents et non redondantes pour la tâche de discrimination à accomplir par notre méthode choisie (ACP). La technique de reconnaissance utilisée au niveau du système de décision opère après chaque acquisition. Dans notre cas, le contrôle et la surveillance de l'eau peut être vu comme un problème de régression ou de classification, où on estime la quantité d'un paramètre physico-chimique ou décider sur l'état de l'eau à travers d'autres paramètres liés à ses caractéristiques. Le module d'apprentissage dans ce cas, est supervisé par un expert, il permet de collecter de manière continue les paramètres relatifs aux différents états de l'eau, pour la mise en œuvre d'une base de connaissance complète. Le but recherché, consiste à valider la méthode utilisée pour l'extraction pour le choix des paramètres d'entrée, et la technique de reconnaissance employée (RNAs) en matière de temps d'apprentissage, l'erreur d'entrainement et de taux de reconnaissance, dans le système de surveillance proposé permettant à la fois le contrôle et l'apprentissage [1].

4.3. Méthodologie de Modélisation et développement

Expérimentalement, on a pu constater que la relation entre les sorties prédictives et les caractéristiques de l'eau brute doit être fortement non-linéaire. Parmi les différents types de modèle de comportement possibles, le modèle à base de réseaux de neurones possède l'avantage de pouvoir intrinsèquement décrire des relations non-linéaires entre les variables d'entrées et celles de sorties d'un système. Dans un premier temps, une analyse des données expérimentales obtenues a permis de mettre en évidence le type d'informations clés (paramètres caractéristiques de l'eau brute) nécessaires à prendre en compte pour la détermination la quantité de l'eau. Cette analyse a donc permis de définir les variables d'entrées du réseau. Une deuxième partie, doit être consacré à l'identification des poids ou paramètres affectés aux connections du réseau, cette étape connue sous le terme d'apprentissage n'est autre qu'une étape d'identification des paramètres d'un modèle non-linéaire. Pour cela, les données des historiques doit être séparées en deux groupes: un groupe de données constituant la base d'apprentissage sur lesquelles portera la détermination des poids et un groupe de données de test non utilisées lors de la phase d'apprentissage mais

servant à « tester » le réseau lors d'une phase de reconnaissance une fois que les poids ont été déterminés.

4.3.1. L'analyse en composantes principales (ACP)

L'ACP est une méthode d'analyse multivariée qui a été souvent utilisée pour le traitement statistique de base des données multidimensionnelles. L'analyse factorielle en composantes principales est un traitement statistique de données dont le but est de représenter et d'expliquer les liaisons statistiques entre les phénomènes. Elle permet d'identifier des variables sous-jacentes, ou facteurs qui expliquent les corrélations à l'intérieur d'un ensemble de variables observées. Elle est souvent utilisée pour réduire un ensemble de données, et dans l'agrégation de l'information, en identifiant un petit nombre de facteurs qui expliquent la plupart des variances observées dans le plus grand nombre de variables manifestes. On peut également utiliser l'analyse factorielle pour résumer, synthétiser, et hiérarchiser l'information contenue dans un tableau de ***n*** lignes (les individus) et *p* colonnes (les variables). Les *n* individus sont décris par un nuage de *p* variables. L'information représentée par ce nuage revient à la dispersion des ***n*** points. Produire un résumé de cette information c'est projeter ces points dans un espace de dimension inférieure à *p* le nombre de variables initiales. Les axes de ce sous-espace sont dits « axes factoriels» ou «facteurs ». Chaque variable *p* porte en elle:

- Une part d'information originale ou part d'inertie.
- Une part d'information originale redondante avec les autres, venant des corrélations entre variables. C'est cette part d'information redondante qui va être regroupée dans le résumé factoriel.
- Les facteurs sont hiérarchisés de la manière suivante:
- L'axe concentre le maximum de l'information: c'est l'axe de la plus grande dimension du nuage de points et il fournit le meilleur résumé dans un espace à une dimension, mais il laisse des résidus de l'information.
- Le $2^{ème}$ axe concentre le maximum de l'information restante, il est orthogonal au premier et c'est le meilleur résumé dans un espace à deux dimensions, Mais, de même il laisse aussi des résidus.

- Le 3ème axe prend encore une part d'information moindre, il est orthogonal au deux premiers. Et ainsi de suite, pour les axes suivants tant que l'on pense qu'ils apportent encore de l'information.

Le nombre de composantes en théorie est égal aux nombres de variables originelles. Mais, en pratique, les premières directions permettent de couvrir un pourcentage élevé (80%, 90%)[3] de toutes les données originelles et sont donc utilisées pour restreindre l'espace d'observation.

4.4. Le réseau de neurones de type RBF

Depuis quelque temps un certain nombre de modèles basés sur les réseaux de neurones ont été développés et appliqués dans le domaine de l'eau. Quelques études récentes ont montré l'efficacité potentielle de cette approche. Dans notre application, un apprentissage (régression et classification) non linéaire de données est opéré en utilisant les réseaux de neurones de fonctions à base radiale (Radial Basis Function Neural Network, RBF). Cette technique à prouvé son succès dans plusieurs domaines d'applications. RBF dispose en fait d'une architecture identique à celle des réseaux de neurones classique PMC, mais compte une seule couche cachée à plusieurs neurones, dont les fonctions d'activation sont des fonctions à base radiale, les plus souvent des gaussiennes, associe à une couche de sortie linéaire. Ce type de fonction est défini dans l'expression suivante :

$$y_i(x) = \exp\left(-\frac{\|x - c_i\|^2}{2\sigma_i^2}\right) \tag{4.1}$$

Un neurone i appartenant à la couche cachée est défini par le centre c_i et le rayon σ_i de la gaussienne associé. La norme $\|.\|$ dénote la distance euclidienne entre le vecteur d'entrée x et le vecteur centre c_i. Ce sont les paramètres de configuration adaptés durant la phase d'apprentissage.

Fig. 4. 2 illustre l'exemple du réseau RBF utilisé dans cette application.

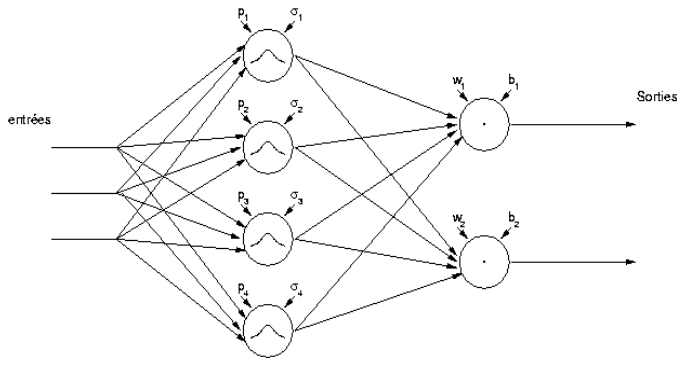


Fig.4. 2. Architecture d'un réseau RBF-NN.

L'apprentissage supervisé par le réseau consiste à déterminer les poids qui minimisant les écarts entre les valeurs de la sortie désirée y_{r_i} et les valeurs de la sortie décidée y_{d_i}. Le minimum du critère quadratique suivant est trouvé :

$$C_w = \frac{1}{N}\sum_{i=1}^{N}(y_{r_i} - y_{d_i})^2 \tag{4.2}$$

avec N représente le nombre d'exemples de la base d'apprentissage.

L'algorithme d'apprentissage consiste à ajouté un a un sur la couche cachée des neurones dont le centre est l'un des exemples de la base d'apprentissage. Dans notre cas, le rayon des gaussiens est constant pour tous les neurones de la couche cachée durant le processus d'apprentissage. La sortie *Z* du réseau RBF est déterminée par :

$$Z = \sum_{i=1}^{N} w_i y_i \tag{4.3}$$

y_i est l'activation de 2 ^{ième} neurone cachée.

