REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE DE TECHNOLOGIE
DEPARTEMENT D'HYDRAULIQUE
N° :……..

DOMAINE : TECHNOLOGIE
FILIERE :HYDRAULIQUE
OPTION : HYDRAULIQUE URBAINE

Mémoire présenté pour l'obtention
Du diplôme de Master Académique

Réalisé par: OUANAS Hanane

et

LAMOUAT Manel

Intitulé

Application des réseaux de neurones artificiels pour l'estimation des lacunes dans les enregistrements de précipitation : Application sur le bassin versant de la Soummam

Encadré par :

Mr. DJERBOUAI.S

Année universitaire : 2021/2022

Remerciement

A l'occasion d'écrire ce mémoire, nous remercions Allah de jouir de la force et la patience pour achever ce travail.

Tout d'abord, un grand merci à notre respecté encadreur Monsieur « Salim Djerbouai » du temps qui nous a accordé et des bons conseils qui nous a donnés au cours de l'encadrement tout en laissant également toute la liberté à développer nos propres idées et d'être toujours présent à nous alerter pour que nous ne s'égarions pas dans de faux chemins.

Notre sincère merci s'adresse aussi à tout le membre du juré qui s'occupera de l'examen de notre travail.

Nous tenons à adresser un grand merci à tous les enseignants du département de l'hydraulique.

Sans oublier enfin à remercier tous nos amis et camarades du département de l'hydraulique.

Dédicace

Je dédie ce modeste travail à :

ma famille qui a tout donné pour que je sois à ce niveau.

A ma Chère Mère Naaima et au meilleur père au monde Amar

Que dieu les garde et leur procure la gaité et le bon heur ainsi qu'une longue vie.

À mon meilleur ami Houda pour son assistance et son soutien.

Manel

Dédicace

Je dédie ce travail

A mes chères parents qui sans eux je n'aurai jamais pu arriver jusqu'ici ; ils m'ont soutenu le long de ce parcours, ils ont veillé à ce que je ne manque de rien, ils m'ont encouragé à tenir bon et à aller jusqu'au bout et ils n'ont cessé de prier pour moi. Aucune dédicace ne saurait exprimer ce que vous méritiez pour les sacrifices que vous n'avez cessé de faire depuis ma naissance, durant mon enfance et même à l'âge adulte.

Maman et papa je vous aime tellement.

A mon cher frère, mes chères soeurs et leurs enfants pour leurs encouragements permanents, et leur soutien moral.

A toute ma famille et mes proches.

A tous mes chers amis pour les aider et les soutenir dans les moments difficiles.

A tous ceux que j'aime et à tous ceux qui m'aiment.

Hanane

Résumé

Les analyses hydrologiques dépendent principalement de la disponibilité des données pluviométriques, bien que les problèmes liés aux données soient communs et qu'il existe différentes raisons pouvant résulter d'un défaut de la station d'enregistrement, de la variation temporelle et spatiale du phénomène pluviométrique et de son complexité physique.

Dans ce travail nous avons fait une comparaison entre les différentes méthodes utilisées pour le comblement des lacunes pluviométriques mensuelles. Deux types de méthodes ont été utilisées dans le présent travail :

- Les méthodes classiques : IDWM, CCWM.
- Les méthodes basées sur l'intelligence artificielle : Les réseaux de neurones artificiels (Apprentissage en profondeur) et les algorithmes génétiques FFSGAM.

L'étude a été menée dans le bassin versant de la vallée du Soummam, qui porte le numéro (15) selon la numérotation de l'Agence nationale des ressources en eau, en utilisant les valeurs de pluie enregistrées dans cinq stations d'enregistrement.

L'évaluation des méthodes utilisées a été faite en se basant sur les meilleurs critères de comparaison, et au final, nous avons conclu que :

Toutes les méthodes ont donné de bons résultats et la méthode ANN a donné des résultats plus performants que toutes les autres méthodes.

Mots-clés: Bassin de Souamam ;Intelligence artificielle ; Réseaux de neurones ;Apprentissage profond ; Algorithmes génétiques ; Méthodes de pondération ; Estimation des données manquantes.

ملخص

تعتمد التحليلات الهيدرولوجية بشكل أساسي على توافر الهطول على الرغم من أن المشاكل المرتبطة في البيانات المفقودة شائعة وتوجد لأسباب محتلفة قد يكون نتيجة خلل في عمل محطة التسجيل أو التباين الزماني او المكاني لظاهرة هطول الأمطار وتعقيدها من الناحية الفيزيائية.

في هذا العمل قمنا بدراسة بعض الطرق لتقدير القيم وسد الثغرات وإجراء مقارنة للعثور على أفضل طريقة والتي هي الطرق الكلاسيكية CCWM , IDWM وطرق تعتمد على الذكاء الإصطناعي من بينها ، الشبكات العصبية الإصطناعية (التعلم العميق .)الخوارزميات الوراثية ; FFSGAM.

أجريت الدراسة في مستجمع مياه واد الصومام الذي يحمل الرقم (15) حسب ترقيم الوكالة الوطنية للموارد المائية ، بإستخدام قيم الأمطار المسجلة في خمس محطات تسجيل .

قمنا بإجراءعمليات تجريبية لكل الطرق على التسجيلات الشهرية وذلك بإستعمال أفضل معايير المقارنة وفي

الأخيراستنتجنا أن :

كل الطرق أعطت نتائج جيدة ,وان الطريقة ANN كانت هي الأفضل.

الكلمات الدالة : حوض المياه ; التعلم العميق ; الذكاء الإصطناعي ; الشبكات العصبية ; التعلم العميق ; الخوارزميات الوراثية; طرق الترجيح ;آلية التعامل مع البيانات المفقودة .

Abstract

Hydrological analyses depend mainly on the availability of rainfall data, although data problems are common and there are different reasons that can result from a failure of the recording station, the temporal and spatial variation of the rainfall phenomenon and its physical complexity.

in this work we studied some methods to estimate values, fill the gaps and make a comparison to find the best method, which are the classical IDWM, CCWM and AI-based methods, including Artificial neural networks (Deep learning), genetic algorithms FFSGAM.

The study was conducted in the Soumam valley watershed, which is numbered (15) according to the numbering of the National Water Resources Agency, using the rainfall values recorded in five recording stations.

The performance of the proposed methods was tested using the best comparison criteria, finally, we concluded that:

All the used methods gave good results and the ANN method was performed better than the others.

Keywords: Soummam catchment ; Artificial intelligence ; neural networks ; Deep learning ; Genetic algorithms ; Weighting methods ; Missing precipitation data estimation.

Table des matières

Chapitre III Méthodes claassique d'estimation des données manquantes

Chapitre IV Méthodes basées sur l'intelligence artificielle

Chapitre V Application sur des séries pluviométriques du bassin versant d'oued Soummam

Chapitre VI Résultats et interprétations

Liste des tableaux

Liste des Figure

Introduction générale

Introduction générale

Récemment, la gestion de l'eau a rencontré des difficultés de gestion en raison de l'augmentation des besoins en eau liée à la croissance démographique, à la croissance urbaine et aux besoins industriels et agricoles.

Toute étude climatique ou hydrologique est basée sur l'exploitation des séries de données recueillies pendant des périodes plus ou moins longues continues ou discontinues Parfois, ces chaînes contiennent de nombreuses lacunes (valeurs manquantes) la conséquence de différents problèmes d'enregistrement, comme une défaillance mécanique dans le cas des pluviomètres automatiques, une absence temporaire d'observateurs dans le cas de pluviomètres manuels ou encore l'arrêt temporaire.

Le but de ce travail est de tester les méthodes basées sur l'intelligence artificielle à savoir les réseaux de neurones artificiels à apprentissage en profondeur ainsi que les algorithmes génétiques, sur l'estimation des données manquantes et de faire une comparaison avec les méthodes classiques pour trouver la meilleure méthode.

- Méthodes classiques CCWM ,IDWM;
- Méthodes basées sur l'intelligence artificielle (les réseaux de neurones artificiels , l'apprentissage en profondeur) , les algorithmes génétiques .