4.5. Simulation

4.5.1. Sélection des descripteurs

La base de données utilisée est constituée de 774 échantillons. Pour chaque échantillon on possède les résultats de mesure réelle mais aussi d'analyses chimiques et physiques effectuées hors-ligne qui constituent un ensemble de 10 descripteurs de la qualité de l'eau brute: pH, T°, C, OD, DBO5, Ca, B, Cl, Mg, MES. La Figure 4. 3 présentes l'évolution des différents descripteurs de la qualité de l'eau brute utilisée dans ce travail. L'Analyse en Composantes Principales appliquée sur l'ensemble de ces données a fournit le tableau et l'histogramme donnés sur la Figue 4.4. On peut remarquer la décroissance rapide des valeurs propres, seules les quatre premières composantes représentent une prise en charge de plus de 87.10 % de l'inertie. L'ensemble des 10 variables est susceptible d'être simplifié et remplacé par 4 nouvelles variables représentées par les 4 premiers axes principaux[1].

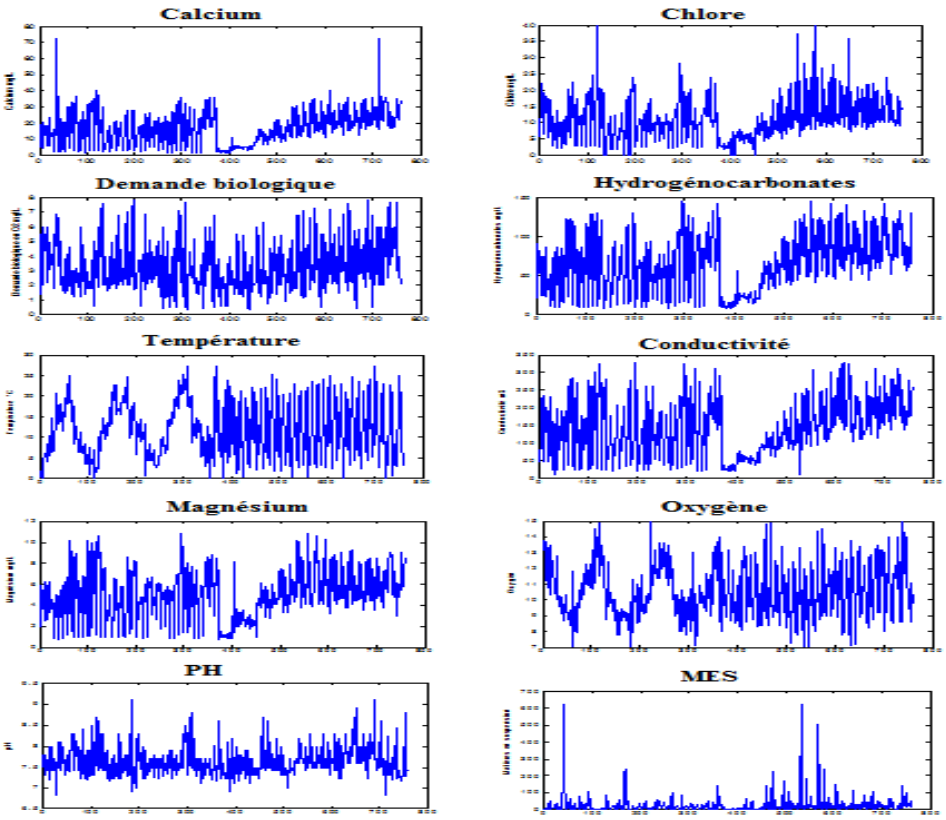


Fig. 4.3. Evolution des paramètres descripteurs de l'eau brute.

Variables	Valeurs propres	En %	% cumulé
1	4.945933	49.46%	49.4
2	1.863958	18.64 %	68.1
3	1.149983	11.50 %	79 .6
4	0.749768	7.50 %	87.1
5	0.663967	6.64%	93.7
6	0.193255	1.93%	95.6
7	0.135361	1.35 %	97.0
8	0.132678	1.33%	98.3
9	0 .088592	0.89 %	99.2
10	0.076506	0.77%	100.0

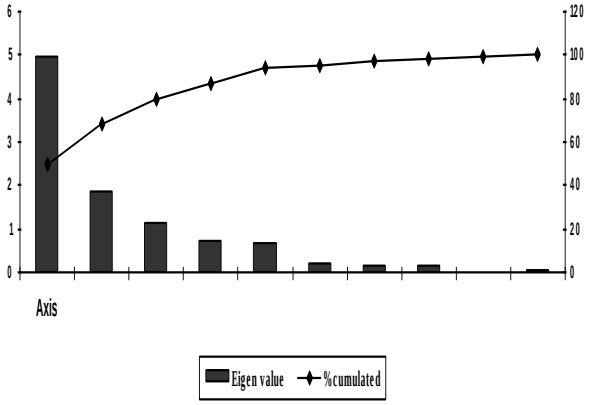


Fig. 4.4. Valeurs et histogrammes des valeurs propres des composantes.

Le tableau 4.1 présente les corrélations de chacune des variables avec les composantes principales. L'interprétation des résultats est la suivante:

Table .4.1 : Contribution de l'inertie totale et corrélation.

Variables	Fact.1		Fact.2		Fact.3		Fact.4	
	Corr.	Tot %	Corr.	Tot %	Corr.	Tot %	Corr.	Tot %
Temp (°C)	0.2031	4 %	**-0.9215**	85%	-0.0529	0%	0.1258	2%
Conductivité (us/cm)	**0.9462**	90 %	0.0370	0 %	-0.0532	0 %	-0.1104	1 %
pH	0.5621	32 %	-0.2547	6 %	-0.0458	0 %	**0.6105**	37 %
[OD] (mg/L)	-0.0114	0 %	**0.9238**	85 %	0.1333	2 %	0.2361	6%
DBO5 (20°C, mg/L)	0.4951	25 %	0.0402	0 %	0.5957	35 %	0.3726	14 %
MES	0.0848	1 %	-0.2038	4 %	**0. 8598**	74 %	- 0. 2974	9 %
Ca (mg/L)	**0.9400**	88 %	0.0708	1 %	-0.0708	0 %	-0.1187	1 %
Hydrogénocarbonates	**0.9295**	86 %	-0.0148	0 %	-0.0619	0 %	-0.1779	3 %
Cl (mg/L)	**0.9035**	82 %	0.2046	4 %	-0.1072	1 %	-0.0121	0%
Mg (mg/L)	**0.9365**	88 %	0.0700	0%	-0.1221	1%	-0.1419	2%
Variance	4.9459	49 %	1.8640	19 %	1.1500	11 %	0.7498	7 %

- L'axe 1 qui représente 49,46% de l'inertie totale est défini positivement et d'une façon nette par 5 variables très groupées Conductivité, Ca, Hydrogénocarbonates, Cl et Mg.

- L'axe 2 (***18.64%***) est défini par deux variables Température et L'oxygène dissous.

- L'axe 3 (***11.50 %***) est défini d'une façon nette par les matières en suspension

- L'axe 4 (***7.50 %)*** représente le pH.

Dans cette étude on peut conclure que pour le contrôle et la surveillance de l'eau en fonction des caractéristiques physico-chimiques, qui soient en plus facilement mesurables en continu, cela nos amène à ne garder que les variables : Conductivité (C), L'oxygène dissous (OD), pH, Matières en suspension (MES) qui sont positivement décorrélés[1].

4.5.2. Régression (capteurs logiciels)

4.5.2.1. Apprentissage e test

Le nombre de neurones d'entrées de réseau est donc 9 représentant les paramètres physico-chimiques de l'eau dans la base complète (10 paramètres), et un variable de sortie à chaque fois dans la régression de données en vue de développement des capteurs logiciels des paramètres non mesurables en continus. Après la réduction de dimension par l'application de

méthode ACP, les entrées de réseau sera 4 variables (base réduite), qui sont : OD, C, pH et MES comme il est montré auparavant.

Afin de procéder aux tests, nous avons tout d'abord, un ensemble de données réelles constituées de 774 échantillons, est présenté. Des bases de données de 443 vecteurs (pour la phase d'apprentissage) et 331 vecteurs (pour la phase de test), constitue de quartes paramètres physico-chimiques (T°, C, pH, MES,…etc) en plus de différents paramètres choisis comme variables de sorties à chaque fois. Notons bien que ces paramètres choisis sont corrélés avec la sortie désirée dans le cas de régression. L'algorithme de corrélation joue donc un rôle prépondérant dans la mesure où on veut développer des capteurs logiciels.

Le réseau de neurones multicouches PMC à apprentissage supervisé est le plus couramment utilisé. L'architecture du réseau de neurone de type RBF est donc choisie et illustrée ci-dessous dans la figure 4.5 représentant les deux cas étudiés (avec réduction -base réduite-), et sans réduction -base complète-). Le choix d'E/S des capteurs logiciels pour la base complète comprend donc 9 paramètres d'entrés et à chaque fois un paramètre restant non mesurable en continu est considéré comme sortie du réseau. Par contre 4 paramètres d'entrés fixés qui sont réduit à partir de l'ACP pour le deuxième cas et les paramètres restant sont considérés comme sorties du réseau constituent un ensemble des capteurs logiciels. Dans les deux cas les paramètres tels que : pH, C, DO et MES n'est jamais considérés comme des capteurs logiciels.

Il s'agit de valider les deux techniques appliquées au contrôle et de surveillance de l'eau potable. Faut-il souligner dans ce cas que la base d'apprentissage représente l'information la plus importante et la plus délicate à constituer. Il s'agit bien d'utiliser une base d'entraînement constituée de données descripteurs de l'eau.

- Couche d'entrée : 4 ou 9 neurones ;
- Couche de sortie : un neurone ;
- La fonction d'activation : Linéaire.

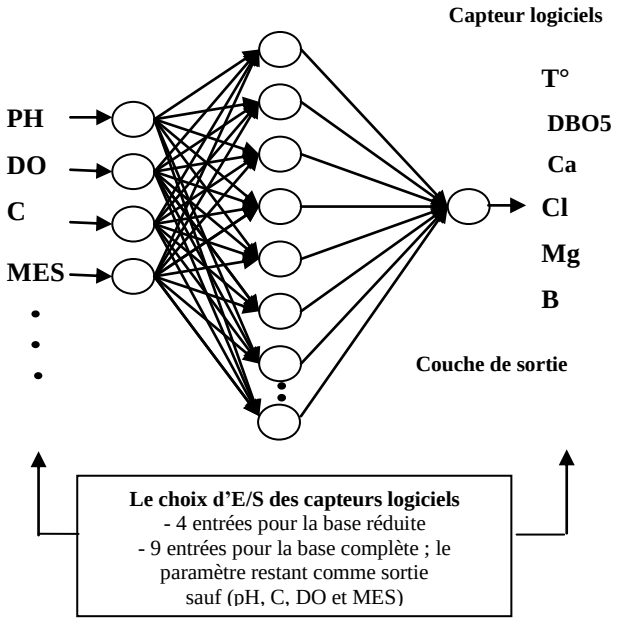


Fig. 4.5. Architecture du réseau (RBF) étudié.

Différents architectures sont testées, l'architecture du réseau de type RBF a prouvée d'être les plus préférée. Celui-ci est retenu pour avoir réalisé de bien meilleurs résultats comparativement aux autres réseaux basés sur des algorithmes d'apprentissage bien connus. L'apprentissage a été réalisé sur les bases de données construites. On présente dans le tableau 4.2 pour les deux bases (complète et réduite) testées. Les résultats correspondant aux différents paramètres d'apprentissage, tels que : le temps d'apprentissage (*T_appr*), et l'erreur d'entraînement (*EQMA*), qui sont tous obtenus à partir d'une base de données de 443 vecteurs.

Table 4.2 : Résultats d'apprentissage – régression -

		Base complète (9 vecteurs)		**Base réduite (4 vecteurs)**	
Base de données	**Capteurs logiciels**	**T_appr (sec)**	**EQMA**	**T_appr (sec)**	**EQMA**
443 vecteurs	DBO5	41.5469	1.1153e-028	43.0781	5.5305e-004
	Mg	49.0781	3.5096e-028	41.8438	4.5147e-005
	T^0	43.9531	1.3157e-027	43.3750	0.0014
	Ca	48.2813	6.8191e-027	39.6875	0.0029
	B	46.7656	8.6515e-026	46.7656	0.0011
	Cl	50.7813	7.4435e-028	41.0938	4.5147e-005

On remarque que l'erreur d'entrainement de la base de données complète meilleur que la base de données réduite, mais en général l'erreur d'entraînement est faible dans les deux cas. Les figures 4.6 et 4.7 montrent leurs résultats de prédiction ponctuelle obtenue sur l'ensemble d'apprentissage de 443 vecteurs indépendant. Suivant les résultats montrés dans le tableau 4.2 les capteurs de Mg, Cl et DBO5 sont les mieux placés dans les deux cas étudiés.

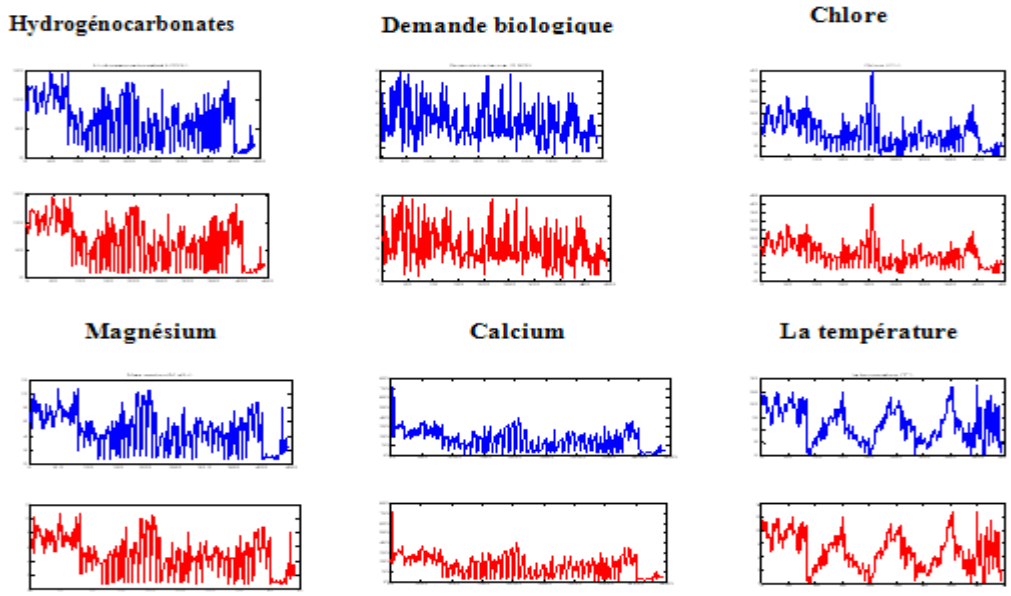


Fig. 4.6. Résultats d'apprentissage (base complète).

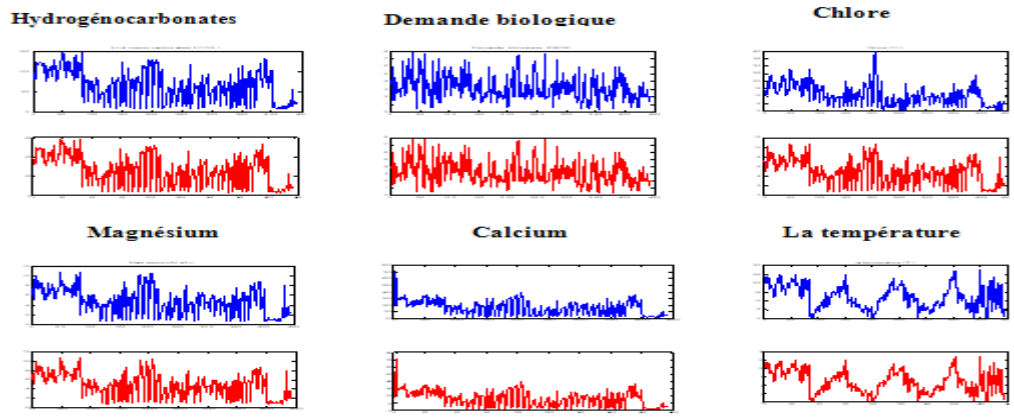


Fig. 4.7. Résultats d'apprentissage (base réduite).