Pour atteindre ces objectifs, ce mémoire a été organisé en six chapitres principaux :

- Le premier chapitre regroupant un aperçu bibliographique, un point de connaissance actuel sur les différentes méthodes utilisées dans l'estimation des données manquantes dans les enregistrements des précipitations ;
- Le deuxième chapitre est consacré à l'étude d'homogénéisation des données pluviométriques ;
- Le chapitre troisième présente les méthodes classiques utilisées pour estimer les données pluviométriques manquantes ;
- Le quatrième chapitre présente la méthode des données manquantes basée sur l'intelligence artificielle ANNs et FFSGAM.

- Le cinquième chapitre est consacré à la présentation de la zone d'étude et à l'application des différentes méthodes d'estimation comblant les lacunes évoquées dans les deux chapitres précédents ;
- Le sixième chapitre présente les résultats et leurs interprétations ;

  En fin, nous terminerons ce travail par une conclusion générale résumant les principaux résultats.

Chapitre I Etude bibliographique

Chapitre I : Etude bibliographique

Introduction

Le processus d'enregistrement et de traitement des données pluviométriques se heurte à un problème fondamental, à savoir l'absence ou la perte dans la série des précipitations, qui nécessite de nombreuses études de synthèse sur les quantités de précipitations annuelles, et si cela n'a pas de conséquences pratiques lorsqu'on dispose de données très nombreuses, cela peut supprimer tout intérêt à l'étude si le nombre de données restantes est trop faible.

L'objectif de cette partie du mémoire, est de fournir à travers une synthèse bibliographique, un point de connaissance actuel sur les différentes méthodes utilisées dans l'estimation des données manquantes dans les enregistrements des précipitations.

I.1.Méthodes utilisées dans le comblement de lacune

L'estimation des données des précipitations manquantes se fait généralement par :

I.1.1.Méthodes classiques

Les Méthodes de pondération classiques (Smith, 1993), méthodes de pondération basées sur la distance (Simanton et Osborn,1980 :Wei et McGuinness, 1973), Méthodes déterministes d'interpolation non-linéaires et stochastique.

La Régression et analyse des séries chronologiques(Salas, 1993).Des variantes de régression sont proposées par Daly et al. (1994 ,2002).

Le guide de l'hydrologie (ASCE, 1996) recommande les deux méthodes nommées normal-ratio et la Méthode de pondération par la distance inverse (IDWM), une étude comparative d'estimation des données de précipitations en utilisant ces deux méthodes peut être trouvée dans Singh et Chowdhury(1986, 1983).

Récemment une étude comparative entre les différentes méthodes de pondération a été faite par Teegavarapua, et Chandramouli (2005), aux Etats unis. Les données pluviométriques de vingt stations pluviométriques sur une période d'observation de 1971 à 2003 sont utilisées

pour tester les méthodes de pondération suivantes : méthode de pondération par la surface inverse, méthode de pondération par le coefficient de corrélation, méthode de pondération par

la surface inverse modifiée, méthode de pondération par l'exponentielle négative de la distance, méthode de pondération rapprochée, méthode d'estimation basée sur les réseau de neurones artificielle, méthode de krigeage. Sur la base de cette étude, ils ont recommandé les trois méthodes suivantes :

La méthode de pondération par le coefficient de corrélation, méthode d'estimation basée sur les réseaux de neurones artificielle et la méthode de krigeage.

Une comparaison a été faite par Pechlivanidis et al., (2005) entre la méthode de pondération par le coefficient de corrélation et la méthode GLM (Modèle linaire généralisé), l'estimation a été faite sur les données journalières de 17 station pluviométriques sur une période d'observation entre 1991 à 2002 dans la région Thames, U K. ils ont trouvé que la méthode GLM donne des résultats meilleurs que ceux obtenus par CCWM sauf dans le cas ou il existe une forte auto-corrélation spatiale.

Des études de Teegavarapu et de Chandramouli (2005) et Tomczak (1998) ont donnée plusieurs variantes de la méthode IDWM.

Teegavarapu(2009) a employé des règles d'association dans les méthodes d'interpolation spatiales pour améliorer l'estimation des données de précipitations.

Les méthodes d'interpolation spatiales qui utilisent l'analyse de tendance par la surface par des équations polynomiales des coordonnées spatiales (wang, 2006). La régression et également applicable pour l'interpolation spatiale, cependant la sélection de la fonction appropriée pose un problème majeur vu le grand nombre des fonctions qui peuvent être utilisées (Sullivan and Unwin, 2003).

Les méthodes d'interpolation dépendant des variance de surface appartenant à la famille des krigeage ont été utilisées dans les l'interpolation spatiale (Vieux, 2001; Grayson and Bloschl, 2001).

En effet, la méthode de krigeage a été utilisée aussi bien pour l'estimation des données manquantes que pour l'interpolation à partir de mesures ponctuelles(Dingman, 2002; Vieux, 2001; Ashraf et al., 1997).

La méthode co-Krigage de radar a été utilisée par Krajewski (1987) pour estimer la pluie moyenne régionale.

Seo et al.(1990a, b) Seo(1996) ont décrit l'utilisation de la méthode co-krigeage et les indicateurs de krigeage , pour l'estimation des données de précipitations manquantes.

La méthode Krigeage ordinairea été utilisée par Teegavarapu(2007) pour l'estimation des pluies journalières.

Malgré toutes les améliorations des méthodes classiques, des limitations des méthodes d'interpolation spatiales existent toujours .Vieux (2001), Grayson et Bloschl (2001), Sullivan et Unwin (2003), Teegavarapu (2007, 2008, 2009) et Brimicombe (2003) ont discuté les limitations de la méthode IDWM et d'autres méthodes d'interpolation spatiale.

I.1.2.Méthodes basées sur l'intelligence artificielle

Récemment, des modèles empiriques basés sur la théorie de l'évolution des principes de la biologie ont été développées parmi lesquelles nous pouvons citer :

Les algorithmes génétiques,Les réseaux de neurones artificiels et la programmation génétiques.Ces méthodes sont utilisées pour le développement et l'application des modèles inductifs.

Les algorithmes génétiques utilisent une procédure de recherche probabiliste qui utilise des méthodes informatiques basées sur les principes d'évolution naturels (Goldberg, 1989).

Les réseaux de neurones artificiels (ANNs) sont des représentations des modèles numériques du processus de fonctionnement du cerveau humain (Zurada, 1992). L'application des réseaux de neurones artificiels dans la domaine de l'hydrologie n'est pas récente (ASCE, 2001a, b; French et al.,1992; Govindaraju and Rao, 2000).

La performance de la fonction universelle des réseaux de neurones est confirmée Par Cybenko (1989) et Hornik et al. (1989).

La programmation génétique (Koza, 1992) peut être utilisée pour créer des programmes informatiques ou des modèles, La sortie d'une programmation génétique est un modèle empirique utilisé comme une fonction d'approximation (Giustolisi and Savic, 2004).

Ilya des limitations dans l'utilisation des algorithmes génétiques pour avoir des fonctions d'approximation, quelques limitations inclues dans les Tavaux de Rogers and Hopfinger (1994) and Shi et al. (1998).

Une nouvelle technique appelée régression polynomiales évolutionnaire a été utilisée pour la recherche des fonctions d'approximations (Giustolisi and Savic, 2004, Giustolisi et al., 2004).
Une nouvelle approche basée sur les algorithmes génétiques nommée FFSGAM (fixed functional set genetic algorithmm ethod) a été récemment développée par (Tufail and Ormsbee, 2006) dans la but de trouver une fonction optimale d'estimation des donnés manquante.
Une étude comparative entre les trois méthodes FFSGAM (fixed functional set genetic algorithmm ethod), la méthode de pondération par la surfaceet la méthode de pondération par la distance inverse été faite par Ramesh S.V. Teegavarapu, et al. 2009. L'estimation a été faite en utilisant les observations pluviométriques journalières de quinze stations pluviométriques dans la ville Kentucky-États-Unis sur la période 1991 à 2002.Sur la base de cette étude, ils ont conclu que la méthode basée sur les algorithmes FFSGAM est meilleure que les méthodes de pondération classiques.

Récemment, l'application des réseaux de neurones à apprentissage en profondeur (LSTM) a été proposée par DJERBOUAI en 2022. L'étude a montrée que les modèles ANN (LSTM) ont données des résultats plus performants que toutes les autres méthodes.