4.5.2.2. Résultats de généralisation (test)

Une fois que le modèle neuronal est déterminé, en procédant à la phase de généralisation, une base de test de 331 vecteurs, est appliquée pour chaque capteur logiciel.

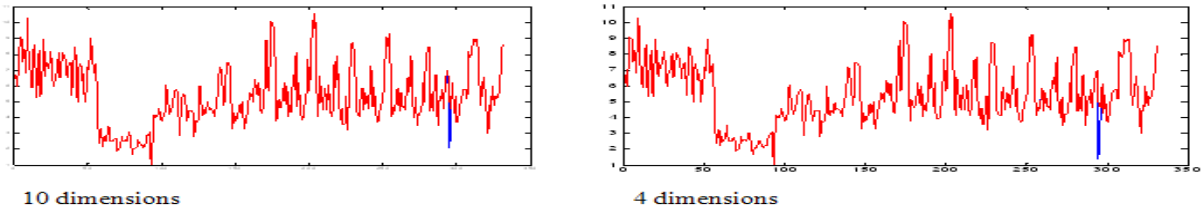
La comparaison entre la sortie calculée par le réseau (valeurs estimées) et les quantités obtenue en sortie (valeurs mesurées) montre que les valeurs estimées sont très porche de la sortie mesuré. Les résultats obtenus dans la phase de test après la validation des deux modèles construit avec les deux bases complète et réduite correspondant aux différents paramètres de test, tels que : le temps de test (*T_test*), et l'erreur quadratique moyen de test (*EQMT*), qui sont tous obtenus à partir d'une base de données de test de 331 vecteurs sont représentés dans le tableau 4.3.

Table .4.3 : Résultats de test.

		Base complète		**Base réduite**	
Base de données	**Capteurs logiciels**	**T_test (sec)**	**EQMT**	**T_test (sec)**	**EQMT**
331 vecteurs	DBO5	0,171	0.0558	0.1719	0,0141
	Mg	0.1719	0.0254	0.1719	0.0396
	T^0	0.2031	0.0812	0.1719	0.0193
	Ca	0.1875	1.5982	0.1719	2.2893
	B	0.1875	9.1354	0.1875	16.442
	Cl	0.2031	0.1162	0,171	0,5865

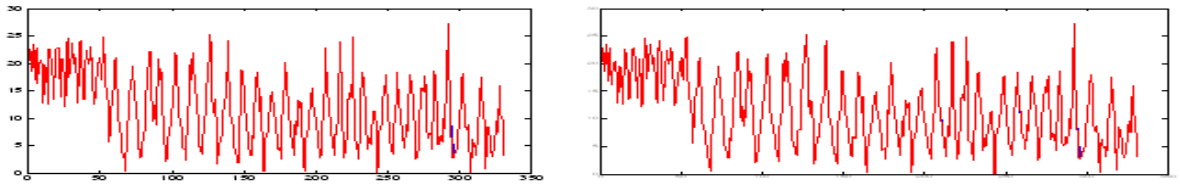
A partir des résultats obtenus dans le tableau précédant, les capteurs logiciels de Ca, B et Cl présente des mauvais résultats en termes de l'erreur quadratique moyenne de test. Par contre les autres capteurs logiciels sont mieux placés en termes d'erreur (0.0141-0.0812). Une erreur entre 1 et 8 %. Le capteur logiciel de Mg présente le meilleur résultat, une erreur entre 2 à 3 % dans les deux cas. Les résultats de test sont présentés dans la figure 4.8.

B



10 dimensions 4 dimensions

Mg



10 dimensions 4 dimensions

T

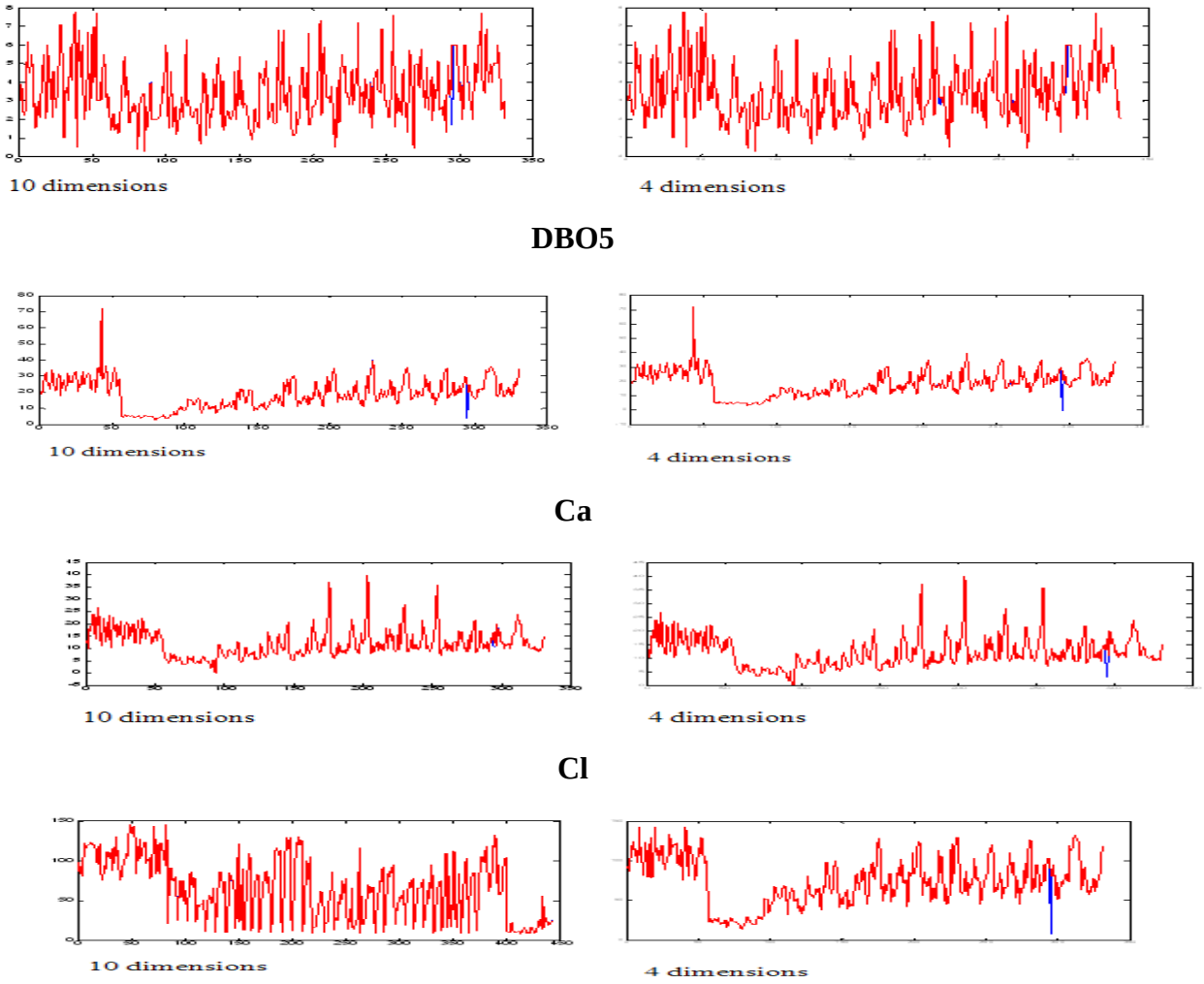


Fig. 4.8. Résultats de test.

4.5.2.3. Evaluation des performances

Dans cette évaluation préliminaire de comparaison des performances des capteurs logiciels établis, nous définissons quelques mesures des performances statistiques comme suit :

- Racine de l'erreur quadratique moyenne (RMSE).
- Erreur quadratique moyenne (MSE).
- Erreur relative moyenne (ERM).
- Erreur absolue moyenne (EAM).

Des résultats statistiques de deux modèles dans les deux phases (apprentissage et test) sont récapitulés dans les tableaux suivants. L'évolution de l'erreur relative dans les deux phases pour les deux modèles, est effectuée. Les tableaux 4.4 j'jusqu'a 4.9 montrent les résultats de généralisation obtenus.