Conclusion

Dans ce chapitre, nous avons découvert nous avons exploré les différentes méthodes les plus importantes utilisées pour estimer les données pluviométriques manquantes.que nous allons travailler avec elle dans les troisième et quatrième chapitres.

Chapitre II Homogénéisation des Données Pluviométrique

Chapitre II :Homogénéisation des Données Pluviométrique

Introduction

Les données pluviométriques sont probablement les données d'entrée les plus importantes pour n'importe quel modèle hydrologique.

Il est donc nécessaire, avant toute utilisation des variables pluviométriques, qu'une analyse statistique des informations qui aide à la prise de décision fondamentale soit faite afin de détecter les anomalies, rechercher la cause et corriger les écarts par des méthodes appropriées .

II.1.Les principales causes d'hétérogénéité

II.1.1.Modification de l'environnement du site de mesure

- Par déplacement de l'appareil

Ceci est un cas fréquent et souvent la station conservera son nom tout en suivant les déménagements de son observateur . Généralement, ces déplacements sont faibles (de l'ordre du kilomètre) mais, ils peuvent provoquer de grandes différences dans la séries de mesures si on modifie l'exposition de l'appareil ou si le changement d'altitude est important.

- Par modification de l'environnement lui-même

Cett en modification peu têtre brusque (construction proche)ou progressive.

- Par changement de la hauteur de l'appareil

On peut trouver des différences de l'ordre de 50% entre un appareil situé à1,20 m du sol et un autre placé au sol. De telles différences sont dues aux turbulences qui se forment autour de l'appareil dans les lieux très exposés aux vents (Le Goulven, 1988).

II.1.2.Erreurs dues à l'appareil

- Modification de la surface réceptrice par construction, échange ou déformation.

Si les engins « standard» ont une surface réceptrice constante et connue.il n'en est pas de même des pluviomètres totalisateurs de fabrication artisanale.

- Erreurs d'étalonnage

Une erreur d'étalonnage peuvent se produire dans le cas des Pluviographes (Le Goulven, 1988).

II.1.3.Erreurs de mesure ou d'enregistrement

- Au niveau de l'éprouvette

Les erreurs peuvent provenir de précisions différentes d'une éprouvette à l'autre ou de lectures incorrectes lorsque l'éprouvette n'est pas verticale ou bien de confusion de chiffres,etc. Le cas le plus typique est celui de confusion d'éprouvette (Le Goulven, 1988).

- Au niveau de l'enregistrement

-manque d'encre,

-erreurs de dates,

-notation erronée des hauteurs mesurées,

-erreurs de transcription, etc (Le Goulven, 1988).

II.1.4.Erreurs lors de l'archivage et de la publication

Lors de la collecte et de la transcription des données brutes, peuvent se produire des erreurs de copie ou des aise et la publication des archives donne lieu à toutes les erreurs d'écriture, (oubli de dates, erreurs de stations,etc.)(Le Goulven, 1988).

II.2.Nécessité d'effectuer des « tests » d'homogénéité

Les diverses causes d'hétérogénéité et les conséquences de celle-ci montrent la nécessité de contrôler rigoureusement les données pluviométriques, ce qui pourrait se faire, dans la plupart des cas,en consultant l'historique de la station.

L'expérience montre qu'un changement de site coincode généralement avec un changement d'observateur et que la confusion d'éprouvettes(ou réglettes) se produit après une interruption des mesures. Cela signifie qu'un bon historique où soient signalés les changements de site, d'observateurs ou d'appareils et ceux de l'environnement, et une vérification des dimensions des pluviomètres et des éprouvettes(ou réglettes), permettraient de résoudre de nombreux problèmes (Le Goulven, 1988).

Malheureusement, ces historiques sont en général inexistants ou d'accès difficile.

L'historique Peut être partiellement reconstitué à partir des documents originaux envoyés par les observateurs, mais cela est insuffisant pour l'analyse d'une série chronologique complète.

D'où la nécessité d'effectuer des tests d'homogénéité sur la base des simples données annuelles et, ultérieurement, s'il ya un problème difficile, de faire une vérification sur le terrain (Le Goulven, 1988).

II.3.Détection des erreurs et correction des anomalies

Le contrôle visuel des données pluviométriques s'avère toujours efficace et permet de déceler à prime abord les hétérogénéités grossières qui peuvent exister et de les corriger.

D'autres hétérogénéités moins évidentes peuvent exister et n'apparaissent pas lors de ce contrôle. Pour celles-ci, il est obligatoire de recourir à certaines méthodes statistiques pour les déceler (Touaibia, 2004).

Les tests d'homogénéités sont nombreux et peuvent être graphiques ou analytiques.

Dans ce travail nous citons les méthodes les plus utilisées à savoir :

II.3.1.Méthodes graphiques

II.3.1.1.Méthode des doubles masses

Cette méthode permet de déceler graphiquement l'hétérogénéité de la série à étudier et de la corriger.

Elle consiste à comparer les pluies (ou toute autre variables) cumulées d'une station A, à propos de laquelle on éprouve des doutes quant à son homogénéité, avec les pluies cumulés d'une station B dont les mesures sont jugées homogènes (Touaibia, 2004).

II.3.1.1.1.Procédé de la méthode des doubles masses

- Sélectionner comme station de base une station dont les observations sont fiables ;
- Faire le cumul des pluies (annuelles, mensuelles, saisonnières) aux stations A et B;
- Porter ces valeurs sur du papier millimétré, avec les valeurs de B en abscisses et les valeurs de A en ordonnées (Figure II.1).

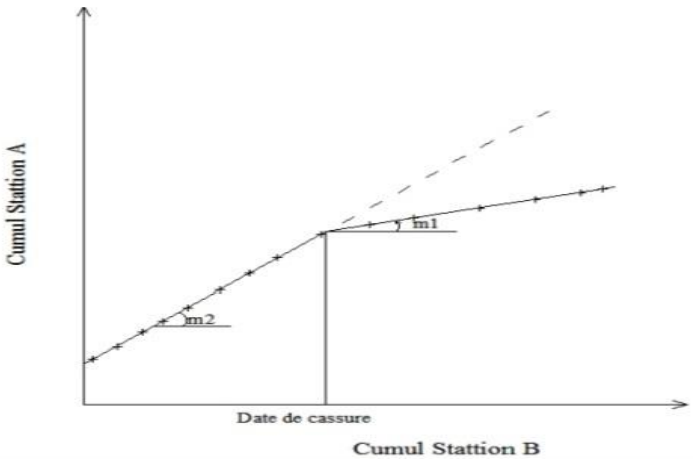


Figure II.1 : Méthode des doubles masses (Touaibia, 2004).

Si les données de la station A contrôlée sont homogène par rapport à celles de la station de base B, la courbe des doubles cumuls avoisine une droite.

Si elle possède une cassure à partir d'un point M, les observations à partir de ce point sont soit fausses soit hétérogènes.

Dans le cas où l'hétérogénéité serait détectée, la correction s'effectuée par modification de la pente de la droite de double cumul des données antérieurs ou postérieurs à la date de la cassure.

Seul le but visé par l'étude en cours indique quelle partie de la série est à corriger.

- Corriger les données observées en multipliant le rapport de pente m_1/m_2 ou m_2/m_1par la valeur erroné respectivement selon que l'on soit après la cassure ou avant (Touaibia,2004).

II.3.1.2.Méthode du cumul des résidus (Paul et Andréf, 1999)

La méthode du cumul des résidus, due à Philippe Bios de l'école nationale supérieure de l'hydraulique de Grenoble, est une extension de l'idée de la méthode des doubles cumuls, à laquelle elle ajoute un contenu statistique autorisant la pratique d'un véritable test d'homogénéité : c'est donc un progrès décisif.

II.3.1.2.1.Procédé de la méthode du cumul des résidus

Soient x_i (série de base), y_i (série à contrôler), l'idée de base consiste à étudier, non pas directement la valeur de $x_i$et y_i(ou $\sum x_i$et $\sum y_i$) mais les cumuls des résidus ε_i de la régression linéaire de y en x :

$Yi= a_0+a_1x_i + \varepsilon_i$ ou encore $\varepsilon_i= y_i – (a_0 + a_1x_i) = y_i – \hat{y}_i$ II.1

De la théorie de la régression il découle que la somme des résidus est nulle et que leur distribution est normale, d'écart type : $\sigma_\varepsilon= \sigma_y\sqrt{1 - r^2}$ II.2

Où r est le coefficient de corrélation linéaire entre X et Y.

Pour un échantillon d'effectif n, le cumul des résidus est défini comme:

$E_0= 0$; $E_j =\sum_i^j \varepsilon_i$ II.3

Quelque soit j =1, n.

Le report graphique des résidus cumulés E_j (en ordonnée) en fonction des numéros d'ordre j des valeurs (en abscisse, j = 0 à n, avec E_0=0) devrait, pour une corrélation avérée entre x et y, donner une ligne partant de 0, oscillant aléatoire auteur de la valeur zéro entre j=0

et j = n. et aboutissant à 0 pour j=n.

La présence d'une inhomogénéité se manifeste par des déviations non aléatoires autour de la valeur nulle.

Bios a décret et testé de nombreux types d'un homogénéités. Il a en outre montré que, pour un niveau de confiance 1- α, le graphe des E_j en fonction de j (j=0 à n) doit être confiné à une ellipse de grand axe n et demi petit axe :

$z_1= -\frac{\alpha}{2}\sigma_\varepsilon \frac{n}{\sqrt{n-1}}$ II.4

Ces développements fournissent un véritable test de l'homogénéité de deux stations.

II.3.2.Méthodes numériques

Plusieurs tests sont utilisés pour s'assurer de l'homogénéité d'une série statistique.

Nous étudierons ici les tests suivants.

II.3.2.1.Test de Wilcoxon (Touaibia, 2004)

C'est plus puissant des tests non paramétriques qui utilise la série des rangs des observations, au lieu de la série de leurs valeurs.

Le test de Wilcoxon se base sur le principe suivant :

Si l'échantillon Y est issu d'une même population que l'échantillon X, l'échantillon XUY (union de X et de Y) en est également issu.

II.3.2.1.1.Procédé du test de Wilcoxon

Soit une série de précipitations de longueur N dont on veut vérifier l'homogénéité, le procédé du test est comme suit:

- On divise la série en deux sous série X, Y de tailles respectivement N_1, N_2, avec $N_1<N_2$
- On classe la série (X+Y) par ordre croissant et on détermine l'origine de chaque valeur.
- On calcul W_x

$$W_x=\sum rangs(X) \quad \text{II.5}$$

- On calcul W_{min} ,W_{max}

$$W_{min}=\frac{(N_1+N_2+1)N_1-1}{2} - u_{1-\alpha/2}\sqrt{\frac{N_1 N_2(N_1+N_2+1)}{12}} \quad \text{II.6}$$

Avec $u_{1-\alpha/2}$: valeur de la variable réduite de Gauss correspondant à une probabilité de 1-α/2

$w_{max}= (N_1+ N_2+ 1)_1 - W_{min}$ II.12

Si $W_{min}<W_x<W_{max}$ l'homogénéité de la série est vérifiée.

II.3.2.2.Test de Mann-Whitney

Il permet de tester l'hypothèse H_0, selon laquelle une série statistique est homogène, c'est-à-dire que éléments qui la constituent proviennent de la même population. En hydrologie, cela veut dire que les conditions qui ont prévalu lors de la collecte des données ou de l'événement du phénomène considéré (pluie, écoulement, évaporation) n'ont pas changé pendant toute la durée de la collecte ou du phénomène.

En d'autres termes, il n'ya pas eu un phénomène extraordinaire qui aurait pu modifier les données hydrologiques considérées comme le changement de site de la station de mesure, la construction d'un barrage qui aurait pu modifier les apports de l'oued, l'urbanisation de la zone étudiée, etc (Touaibia, 2004).

II.3.2.2.1.Précédé du test de Mann-Whitney

Pour appliquer le test de Mann-Whitney on procédé comme suit (Sari, 2002) :

- On divise notre en deux sous-ensembles de tailles respectives N_1N_2, avec $N_2>N_1$. La taille de la taille de l'échantillon originale est $N= N_1+ N_2$
- On classe les valeurs par ordre croissant de 1 à N et l'on note les rangs $R(x_i)$, des éléments du premier sous ensemble et ceux $R(y_i)$ des éléments du second sous-ensemble dans l'échantillon original.
- On définit K et S comme suit :

$$K = L - \frac{N_1(N_1+1)}{2} \quad \text{II.7}$$

$$S= N_1N_2-K \quad \text{II.8}$$

Avec:

$L = \sum_{l=1}^{N_1} R(X_i)$: C'est-à-dire la somme des rangs des éléments du premier échantillon dans l'échantillon original.

K : est la somme des nombres des dépassements de chaque élément du second échantillon par ceux du premier échantillon.
S : est la somme des nombres de dépassements des éléments du premier sous ensemble (ou échantillon) par ceux du second.

- On calcul :

$$\overline{K} = \overline{S} = \frac{N_1N_2}{2} \quad \text{II.9}$$

$$S_K=S_S = \frac{N_1N_2}{2}(N_1 + N_2 + 1) \quad \text{II.10}$$

- On peut alors tester l'hypothèse H_0 selon laquelle les deux sous échantillon

- proviennent de la même population au niveau de signification α en comparant la grandeur:

$$T = \left|\frac{k-\bar{k}}{S_K}\right| \qquad \text{II.11}$$

Avec la variable centré réduite ayant une probabilité au dépassement α/2. Si $T < Z_{\alpha/2}$, on accepte H_0.

II.3.2.3.Test de la médiane (ou test de Mood)

Ce test permet de vérifier si une série de données est homogène.

II.3.2.3.1.Procédé du test de la Médiane (Touaibia, 2004)

On classe l'échantillon par ordre croissant .

- On détermine sa médiane M (la médiane une constante de telle sorte que 50 % des x_i lui soient inférieures et 50 % des x_i lui soient supérieures.
- On remplace la série des valeurs non classées par une suite de signe :

+ Pour les $x_i > M$

- Pour les $x_i < M$

- On calcul les quantités N_s et T_s , avec :

N_s: Nombre total de séries + ou –

T_s: Taille de la plus grande série de + ou de –

$N_s$suit approximativement une loi normale de moyenne $\frac{1}{2}(N + 2)$ et de variance $\frac{1}{4}(N - 4)$ et T_s suit une loi binomiale, ceci permis d'établir pour un seuil de signification compris entre 91 % et 95 %, les conditions du test sont les suivantes :

$$N_s > \frac{1}{2}\left(N + 1 - u_{1-\frac{\alpha}{2}}\sqrt{N+1}\right) \qquad \text{II.12}$$

$$T_s < 3.3(\log_{10}(N) + 1) \qquad \text{I1.13}$$

Si les conditions du test sont vérifiées, on conclut que la série à étudier est homogène au seuil de signification 1- α.

Conclusion

Dans ce chapitre nous avons vu quelques méthodes de détection et de correction des hétérogénéités dans les séries pluviométriques. Pour les méthodes d'extension des séries pluviométriques nous allons les traités dans les chapitres suivants.

Chapitre III Méthodes classiques d'estimation des données manquantes

Chapitre III : Méthodes classiques d'estimation des données manquantes

Introduction

Dans le domaine de l'hydrologie, faire face à des enregistrements pluviométriques avec des données manquantes est inévitable. Le comblement des lacunes pluviométriques a toujours été une tâche difficile en raison de la variabilité spatiotemporelle de la pluviométrie.

Dans ce chapitre, les différentes méthodes classiques utilisées pour l'estimation des données pluviométriques manquantes sont présentées.

III.1.Méthodes simples

Si on dispose des données complètes des stations voisines, on peut alors utiliser les méthodes suivantes (Llamas, 1993):

- Remplacer la valeur manquante par celle de la station la plus proche. Cette méthode est généralement utilisée pour compléter les pluies annuelles ;
- Remplacer la valeur manquante par une simple moyenne arithmétique des stations voisines. Cette méthode est utilisée lorsque les précipitations moyennes annuelles de la station X (dont on veut compléter l'information) sont égales aux moyennes annuelles des stations voisines à 10% prés ;
- Si l'écart entre les précipitations moyennes annuelles de la station X et celles des stations voisines est supérieur à 10 %, alors les précipitations manquantes de X peuvent être estimées par les moyennes pondérées par les tendances annuelles des stations voisines, donnée par la formule suivante :

$$P_X = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\bar{P}_X}{\bar{P}_i}P_i\right) \qquad \text{III.1}$$

Où

P_X : Valeur manquante ;

n : Nombre de stations de références ;

P_i : Précipitation à la station i, correspondante à P_X;

$\overline{P_i}$: Précipitation moyenne annuelle à la station i ;

$\overline{P_X}$: Précipitation moyenne annuelle à la station.