Tableau. 4.4. a, b, c capteurs logiciels de Cl.

- a -

	Exemples d'apprentissage					Exemples de test				
	valeur moyenne (mg/l)	**Standard deviation (mg/l)**	**Minimum (mg/l)**	**Maximum (mg/l)**	**R^2**	**valeur moyenne (mg/l)**	**Standard deviation (mg/l)**	**Minimum (mg/l)**	**Maximum (mg/l)**	**R^2**
Mesuré	10,63	6,490	0	40		12,026	5,648	0	37,2	
Base complète	10,63	6,490	0	40	0.99	12,007	5,641	0	37,2	0.99
Base réduite	10,63	6,490	0	40	0.99	11,983	5,662	0	37,2	0.98

- b-

	Exemples d'apprentissage				Exemples de test			
	RMSE	**MSE**	**ERM**	**EAM**	**RMSE**	**MSE**	**ERM**	**EAM**
Base complète	2.72e-14	7,44e-28	0	0	0,3408	0,116	0,001	0,018
Base réduite	0,0067	1,11e-5	0	0	0,118	0.0141	0,0024	0,042

-c-

	Exemples d'apprentissage			Exemples de test		
	0.05	**0.1**	**0.25**	**0.05**	**0.1**	**0.25**
Base complète	100	100	100	99.69	99.69	99.69
Base réduite	100	100	100	99.39	99.39	99.69

Tableau. 4.5. a, b, c capteurs logiciels de DBO5.

- a -

	Exemples d'apprentissage					Exemples de test				
	valeur moyenne (mg/l)	**Standard deviation (mg/l)**	**Minimum (mg/l)**	**Maximum (mg/l)**	**R^2**	**valeur moyenne (mg/l)**	**Standard deviation (mg/l)**	**Minimum (mg/l)**	**Maximum (mg/l)**	**R^2**
Mesuré	3,16	1, 53	0,3	7,8		3, 413	1,55	0,3	7,8	
Base complète	3,16	1,53	0,3	7.8	1	3,400	1,542	0.32	7.8	0.99
Base réduite	3,16	1,53	0,3	7,8	0.99	3,407	1,539	0.3	7,8	0.98

- b -

	Exemples d'apprentissage				Exemples de test			
	RMSE	**MSE**	**ERM**	**EAM**	**RMSE**	**MSE**	**ERM**	**EAM**
Base complète	1.05e-14	1,11e-13	0	0	0.2362	0.0558	0,0021	0,012
Base réduite	0.0235	5,53e-4	0	0	0,118	0.0141	9,66e-4	0,008

- c -

	Exemples d'apprentissage			Exemples de test		
	0.05	**0.1**	**0.25**	**0.05**	**0.1**	**0.25**
Base complète	100	100	100	99.69	99.69	99.69
Base réduite	99.54	99.54	100	99.09	99.09	99.69

Tableau. 4.6. a, b, c capteurs logiciels de Ca.

- a -

	Exemples d'apprentissage					Exemples de test				
	valeur moyenne (mg/l)	**Standard deviation (mg/l)**	**Minimum (mg/l)**	**Maximum (mg/l)**	**R^2**	**valeur moyenne (mg/l)**	**Standard deviation (mg/l)**	**Minimum (mg/l)**	**Maximum (mg/l)**	**R^2**
Mesuré	16,21	10,086	2	36,4		18,815	8,918	3	36	
Base complète	16,21	10,086	2	36,4	1	18,745	8,944	3	36	0.98
Base réduite	16,21	10,086	2	36,4	0.99	18,735	8,963	0,504	36	0.97

- b -

	Exemples d'apprentissage				Exemples de test			
	RMSE	**MSE**	**ERM**	**EAM**	**RMSE**	**MSE**	**ERM**	**EAM**
Base complète	8,25e-14	6.81e-29	0	0	1,2641	1,5982	0,0025	0,069
Base réduite	0.053	0.0029	0	0	1,513	2,2893	0,0029	0,193

- c -

	Exemples d'apprentissage			Exemples de test		
	0.05	**0.1**	**0.25**	**0.05**	**0.1**	**0.25**
Base complète	100	100	100	99.39	99.39	99.39
Base réduite	99,69	100	100	99.39	99.39	99.77

Tableau. 4.7. a, b, c capteurs logiciels de T°.

- a -

	Exemples d'apprentissage					Exemples de test				
	valeur moyenne (mg/l)	**Standard deviation (mg/l)**	**Minimum (mg/l)**	**Maximum (mg/l)**	**R^2**	**valeur moyenne (mg/l)**	**Standard deviation (mg/l)**	**Minimum (mg/l)**	**Maximum (mg/l)**	**R^2**
Mesuré	12,073	6.306	0,2	25,3		11,662	6,182	0,3	25,3	
Base complète	12,073	6.306	0,2	25,3	1	11,678	6,166	0,3	25,3	0,99
Base réduite	12,073	6.306	0,2	25,3	0.99	11,669	6,173	0,3	25,3	0,99

- b -

	Exemples d'apprentissage				Exemples de test			
	RMSE	**MSE**	**ERM**	**EAM**	**RMSE**	**MSE**	**ERM**	**EAM**
Base complète	3,62e-14	1.31e-27	0	0	0.2849	0.0812	0 ,0055	0,0156
Base réduite	0.037	0.0014	0	0	0,138	0,0193	-0,0026	0,0105

- c -

	Exemples d'apprentissage				Exemples de test			
	0.05	**0.1**	**0.2**	**0.25**	**0.05**	**0.1**	**0.2**	**0.25**
Base complète	100	100	100	100	99.69	99.69	99.69	99.69
Base réduite	99,69	100	100	100	99.69	99.69	99.69	99.69

Tableau. 4.8. a, b, c capteurs logiciels de Mg.

- a -

	Exemples d'apprentissage					Exemples de test				
	valeur moyenne (mg/l)	**Standard deviation (mg/l)**	**Minimum (mg/l)**	**Maximum (mg/l)**	**R^2**	**valeur moyenne (mg/l)**	**Standard deviation (mg/l)**	**Minimum (mg/l)**	**Maximum (mg/l)**	**R^2**
Mesuré	4,81975	2,411	0,7	10,8		5 ,4445	1,863	1	10,8	
Base complète	4,81974	2,411	0,7	10.8	1	5,4368	1,872	1	10,8	0,99
Base réduite	4,81975	2,411	0,7	10,8	0.99	5,4375	1,876	1	10,8	0,98

- b-

	Exemples d'apprentissage				Exemples de test			
	RMSE	**MSE**	**ERM**	**EAM**	**RMSE**	**MSE**	**ERM**	**EAM**
Base complète	1.87e-14	3,50e-28	0	0	0,1593	0,0254	0 ,0017	0,0087
Base réduite	0,0067	4,51e-5	0	0	0,1989	0,0396	0,0021	0,0115

-c -

	Exemples d'apprentissage				Exemples de test			
	0.05	**0.1**	**0.2**	**0.25**	**0.05**	**0.1**	**0.2**	**0.25**
Base complète	100	100	100	100	99.69	99.69	99.69	99.69
Base réduite	100	100	100	100	99.39	99.39	99.39	99.39

Tableau. 4.9. a, b, c capteurs logiciels de B.