III.2.MéthodeIDWM (Inverse distance weightingméthod)

La méthode de pondération par ladistance inverse (reciprocal-distance), (Wei and McGuinness, 1973) est très utilisée pour l'estimation des données de précipitations manquantes. Cette méthode estime la valeurd'observation manquante,P_m, en utilisant les valeurs observées dans d'autres stations et la distance d_{mi}, elle est donnée par Teegavarapu et al., 2009):

$$P_m = \frac{\sum_{i=1}^{n} P_i d_{mi}^{-k}}{\sum_{i=1}^{n} d_{mi}^{-k}} \qquad \text{III.2}$$

Où :

P_m : Observation dans la station de base m ;

P_i : Observation dans la station i ;

n : Nombre de stations ;

d_{mi} : Distance entre la station i et la station m ;

k :varie entre de 1 à 6.La valeur la plus utilisée de k est 2.

III.3.MéthodeCCWM (Coefficient of correlation weighing method)

Le succès de la méthode de pondération par la distance inverse (IDWM) dépend fortement de l'existence d'une forte auto-corrélation spatiale positive (Teegavarapu et al., 2009).

Le coefficient de corrélation permet de quantifier la force de l'auto-corrélation spatiale, donc on peut l'utiliser comme un coefficient de pondération. Dans la méthode CCWM, les coefficients de pondération (d_m) sont remplacés par des coefficients de corrélations, on obtient la formule suivante (Teegavarapu et al, 2009) :

$$P_m = \frac{\sum_{i=1}^{n} P_i R_{mi}}{\sum_{i=1}^{n} R_{mi}} \qquad \text{III.3}$$

Où :

P_m : Observation dans la station de base m ;

P_i : R_{mi} : Coefficient de corrélation spatial entre la station i et celle de base m.

III.4.Méthode basée sur la corrélation

III.4.1.Généralités (Sari, 2002,Touaibia, 2004)

La régression et la corrélation consistent en l'étude des liaisons existent entre deux ou plusieurs variables. En hydrologie, elles constituent l'outil mathématique le plus ancien et le plus largement utilisé, dont les buts sont multiples:

- Extension dans le temps des séries d'observations hydrologiques de courtes durées ;
- Prévision des grandeurs hydrologiques (écoulement à partir des conditions hydrométéorologiques observées : pluies, températures…..);
- Extension géographique à des bassins non observés de caractéristiques hydrologiques déterminées sur divers bassins de régime analogue ;
- Etude de la dépendance entre les valeurs successives d'une série de données hydrologiques (série chronologiques).

Certaines grandeurs hydrologiques peuvent à la foi ne pas être indépendante et cependant ne pas être liées par une relation fonctionnelle : on dit qu'il existe entre elles une dépendance stochastiques (c'est-à-dire processus soumis au hasard) et font l'objet d'une étude statistique.

Une dépendance rigoureusement fonctionnelle correspond à une conception théoriquen'est jamais vérifiée en hydrologie.

On dit qu'il ya une corrélation entre deux variables observées, lorsque les variations des deux variables se produisent dans le même sens (corrélation positive), ou lorsque les variations sont dans le sens contraire (corrélation négative).

III.4.2.Définitions

- **Régression**: c'est une méthode de recherche d'une relation exprimant le lieu entre une variable dépendante Y et une ou plusieurs variables dites indépendantes.
- **Corrélation :** c'est une méthode de recherche de la liaison qui existe entre deux variables aléatoires.

On peut calculer la corrélation existant entre n'importe quelles variables aléatoires. Des corrélations très élevées mais qui n'ont aucune signification sont très fréquentes, donc on Observation dans la station i n'entreprend une corrélation que lorsque la dépendance entre les variables peut être expliquée (Touaibia, 2004).

- **Diagramme de dispersion**

L'existence d'une corrélation entre deux variables peut être décelée graphiquement. Il s'agit de reporter les couples d'observations (x_i,y_i) sur un graphique en prenant pour abscisse la variable x, et pour ordonnée la variable y. Chaque point du graphique représente simultanément la valeur x_i, et la valeur y_i .Le graphique résultant constitue un nuage de points appelé : Diagramme de dispersion.

III.4.3.Choix du modèle de régression

Lorsque le diagramme de dispersion est linéaire ou approximativement linéaire, on peut s'efforcer de rechercher l'équation de la droite qui s'y ajuste le mieux. Cette droite de régression de régression de Y en X est généralement déterminée par la méthode des moindres carrées. Dans la pratique, on s'efforce toujours de trouver une régression linéaire même s'il faut faire une transformation dans la relation fonctionnelle. Les différents modèles existant sont(Touaibia, 2004):

Le modèle linéaire représenté par l'équation de la droite : $Y=A+BX$

Les modèles curvilinéaires, à savoir :

Le modèle puissance $Y=AX^B$

Le modèle exponentiel $Y=Ae^{BX}$

Le modèle parabolique $Y=A+BX+CX^2$

III.4.4.Conditions préalables à l'homogénéisation par la Régression

La mise en œuvre d'une opération d'homogénéisation par régression exige que certaines conditions soient satisfaites, à savoir(Laborde,2003) :

- Il faut que la relation soit linéaire ou linéairesable;

- ❖ Il faut que les variables confrontées suivent une loi normal pour qu'on puisse estimer les variances des échantillons étendus et le gain d'information ainsi obtenu ;
- ❖ Il faut que les réalisations successives des variables soient indépendantes.

n'entre prend une corrélation que lorsque la dépendance entre les variables peut être expliquée (Touaibia, 2004).

III.4.5.Régression linéaire simple

III.4.5 .1.Coefficient de corrélation

C'est l'indice qui mesure l'intensité de la liaison linéaire entre deux variables, qui est un nombre sans dimension, il est donné par :

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad \text{III.4}$$

En raison de la symétrie de sa définition, le coefficient de corrélation mesure aussi bien l'intensité de la liaison entre y et x qu'entre x et y.

Le coefficient de corrélation est indépendant des unités de mesure de x et de y.

La valeur du coefficient de corrélation peut varier entre -1,(corrélation négative et parfaite) et +1 (corrélation positive et parfaite). Plus les points sont étroitement alignés selon une droite, plus la valeur du coefficient de corrélation sera élevée et s'approchant de +1 ou -1 selon le cas).

III.4.5.2.Droite de régression-méthode des moindres carrés

La droite de régression c'est la droite qui s'ajuste le mieux aux observations, elle constitue un outil de prévision. On pourra estimer ou prévoir, à l'aide de cette équation, les valeurs d'une variable à partir des valeurs prises par l'autre variable.

Pour la régression linéaire, la droite de régression de Yen X est généralement déterminée par la méthode des moindres carrés, qui consiste à minimiser la somme des carrés des écarts entre les points observés et les points correspondants sur la droite.

Soit un échantillon de n couples d'observations (x_i,y_i) et soit l'équation de la droite :

$$\hat{y}=b_0 + b_1 x_i \quad \text{III.5}$$

Où :

b_0 : Ordonnée à l'origine;

b_1 : Pente de la droite ;

$\hat{y}$: Représente la valeur estimée de la variable dépendante pour une valeur particulière x_i de la variable explicative (indépendante).

Soit ei l'écart verticale entre la valeur observée y_i et l'estimation $\hat{y}$obtenue par la droite de régression pour $X=x_i$.

$$ei= y_i- \hat{y} = y_i- b_0 - b_1 x_i, \text{ pour } i=1,\ldots.n. \quad \text{III.6}$$

La Somme des carrés de ces écarts pour l'ensemble des points est égale à:

$$S= e1^2 + e2^2+\ldots\ldots\ldots\ldots en^2 = \sum_{i=1}^{n}(yi - \hat{y})^2 = \sum_{i=1}^{n}(yi - b0 - b1xi)^2 \quad \text{III.7}$$

La méthode des moindres carrés permet de déterminer les expressions b_0 et b_1 de telle sorte que la somme S soit minimale. La droite obtenue est dite droite des moindres carrés, ou droite de régression. On trouve:

$$\begin{cases} b_1 = \frac{\sum_{i=1}^{n}(x_i y_i)+n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = r\frac{S_x}{S_y} & \text{III.8} \\ b_0 = \bar{y} - b_1\,\bar{x} & \text{III.9} \end{cases}$$

Où :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i ; \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i ; S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \quad ; S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$$

III.4.5.3.Conduite des calculs pour l'extension des séries de totaux pluviométriques annuels

Soient deux variables x et y , x observée n fois et y observée k fois avec n> k. Soit k le nombre de couples (x,y). On se propose, à partir de ces k couples d'établir la droite de régression de y en x puis, à partir des valeurs de x, reconstituer les (n-k) valeurs de y non-observées(Sari,2002).

Soient :$\bar{x}_k$;$\bar{y}_k$;σ_x; $\sigma_y$les moyennes et les écart-types déterminés à partir des k couples ainsi que le coefficient de corrélation r_k correspondant.

La régression de y en x s'écrit:

$$\hat{y}_j = r_k \frac{\sigma_y}{\sigma_x}\left(x_j - \bar{x}_k\right) + \bar{y}_k \qquad \text{III.10}$$

$k < j \leq n$

Ainsi seront reconstituée les (n-k) valeurs de y qui manquent.

L'estimation de la moyenne des y de l'échantillon étendu$\bar{\hat{y}}_n$ peut s'obtenir directement de$\bar{x}_n$ comme suit:

$$\hat{y}_j = r_k \frac{\sigma_y^k}{\sigma_x^k}(\bar{x}_n - \bar{x}_k) + \bar{y}_k \qquad \text{III.11}$$

On peut estimer l'écart-type de l'échantillon étendupar :

$$\hat{\sigma}_y^2 = \sigma_y^2 + r_k^2 \frac{\sigma_y^2}{\sigma_x^2}(\sigma_x^2 - \sigma_x^2) \qquad \text{III.12}$$

III.5.5.4.Moyen d'appréciation du gain obtenu par l'extension

Le bénéfice de l'extension de la série Y à l'aide de la série X est d'autant plus grand que le coefficient de corrélation est élevé. Ce bénéfice a été traduit par R.Véron en efficacité relative E, qui s'exprime selon l'équation suivante(Sari,2002) :

$$E = 1 + \left(1 - \frac{K}{n}\right)\left(\frac{1-(k-2)r^2}{k-3}\right) \qquad \text{III.13}$$

Où :

r : C'est le coefficient de corrélation calculé sur k années ;

E : Efficacité relative de qui varie de k/n à n.

Ce bénéfice est traduit, en utilisant E sous la forme d'un gain réel d'information que l'on exprime à l'aide du nombre d'années<< efficaces>> ou <<fictives>> ń, à laquelle correspond l'échantillon y étendu.

ń varie de k (aucun gain, car corrélation nulle entre y et x avec r = 0) à n (gain maximum, liaison fonctionnelle entre x et y et r = 1).

$$ń = \frac{k}{E} \qquad \text{III.14}$$

On admet que la série y étendue correspond en poids d'information à ce que donnerait une série y réellement observée durant ń années (Sari, 2002).

Remarque

L'extension de séries hydrologiques par la à régression n'est admissible qu'à la condition que les liaisons soient linéaires entre séries courtes et longues et également que les variables soient normales .Ce qui n'est pas le cas pour les variables mensuelles et saisonnières.

Si la normalité des variables n'est pas sûre, mais si la linéarité existe, on peut adapter la méthodologie de l'extension comme suit (Laborde, 2003):

a) On fait d'abord l'extension telle que décrite sur les totaux annuels.

b) On établit ensuite graphiquement les liaisons linéaires entre séries mensuelles, ou pluie mensuelles, Y à étendre et séries X de base, ceci pour la période commune de k années.

c) On estime point par point sur la droite de régression les n-k valeurs de la séries **Y** non observées ; ces deux opérations correspondent à l'application de l'équation (II.10) mais dans laquelle, les lois n'étant pas normales, le coefficient k^y **xy** n'a plus la signification d'un coefficient de corrélation.

d) On est alors en possession de plusieurs séries mensuelles ou pluie mensuelles de la station Y chacune desquelles composées de k valeurs observées et n-k reconstituées.il faut maintenant faire les sommes des n-k valeurs reconstituées par mois ou groupe de mois afin d'obtenir les n-k valeurs annuelles correspondantes.

e) On confronte enfin pour chaque année j de la période étendue de k à n années, le total annuel Pj obtenu directement ci-dessus en (a) et le total annuel P`j obtenu par sommation des valeurs mensuelles ou pluie mensuelles, puis l'on corrige ces dernières valeurs du produit Pj/P`j afin de les rendre homogènes avec l'estimation globale Pj faite à l'échelle annuelle.

III.4.6.Régression double (Laborde,2003)

III.4.6.1.Equation de la régression double linéaire

Soit une variable z que l'on désire expliquer à partir de deux variables x et y. On se propose de trouver une relation linéaire de la forme :

$$z = ax + by + c + \varepsilon \qquad \text{III.15}$$

Les paramètres a, b et c étant déterminés de façon à minimiser la somme des carrés des écarts ε.

$$\varepsilon_i{}^2 = (z_i - ax_i - by_i - c_0)^2 \qquad \text{III.16}$$

Ecrire que $\Sigma\varepsilon_i{}^2$ est minimum revient à écrire que les dérivées partielles de $\Sigma\varepsilon_i{}^2$ par rapport à a, b et c que l'on veut déterminer sont nulles:

$$\begin{cases} \frac{\partial \Sigma \varepsilon_i{}^2}{\partial a} = 0 = 2\Sigma x_i(z_i - ax_i - by_i - c_0) = 2\Sigma x_i \varepsilon_i = 0 \\ \frac{\partial \Sigma \varepsilon_i{}^2}{\partial b} = 0 = 2\Sigma y_i(z_i - ax_i - by_i - c_0) = 2\Sigma y_i \varepsilon_i = 0 \\ \frac{\partial \Sigma \varepsilon_i{}^2}{\partial c} = 0 = 2\Sigma(z_i - ax_i - by_i - c_0) = 2\Sigma \varepsilon_i = 0 \end{cases}$$

On remarque comme pour la régression simple, que la méthode des moindres carrés donne pour solution des paramètres a, b et c tels que les erreurs soient indépendantes de x et de y (orthogonalité établie par les deux premières équations) et nulles en moyenne (troisième équation).

La résolution de ce système de trois équations à trois inconnues ne présente pas de difficultés. Les paramètres a, b et c peuvent s'exprimer en fonction des moyennes, écarts-types et coefficients de corrélation de x, y et z

Où :

$$\bar{x} = \frac{\Sigma x}{n} \; ; \qquad \sigma_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}}$$

$$\bar{y} = \frac{\Sigma y}{n}; \sigma_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n}}$$

$$\bar{z} = \frac{\Sigma z}{n}; \sigma_z = \sqrt{\frac{\Sigma(z_i - \bar{z})^2}{n}}$$

$$p = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{(n-1)\sigma_x\sigma_y} \; ; r_1 = \frac{\Sigma(z-\bar{z})(x-\bar{x})}{(n-1)\sigma_z\sigma_x} \; ; \; r_2 = \frac{\Sigma(z-\bar{z})(y-\bar{y})}{(n-1)\sigma_z\sigma_y}$$

Ces trois coefficients de corrélation seront appelés par la suite coefficients de corrélation totale entre x et y, z et x, et z et y.