-a-

	Exemples d'apprentissage					Exemples de test				
	valeur moyenne (mg/l)	**Standard deviation (mg/l)**	**Minimum (mg/l)**	**Maximum (mg/l)**	**R^2**	**valeur moyenne (mg/l)**	**Standard deviation (mg/l)**	**Minimum (mg/l)**	**Maximum (mg/l)**	**R^2**
Mesuré	65,10	37,114	8	146		74,346	30,340	22	143	
Base complète	65,10	37,114	8	146	1	74,180	30,448	22	143	0,99
Base réduite	65,10	37,114	8	146	0.99	74,123	30,553	22	143	0,98

-b-

	Exemples d'apprentissage				Exemples de test			
	RMSE	**MSE**	**ERM**	**EAM**	**RMSE**	**MSE**	**ERM**	**EAM**
Base complète	2,94e-13	8,65e-26	0	0	3,022	9,135	0,002	0,166
Base réduite	0,0331	0,001	0	0	4,054	16,44	0,0027	0,2258

-b-

	Exemples d'apprentissage				Exemples de test			
	0.05	**0.1**	**0.2**	**0.25**	**0.05**	**0.1**	**0.2**	**0.25**
Base complète	100	100	100	100	99.69	99.69	99.69	99.69
Base réduite	100	100	100	100	99.39	99.39	99.39	99.39

En effet, et d'après les résultats obtenus dans les tableaux précédents, on trouve que le coefficient de corrélation entre les valeurs estimées et réelles est tout à fait satisfaisant. Cela est exprimée par la forte valeur du coefficient de corrélation (R^2= 0.97-0.99), ainsi obtenu. La précision de l'estimation des valeurs est confirmée par l'absence de grandes erreurs exceptionnelles. On peut constater que la moyenne et l'écart type des valeurs estimées et réelles dans tous les cas sont proches les uns des autres, ce qui indique une finesse globale de modélisation. Les valeurs minimales et maximales sont bien assorties par les deux bases. L'évolution de l'erreur relative dans les deux phases (apprentissage et test) obtenue par les deux modèles, est montrée ci-dessous dans les tableaux de résultats. L'analyse des résultats obtenus, montre que les deux modèles présentés par les deux bases de données ont bonne capacité d'apprentissage vis à vis de l'erreur d'entraînement. Le temps d'apprentissage est relativement faible. Ils ont une bonne aptitude en généralisation que l'on justifié par les taux de reconnaissance obtenus (supérieur à 99 %). Pour des erreurs supérieures à 25%, il est recommandé d'utiliser plus d'entrées pour mieux caractériser l'objet à mesurer. Dans la marge d'erreur de 25 %, les tests ont conduit à un taux de reconnaissance supérieur à

99.69%, pour les deux bases. Il s'avère satisfaisant relativement à la marge d'erreur indiquée. L'erreur relative moyenne de test calculé par les deux bases est nulle.

4.5.3. Classification binaire (Deux classes : potable et non potable)

Il s'agit de valider la technique RNAs (RBF) appliquée au contrôle de qualité de l'eau potable dans le sens de bien défini l'état de l'eau. Faut-il souligner dans ce cas que la base d'apprentissage représente l'information la plus importante et la plus délicate à constituer. Il s'agit bien d'utiliser une base d'entraînement constituée de données relatives aux différents états qualitatifs de l'eau suivant les normes recommandées. Dans un but de simulation, les bases de données réelles complète et réduite constituées des paramètres physico-chimiques (pH, MES, C, DO,…..), ont été utilisées. Dans ce cas là, le problème de contrôle et de surveillance des eaux est vu comme un problème de classification des données. Les classes sont comme suit : exemples positifs pour le cas potable, et négatifs pour le cas non potable [3].

4.5.3.1. Résultats d'apprentissage et de test

Ils sont testés et validés en utilisant l'algorithme d'apprentissage et de test par l'architecture du réseau RBF. Les résultats de l'apprentissage et de test avec les deux bases de données réelles précédemment utilisées (complète et réduite), sont présentés dans les tableaux 4.10 et 4.11 respectivement. Les paramètres tels que : le temps d'apprentissage (*T_appr*), l'erreur d'entraînement (*Er*) pour l'apprentissage, et le taux de reconnaissance (*Taux_rec*), et le temps de test (*T_test*), pour le test, sont évalués.

Table. 4.10 : Résultats d'apprentissage.

Les bases	T_appr (sec)	Er	Taux de reconnaissance (Taux_rec)
Base complète	41.9688	5.4282e-029	100%
Base réduite	40.5469	2.9106e-029	100%

Table. 4.11 : Résultats de test.

Les bases	T_test (sec)	Taux de reconnaissance (Taux_rec)
Base complète	0.1719	99,09 %
Base réduite	0.1563	99,09 %

Les résultats d'apprentissage portés dans les tableaux précédents montrent que l'utilisation de deux bases de données donne les mêmes résultats entre autre, le taux de reconnaissance. Soit un temps d'apprentissage relativement court et l'erreur d'entraînement reste relativement faible. Les résultats de test montrés une bonne adéquation de l'algorithme testé pour ce type d'application. Un taux de reconnaissance plus de 99,09 % est donc obtenu, avec même une erreur de 0,0363. Ces résultats affichent un bon taux de reconnaissance lorsqu'en fait réduire les dimensions d'entrées. La réduction de dimension porte les mêmes résultats que les bases de données complètes.

4.5.3.2. Discussion des résultats

a). Analyse

On sait bien que l'algorithme d'apprentissage des RNAs générale influe sur la généralisation qui représente la tâche accomplie par le réseau une fois que son apprentissage achevé. Celle-ci est aussi influencé essentiellement par quatre facteurs : la complexité du problème, l'algorithme d'apprentissage, la complexité de l'échantillon (le nombre d'exemples) et enfin la complexité du réseau (nombre de poids). La complexité du problème est déterminée en partie par sa nature même : on peut parler de « complexité intrinsèque ». Par ailleurs, l'algorithme d'apprentissage influe sur la généralisation par son aptitude à trouver un minimum local assez profond, sinon le minimum global. Un facteur influent sur la généralisation est la complexité du réseau. On peut constater que le modèle ayant très peu de paramètres n'a pas assez de flexibilité pour réaliser un apprentissage correct des exemples d'apprentissage. Les erreurs d'apprentissage et de test sont toutes deux importantes : c'est la situation de *sous-apprentissage*. En revanche, le modèle constitué de nombreux paramètres, lisse parfaitement les exemples d'apprentissage. Il commet donc une erreur faible sur ces données, mais probablement une erreur plus importante sur les données de test. C'est la situation de *sur-apprentissage.* Finalement, le modèle possédant un nombre de paramètres modérés réalise un bon compromis entre précision d'apprentissage et bonne généralisation.

Le problème de la généralisation est souvent vu sous trois perspectives différentes. Dans la première, la taille du réseau est fixée (en accord avec la complexité du problème). Dans notre cas, les neurones d'entrée, où chaque neurone représente un paramètre physico-chimique et un neurone de sortie décrivent dans un premier temps notre réseau (régression ou classification). La question qui se pose est de : combien d'exemples d'apprentissage sont

nécessaires pour atteindre une bonne généralisation ? Cette perspective est intéressante dans les applications où l'on a la possibilité d'acquérir autant d'exemples que l'on veut. Dans le cas d'un système multicapteur (notre cas précis), on peut acquérir autant d'exemples qu'on veut, cependant une base de 744 vecteurs par exemple est-elle suffisante ? La deuxième perspective c'est quand nous supposons que le nombre d'exemples d'apprentissage est fixé ; la question qui se pose dans ce cas est : quelle est la taille du réseau qui donne la meilleure généralisation de ces données ? Dans notre application, quel est le réseau pris parmi les différents réseaux testés, qui donne la meilleure généralisation ? Est ce un réseau à une seule couche cachée est suffisant ? On est conduit à adopter finalement ce point de vue puisqu'on est devant l'impossibilité d'avoir une base de connaissance aussi complète qu'elle soit. Un enrichissement continuel de cette base avec le temps est pratiquement indispensable. Cela dépend de beaucoup de paramètres aussi bien climatiques que géographiques. Il importe alors dans cette situation de déterminer la taille du réseau qu'il faut pour décrire au mieux les données en notre possession. Cependant, tous les différents réseaux validés peuvent acquérir des données d'apprentissage, et l'erreur d'entraînement est plus faible presque dans tous ces réseaux. La variante de l'estimation due à la taille finie de l'échantillon induit un écart entre la capacité réelle de généralisation et la capacité estimée (risque empirique). Dans la troisième perspective, on se donne des complexités d'échantillon et de modèle et on cherche pour une probabilité fixée, l'écart maximum entre la vraie capacité de généralisation et la capacité de généralisation estimée à partir de l'échantillon. La théorie de Vapnik, principalement dans le développement de la théorie de l'apprentissage statistique, permet de répondre à la première et à la troisième question. Les notions de *dimension de Vapnik-Chervonenkis* et de la théorie des courbes d'apprentissage permettent d'établir un lien entre la complexité de l'échantillon et la complexité du réseau. Si on revient aux principes théoriques de base, les réseaux de neurones sont basés sur le principe de la minimisation du risque empirique (MRE). Ce principe se traduit par les méthodes connues, pour la régression par exemple, on minimise le nombre d'erreurs en apprentissage. On peut minimiser le risque empirique (par une règle d'apprentissage) après le choix d'une architecture d'un réseau, soit fixer la valeur du risque empirique (idéalement, à la valeur 0). Dans notre cas, Malgré la diversité des architectures de réseaux de neurones, surtout qu'il n'existe pas une règle bien précise pour fixer le nombre de neurones et de couches cachées dans un réseau, sauf les réseaux de type RBF qui on peut dit qui ont eu une architecture pratiquement standard. Ce problème de choix est posé et reste le principal inconvénient dans l'utilisation des RNAs. Ce compromis de choix entre le nombre d'itérations (temps d'apprentissage) et l'erreur d'entraînement (pour la phase de