On a alors, tout calcul fait :

$$\begin{cases} a = \frac{r_1 - r_2 p}{1 - p^2} \frac{\sigma_z}{\sigma_x} & \text{III.17} \\ b = \frac{r_2 - r_1 p}{1 - p^2} & \text{III.18} \\ c = \bar{z} - a\bar{x} - b\bar{y} & \text{III.19} \end{cases}$$

III.4.6.2.Coefficient de corrélation multiple et variance résiduelle

Pour la corrélation simple, on a vu que le coefficient de corrélation totale mesurait la dispersion des écarts ε_i. On peut donc construire de même un coefficient de corrélation multiple R qui mesurera la dispersion desε_i:

$$\varepsilon_i = (z_i - ax_i - by_i - c_0)$$

$$R^2 = 1 - \frac{\partial \Sigma {\varepsilon_i}^2}{\sigma_z^2} \quad \text{III.20}$$

On montre alors que R peut se déduire de r1, r2 et p par l'expression :

$$R^2 = \frac{r_1^2 + r_2^2 - 2pr_1r_2}{1 - p^2} \quad \text{III.21}$$

Si x et y sont des variables indépendantes, le coefficient p est nul et l'expression précédente se simplifie en :

$$R^2 = r_1^2 + r_2^2 \quad \text{III.22}$$

Par définition même de R, on montre que l'écart-type$\sigma_{\varepsilon i}$ que l'on peut également noter$\sigma_{z_{xy}}$ est :

$$\sigma_{z_{xy}} = \sigma_{\varepsilon i} = \sigma_z \sqrt{1 - R^2} \quad \text{III.23}$$

III.4.6.3.Notion de coefficient de corrélation partielle

Nous avons admis que z dépendait à la fois de x et y. Les coefficients de corrélation totale $r_1$et $r_2$entre z, x et y rendent donc mal compte de la liaison entre 2 variables puisque l'on ne tient pas compte de l'influence de la troisième. L'idée est donc de mesurer non pas la corrélation totale entre z et x mais entre z corrigé des variations de y et x.

On définit donc un coefficient de corrélation partielle entre x et z corrigé des variations de y (noté,r_{xz_y}).

Tous calculs faits, les expressions des coefficients de corrélation partielle sont

$$r_{zx_y}{}^2 = \frac{R^2 - r_2^2}{1 - r_2^2} \qquad \text{III.24}$$

$$r_{zy_x}{}^2 = \frac{R^2 - r_1^2}{1 - r_1^2} \qquad \text{III.25}$$

III.4.7.Régressions linéaires multiples(Laborde, 2003)

III.4.7.1.Mise en équation

Supposons que l'on cherche à expliquer une variable y à partir de k variables x. Si y et les x sont tirés d'une loi de Gauss à k+1 dimensions, les paramètres de cette loi de distribution sont :

Les moyennes marginales : $\bar{y}, \bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots \ldots \bar{x}_i, \ldots \ldots \bar{x}_k$

Les écarts-types marginaux : $\sigma_y, \sigma_{x1}, \sigma_{x2}, \sigma_{x3} \ldots \ldots \sigma_{xi}, \ldots \ldots \sigma_{xk}$

Les coefficients de corrélation totale, soit la matrice [r] :

$$\begin{vmatrix} 1 & r_{yx1} & r_{yx2} & \ldots\ldots & r_{yxj} & \ldots\ldots & r_{yxk} \\ r_{x1y} & 1 & r_{x1x2} & \ldots\ldots & r_{x1xj} & \ldots\ldots & r_{x1xk} \\ r_{x2y} & r_{x2x1} & 1 & \ldots\ldots & r_{x2xj} & \ldots\ldots & r_{x2xk} \\ & & & \ldots\ldots & \ldots\ldots & \ldots\ldots & \\ r_{xiy} & r_{xix1} & r_{xix2} & \ldots\ldots & r_{xixj} & \ldots\ldots & r_{x2xk} \\ & & & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ r_{xky} & r_{xkx1} & r_{xkx2} & \ldots\ldots & r_{xkxj} & \ldots\ldots & 1 \end{vmatrix}$$

distribution conditionnelle des y liés par les k x_i est donnée par:

$$\hat{y}_{xi} = a_0 + a_1 x_1 + \cdots + a_i x_i + \cdots a_k x_k \qquad \text{III.26}$$

Il reste alors à évaluer les k+1 coefficients de régression a_i, le coefficient de corrélation multiple R, et les k coefficients de corrélation partielle $r_{yxi_{x1,x2,x3 \ldots\ldots xk.}}$

On détermine alors les a_i par la méthode des moindre carré.

Les dérivées partielles par rapport aux k+1 paramètres devront donc être nulles :

1 équation :$\frac{\partial \varepsilon^2}{a_0} = -2\sum(y - a_0 - a_1 x_1 \ldots \ldots - a_k x_k) = 0$ III.27

(Erreur nulle en moyenne et par conséquent $a_0 = 0$)

k équations du type : $\frac{\partial \varepsilon^2}{a_i} = -2\sum x_i(y - a_0 - a_1x_1 \ldots \ldots - a_kx_k) = 0$ III.28

III.4.7.2.Coefficients de régression, de corrélation multiple et de corrélation partielle

Si dans la matrice [r], on note les lignes et colonnes de 0 à k, les différents paramètres s'expriment en fonction du déterminant de [r] noté Δ et des déterminants Δij des mineurs de [r] obtenus en supprimant la i ème ligne et la j ème colonne de [r],(On utilisera le signe + si i+j est pair et le signe - si i+j est impair).

Les coefficients de régression sont donnés alors par :

$a_i = \frac{\sigma_y}{\sigma_{xi}} \frac{\Delta_{0i}}{\Delta_{00}}$ (k fois) III.29

$a_0 = \bar{y} - \sum_{i=1}^{k} \bar{x}_i$ III.30

Enfin le coefficient de corrélation multiple R est donné par :

$R^2 = 1 - \frac{\Delta}{\Delta_{00}}$ III.31

III.4.7.3.Seuils de signification

Pour le coefficient de corrélation multiple, on considère que R est significatif si la quantité :

$F = \frac{n-(k+1)}{K} \frac{R^2}{1-R^2}$ III.32

Est significativement supérieure à 1.

Pour les seuils de signification des coefficients de corrélation partielle, on utilisera, les tables de Student. Le nombre de degré de liberté ν égal à ν = n - k – 1.

II.4.8.Régression non linéaire

Vue sa complexité, dans de nombreux cas on pourra se sortir d'affaire en linéarisat la fonction envisagée, ainsi par exemple :

- Hyperbole : $y = \frac{x}{ax-b} \rightarrow \frac{1}{y} = a - \frac{b}{x}$;
- Exponentielle : $y = ae^{bx} \rightarrow \ln y = \ln a + bx$;
- Puissance : $y = ax^b \rightarrow \ln y = \ln a + b \ln x$.