généralisation), avec la considération du taux de reconnaissance, rentre dans le choix de l'algorithme d'apprentissage et l'architecture du réseau le plus préférable pour notre application. D'après les résultats obtenus (taux de reconnaissance plus de 99% à 25 %), l'erreur en apprentissage qui représente le risque empirique est faible, avec l'utilisation de réseau de type RBF qui le plus performant par rapport au réseau de neurone artificiels.

b). Evaluation

Les résultats de simulation caractéristiques obtenus dans les tests de validation des deux modèles des bases différentes sont résumés ci-dessous dans le Table. 4.12**.**

Table .4.12 : Tableau comparatif des résultats obtenus pour les deux modèles.

	Régression (DBO5)		**Classification**	
Propriétés	**Base complète**	**Base réduite**	**Base complète**	**Base réduite**
Temps d'entraînement	41.5469	43.0781	41.9688	40.5469
Temps de test	0,171	0.171	0.1719	0.1563
Erreur d'entraînement (*EQMA*)	1.1153e-028	5.5305e-004	5.4282e-029	2.9106e-029
Erreur de test	0.0558	0,0141	-	-
Taux de reconnaissance	100	100	99,09 %	99,09 %
Coefficient de détermination R^2	0.99	0.98	-	-

D'après ce tableau comparatif, il apparaît que sur le plan décisionnel, les deux modèles (base complète et réduite) présentent de bons résultats, avec des taux de reconnaissance supérieur à 99.09 %. Le temps de calcul de la phase d'apprentissage est reste stable. Le timing correspondant à la phase de test correspond à la base réduite lui confère l'avantage d'une intégration dans un système de surveillance dynamique (moins de capteurs en entrée). Les caractéristiques affichées dans les résultats obtenus soulignent l'intérêt théorique et pratique de l'utilisation des réseaux de neurones RBF dans les deux cas régression et classification des données d'une part, et d'autre part la réduction de dimension entre autre l'analyse multivarie en général pour ce type d'application.

c). Comparaison

Une des principales caractéristiques des données environnementales réside dans le fait qu'on dispose d'un nombre de données de sorties très faible devant un grand nombre d'entrées. Un enjeu majeur du traitement statistique de ces données est l'analyse multivariée à but décisionnel. D'un point de vue statistique, ce grand nombre de variables devant ln petit nombre d'observations rend l'analyse multivariée difficile. Une façon de contourner ce fléau de la dimension consiste à réduire cette dimension. La régression et la classification supervisée est vue comme un problème de reconnaissance de formes avec peu d'observations

et beaucoup de variables d'entrées dans un système pour l'aide à la décision. L'analyse en composantes principale (ACP) comme une technique de réduction de dimension entre autre l'analyse multivarie. D'après les résultats obtenus, notre travail a permet d'illustrer la pertinence parmi de ces approches proposées lorsqu'elles sont appliquées à l'analyse de deux jeux de données réels différents (complète et réduite). La régression et la classification supervisée obtenue en combinant les techniques d'analyse multivariée (ACP) et les réseaux de neurones artificiels de type RBF comme technique de reconnaissance de formes sont appliquées au domaine de surveillance des eaux expriment les variables physico-chimiques indicatrices à permet de montrer l'effet de l'analyse multivariée sur la régression et la classification supervisée. Ces résultats affirment que l'analyse multivarie par l'ACP dans notre application donne un taux très acceptable et parfois équivalent aux celles de sans réduction de dimension, mais dans la réalisation pratique on choisi un système multicapteur opérant avec moins de voies (capteurs physiques) en entrée mais donne des bonnes résultats que celle de plusieurs voies, entre autre un coût plus faible.

Conclusion

Ce dernier chapitre est dédie à une étude en simulation concernant la technique statistique ACP permettant de sélectionner les entrées et éliminer les informations redondantes, ainsi que la mise en œuvre de la technique d'apprentissage statistique RBF appliquées dans le domaine de contrôle et de surveillance des eaux potables. Cette étude a permis la validation et l'évaluation des performances de deux méthodes présentées (ACP et RBF). Une comparaison était effectuée pour voir l'effet de l'analyse multivarie (réduction de dimension) dans la régression et la classification des données dans le but d'une comparaison des performances. Les paramètres liés au taux de reconnaissance, au temps d'apprentissage, à l'erreur d'entraînement, ont été les facteurs pertinents qui nous ont permis d'évaluer les méthodes étudiées dans les deux champs d'applications.

CONCLUSION GENERALE

Dans notre travail présenté dans ce mémoire a été consacré à la mise en œuvre de la méthode « *Analyse en composantes principales,* ACP» de l'analyse multivariée appliquées à la reconnaissance de formes dans le domaine de contrôle et la surveillance des eaux brutes. Cette étude découle des progrès technologiques importants qui ont été enregistrés ces dernières années, dans le but et l'intérêt d'une surveillance moderne et une meilleure efficacité de la qualité des eaux propres. A cet effet, notre modeste travail peut être considéré comme une contribution aux solutions proposées, pour résoudre des problèmes d'intérêt stratégique à préoccupation nationale, utilisant des outils modernes à base de techniques avancées.

Les divers dispositifs et outils de surveillance dans le domaine de l'eau existants actuellement de par le monde, sont réalisés dans le but d'assurer une surveillance permanente et efficace. C'est dans l'esprit et l'intérêt considérable que présente le contrôle de la qualité de cette ressource dans les usines de production et de distribution de l'eau, que nous avons tenté dans ce travail d'exposer notre application. Le système appelé multicapteur proposé, permet à la fois d'assurer le contrôle permanent et l'apprentissage. Ce travail concerne l'étude d'un tel système à partir de la donnée des caractéristiques physico-chimiques de l'eau brute telles que le pH, la température, etc.

L'aspect novateur de ce travail réside dans l'intégration de diverses techniques dans un système global comprenant le contrôle automatique de l'eau. Ce contrôle à été précédé d'une analyse statistique (Analyse en Composantes Principales), permettant de déterminer les corrélations existantes entre les variables caractéristiques de l'eau brute puis de ne conserver que les caractéristiques apportant réellement une information pertinente. Notre travail s'est basé sur l'exploitation des données réelles. Les modèles exposés lors de cette étude présentent de bonnes performances en matière de taux de reconnaissance. Un intérêt d'usage et

d'application de ces techniques dans ce domaine est donc bien justifié. La régression (*capteur logiciel*) et la classification (*surveillance*) ce qui présente un souci majeur pour l'obtention d'un modèle optimale. Le taux de reconnaissance de calcul imparti à la phase d'apprentissage est très court, ce qui lui confère l'avantage d'une intégration dans un système de surveillance dynamique. Il faut toute fois souligner le fait que le principal souci pour l'application de ce type de modèle, est l'obtention d'un réseau « *optimal* ». Ceci met évidemment en jeu le nombre et le type d'exemples à utiliser dans la base d'apprentissage pour un moindre coût. Un enrichissement continuel de la base de données avec le temps est pratiquement indispensable, car cela dépend de beaucoup de paramètres aussi bien climatiques que géographiques. La régression et la classification supervisée obtenue en combinant les techniques d'analyse multivariée (ACP) et les réseaux de neurones artificiels de type RBF comme technique de reconnaissance de formes sont appliquées au domaine de surveillance des eaux expriment les variables physico-chimiques indicatrices à permet de montrer l'effet de l'analyse multivariée sur la régression et la classification supervisée. Ces résultats affirment que l'analyse multivariée par l'ACP dans notre application donne un taux très acceptable et parfois équivalent aux celles de sans réduction de dimension, mais dans la réalisation pratique on choisi un système multicapteur opérant avec moins de voies (capteurs physiques) en entrée mais donne des bonnes résultats que celle de plusieurs voies, entre autre un coût plus faible. Plusieurs perspectives peuvent être envisagées, d'une part concernant la méthode de régression et de classification, qui peut être complémenté avec l'utilisation d'autres algorithmes en parallèle, d'autres part concernant l'entrée du système en ajoutant de nouveaux capteurs en entrée pour des paramètres non mesurable en continu; des capteurs logiciels entre autres.

Bibliographie

[1] **A. BAIRA** « étude corrélative des paramètres physico-chimiques pour le contrôle et la surveillance des eaux propres » Université de M'sila, JUIN 2011

[2] **M.LADJAL**« traitement et fusion multisensorielle appliques a la surveillance des eaux potables» *Thèse de Magister* Université de M'sila,

[3] **T. SERAICHE , S.ALLOUACHE.** « Application des réseaux de neurones artificiels surveillance dans le domaine de surveillance des eaux potables », Université de M'sila, 2006

[4] **H. Hernandez** «développement d'un capteur logiciel pour la prédiction d'un coagulant dans une station de traitement d'eau potable en vue de son diagnostic » Thèse de Doctorat de l'INSA de Toulouse 2006.

[5] « Nouveaux risques sanitaires et nouveaux enjeux pour le contrôle de la qualité des eaux potables **»** Article de synthèse, Les Technologies de laboratoire N°3 Mars –Avril 2007Université Paris Sud 11.

[6] «Contrôles de la qualité des eaux **»**, un magazine.

[7] JORADP n°18 du 23/03/2011 décret exécutif n°11-125 du 22/03/2011 relatif à la qualité de l'eau de consommation humaine en Algérie.

[8] « Conseils pour un approvisionnement en eau potable salubre dans les secteurs de compétence fédérale » Version 1 partie 2.

[9] **N.Valentin**, « Construction d'un capteur logiciel pour le contrôle automatique du procédé de coagulation en traitement d'eau potable », Thèse doctorat, Laboratoire des eaux UTC, 2000.

[10]. «L'eau potable » Pdf/Cours univ-reunion.

[11] **M. Zemouri**, « Contribution à la surveillance des systèmes de production à l'aide des réseaux de neurones dynamique, Application à la e-maintenance », Thèse de doctorat, Université de Franche-Comté, 2003.

[12] **A.Balamane.** « Sélection D'attributs par dimension fractale **»**, Université de Québec, décembre 2007.

[13] **S. AMBAPOUR** « Introduction à l'analyse des données » document de travail L'analyse de données Pdf/ Polycopié de cours ENSIETA - Réf. : 1463

[14] **G. Lelandais** «Analyse comparative, intra et inter espèces, de transcriptases de levures» Thèse Doctorat, Université Paris , septembre 2005

[15] **C. Duby, S. Robin** « Analyse en Composantes Principales » université Paris, juillet 2006

[16] **A. Baccini** «Statistique Descriptive Multidimensionnelle», Université Paul Sabatier, Toulouse, mai 2010

[17] «L'analyse en composantes principales» (pdf).

[18] **B.M.Dorsaf. «**Réduction de données pour le traitement d'images**»,** thèse Magister ,université de Mentouri Constantine, 2009.

[19] **R. Genuer** « forêts aléatoires : aspects théoriques, sélection de variables et applications» université pris novembre 2010
[20] **M. BOUBOU** « Contribution aux méthodes de classification non supervisée via des approches pretopologiques et d'agrégation d'opinions» Thèse de Doctorat de, Université Claude Bernard - Lyon I, Novembre 2007
[21] **C.BOUVEYRON** « modélisation et classification des données de grande dimension application à l'analyse d'images » thèse, université joseph fourier – grenoble 1, 2006
[22] **M. Maazaoui** «Classification audio par approche hybride SVM-HMM», Mémoire de Master Tunis, Décembre 2008
[23] **R. Rakotomalala** « Pratique de la Régression Linéaire Multiple Diagnostic et sélection de variables» Université Lumière Lyon 2, Jul-201
[24] **A. Guyader**« Régression linéaire» 1ere éd, Edition, Université Rennes 2, Paris, 2010
[25] **M. R. Zemouri**, « Contribution à la surveillance des systèmes de production à l'aide des réseaux de neurones dynamique, Application à la e-maintenance », Thèse de doctorat, Université de Franche, Comté, 2003.

[26] **A. BENSAAD, F. FERTAS,** «reconnaissance automatique de visage par réseaux de neurones», Université de M'sila, 2005.

[27] **F. BABUS** «Contrôle de processus industriels complexes et instables par le biais des techniques statistiques et automatiques», Thèse de doctorat , Université d'Angers ,2008

[28] **M.AMMAR** «**M**ise en œuvre de réseaux de neurones pour la modélisation de cinétiques réactionnelles en vue de la transposition batch/continu. Thèse de Doctorat, Paris ,juillet 2007
[29] **R.DUBOIS** « Application des nouvelles méthodes d'apprentissage à la détection précoce d'anomalies en électrocardiographie », thèse de doctorat de l'université paris 6 janvier 2004
[30] **A.SEGHIOUR**« Contrôle Non Destructif a base d'Ultrasons Application au contrôle de qualité des matériaux de construction» Mémoire de Master**,** Université de M'sila, 2011
[31] **David G. Stork** « Classification Toolbox» For use with MATLAB®

[32]**Howard Demuth** «Neural Network Toolbox» *Version 4*User's Guide

[33] «*Numerical Analysis*» Second Edition, Using MATLAB®

MINISTERE DE LENSEGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

Université de M'sila
Faculté de Technologie
Département d'Electronique
Option : Instrumentation et Maintenance Industrielle

Année Universitaire : 2011/2012

Proposé et dirigé par : Mohamed LADJAL

Étudié par : Ameur SLIM

Intitulé : Contribution de l'analyse multivariée à l'étude de la régression et la classification supervisée des données environnementales.

Résumé :

Une des principales caractéristiques des données environnementales réside dans le fait qu'on dispose d'un nombre de données de sorties très faible devant un grand nombre d'entrées. Un enjeu majeur du traitement statistique de ces données est l'analyse multivariée à but décisionnel. D'un point de vue statistique, ce grand nombre de variables devant ln petit nombre d'observations rend l'analyse multivariée difficile. Une façon de contourner ce « fléau » de la dimension consiste à réduire cette dimension. Dans ce travail, la régression et la classification supervisée est vue comme un problème de reconnaissance de formes avec peu d'observations et beaucoup de variables d'entrées dans un système pour l'aide à la décision. Parmi les techniques de réduction de dimension en trouve : l'analyse en composantes principale (ACP), l'analyse en composantes indépendantes (ACI), …. L'objectif de ce mémoire consiste à illustrer pertinence parmi de ces approches proposées lorsqu'elles sont appliquées à l'analyse de deux jeux de données réels différents. La régression et la classification supervisée obtenue en combinant les techniques d'analyse multivariée et les réseaux de neurones artificiels comme technique de reconnaissance de formes sont appliquées au domaine de surveillance des eaux. On cherche une évaluation complète expriment les variables indicatrices dans notre application et voir l'effet de l'analyse multivariée sur la régression et la classification supervisée. Cette décision est basée sur l'application de la technique choisie et sur l'interprétation des informations réduites et complètes obtenues sur tout l'ensemble de deux bases de données.

Mots clés :

Réduction de dimension, Analyse multivariée, régression, classification, réseaux de neurones artificiels, simulation.