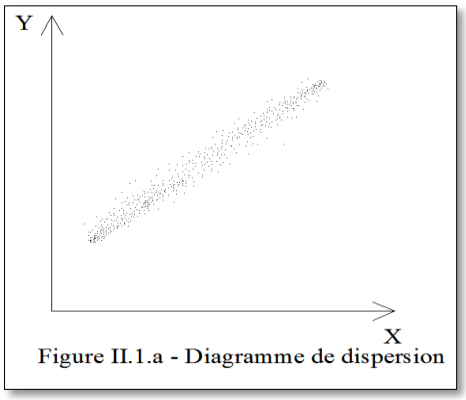


Figure II.1.a - Diagramme de dispersion

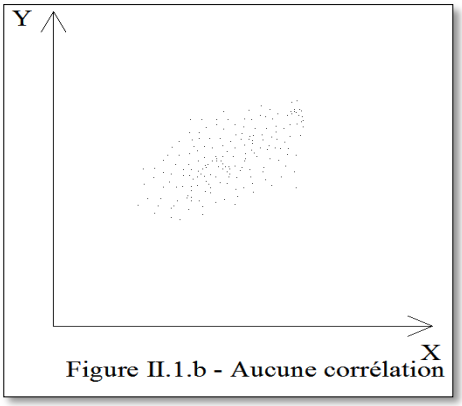


Figure II.1.b - Aucune corrélation

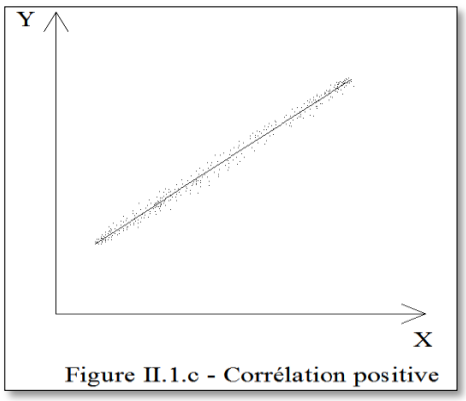


Figure II.1.c - Corrélation positive

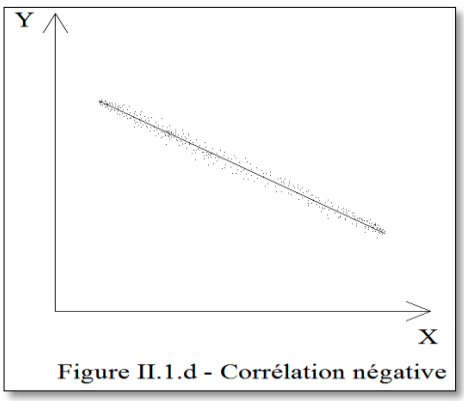


Figure II.1.d - Corrélation négative

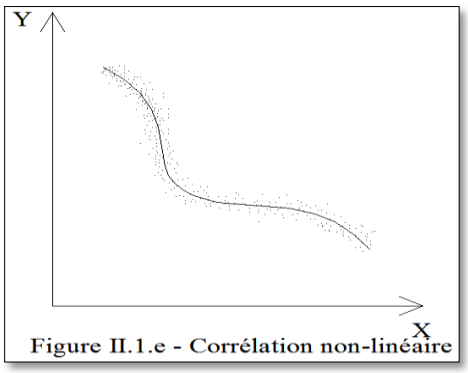


Figure II.1.e - Corrélation non-linéaire

.

Figure III.1.schémas explicatifs

Conclusion

Dans cette partie du travail, les concepts des méthodes classiques utilisées sont expliqués et dans le prochain chapitre, nous verrons d'autres méthodes basées sur l'intelligence artificielle.

Chapitre IV Méthodes basées sur l'intelligence artificielle

Chapitre IV : Méthodes basées sur l'intelligence artificielle

Introduction

L'intelligence artificielle est une branche de l'informatique, qui consiste à résoudre des problèmes pratiques en se basant sur des approches inspirées de la biologie humaine. Ces derniers temps, plusieurs méthodes basées sur la théorie du développement des principes de la biologie, ont été développées.

Dans notre travail nous nous intéressons à l'application de ces techniques à l'estimation des données manquantes dans les enregistrements de précipitations. Deux méthodes ont été choisies, qu'on va présenter dans le présent chapitre, à savoir :

- Les algorithmes génétiques
- Réseaux de neurones artificiels (Deep learning).

IV.1.Les algorithmes génétiques

IV.1.1.Aperçu sur les algorithmes génétiques (AG)

Les algorithmes génétiques sont classés parmi la rubrique générale de la programmation évolutive, qui représentent l'optimisation à travers un processus analogue à celui de la mécanique de la sélection naturelle ou la génétique naturelle dans les sciences biologiques (Goldberg, 1989). Trois processus heuristiques de reproduction, croisement et mutation sont appliqués probabilistiquement à des variables de décision discrètes ayant subi un codage binaire. Au lieu de générer des progressions de solutions uniques comme les autres algorithmes d'optimisation, l'algorithme génétique produit des groupes ou populations de solutions dont la « progéniture » montre des niveaux croissants de précision (valeurs de la fonction objectif)(SOUAG-GAMANE, 2007).

L'algorithme génétique a été utilisé pour résoudre différents problèmes de gestion des systèmes de réservoirs et des ressources en eau en général par plusieurs auteurs tels que : Otero et (1995) ; Olivera et Louks (1997) ; Sharif et Wardlaw (2000) ; Ilich (2001)(SOUAG-GAMANE, 2007).

- Les principaux avantages des algorithmes génétiques sont :

− ils peuvent être directement associés aux modèles de simulation hydrologiques ou de qualité des eaux sans exiger des hypothèses simplificatrices dans le modèle ou de calcul de dérivées ;

− des mesures du taux de recouvrement après l'occurrence d'une défaillance, par exemple, ou la sévérité des conséquences d'une défaillance qui sont difficiles à inclure explicitement dans les méthodes algorithmiques, sont facilement introduites dans un modèle associé algorithme génétique-simulation ;

− la grande capacité des algorithmes génétiques à résoudre des problèmes hautement non linéaires et non convexes.

- L'inconvénient des algorithmes génétiques réside dans :

− les exigences de calcul coûteuses des algorithmes génétiques les rendent mal appropriés pour les problèmes d'optimisation stochastique implicite ou explicite des systèmes de réservoirs, à moins de paramétrer dans un certain sens les politiques de gestion ;

− leur difficulté à expliquer explicitement les contraintes (particulièrement les contraintes d'inégalité) et à maintenir des solutions réalisables dans la population.

Bouchart et Hampart-Zoumian (1999) décrivent une application des algorithmes génétiques pour identifier les séquences de débits appropriées pour l'entraiment d'un modèle d'apprentissage renforcé ; l'apprentissage renforcé procure des stratégies, pour la résolution des problèmes, similaires à des problèmes de programmation dynamique de grande échelle sans avoir besoin d'une connaissance explicite de la fonction de probabilité de transition d'état(SOUAG-GAMANE, 2007).

Cai et al. (2001) décrivent une application des algorithmes génétiques dans la résolution de problèmes non linéaires de grande échelle de planification des ressources en eau sur plusieurs périodes. Les algorithmes génétiques optimisent sur un nombre limité de variables couplées tel que, lorsqu'elles sont fixées, ils permettent la décomposition du problème original en plusieurs petits problèmes de programmation linéaire (SOUAG-GAMANE, 2007).

IV.1.2.Présentation de la méthode FFSGAM

Cette méthode est basée sur le développement d'un modèle inductif, en utilisant les algorithmes génétiques et l'optimisation afin d'obtenir une fonction optimale d'estimation des données manquantes de précipitations.

Le processus de recherche de la fonction optimale en utilisant la méthode FFSGAM se fait en deux étapes.

a) Rechercher les fonctions optimales des variables de décision (ou modèle inputs) en utilisant les algorithmes génétiques (AG) ;
b) Evaluer les coefficients de la fonction optimale choisit en utilisant l'optimisation.

IV.1.2.1.Principe de la méthode FFSGAM

Le principe de la méthode FFSGAM peut être expliqué par l'exemple suivant:

Soient (y) une variable dépendante, et (x_1, x_2) deux variables indépendantes, pour obtenir la fonction empirique qui relie les deux variables indépendantes (x_1, x_2) par la variable dépendante (y) en utilisant FFSGAM, on procède comme suit :

La fonction prédéfinie des variables indépendantes (x_1, x_2) est donnée par l'expression IV.1 :

$$y=a_1F(x_1)\ @b_1\ F(x_2) \qquad \text{IV.1}$$

Où:

Les coefficients a_1 b_1= {nombres réels}

L'operateur mathématique @= {+,-,*, /,^}

F ( )=0, 1, x, log(x),e^x,sin(x),1/x,etc.

La fonction recherchée est une fonction explicite de y (modèle output)en fonction des deux variables dépendantes (x_1 et x_2) (modèle inputs), et la précision de la fonction est basée sur l'erreur moyenne quadratique.

Plus d'une fonction peuvent être obtenues en variant les paramètres et la structure de lafonction prédéfinie.

Les opérations de base des algorithmes génériques (sélection, croisement et mutation) sont utilisées afin de sélectionner les meilleures (simple et faciles à utiliser) fonctions optimales.

Le modèle FFSGAM commence par une sélection aléatoire de solutions pour constituer la population initiale, chaque solution représente une équation explicite pour la variable y.

Les algorithmes génériques travaillent sur une population de solutions possibles en essayant de trouver la solution optimale.

Dans chaque génération, certaine population de solutions améliorent la précision et d'autres sont plus mauvaises.

Les meilleures solutions sont utilisées pour la génération suivante de population afin de continuer le processus de recherche.

D'une génération à une autre le modèle FFSGAM continue de trouver des solutions, il se peut que certains individus soient plus mauvais que leurs parents.

Les solutions améliorées tendent à suivre le processus, et ceux qui sont mauvaises tendront à s'éteindre dans le processus.

A la fin du nombre de génération spécifié, la fonction ayant la précision la plus élevée est sélectionnée comme la structure optimale de l'expression explicite recherchée. Les coefficients de la fonction sont obtenus par l'optimisation.

La méthode FFSGAM peut être illustrée par l'organigramme suivant

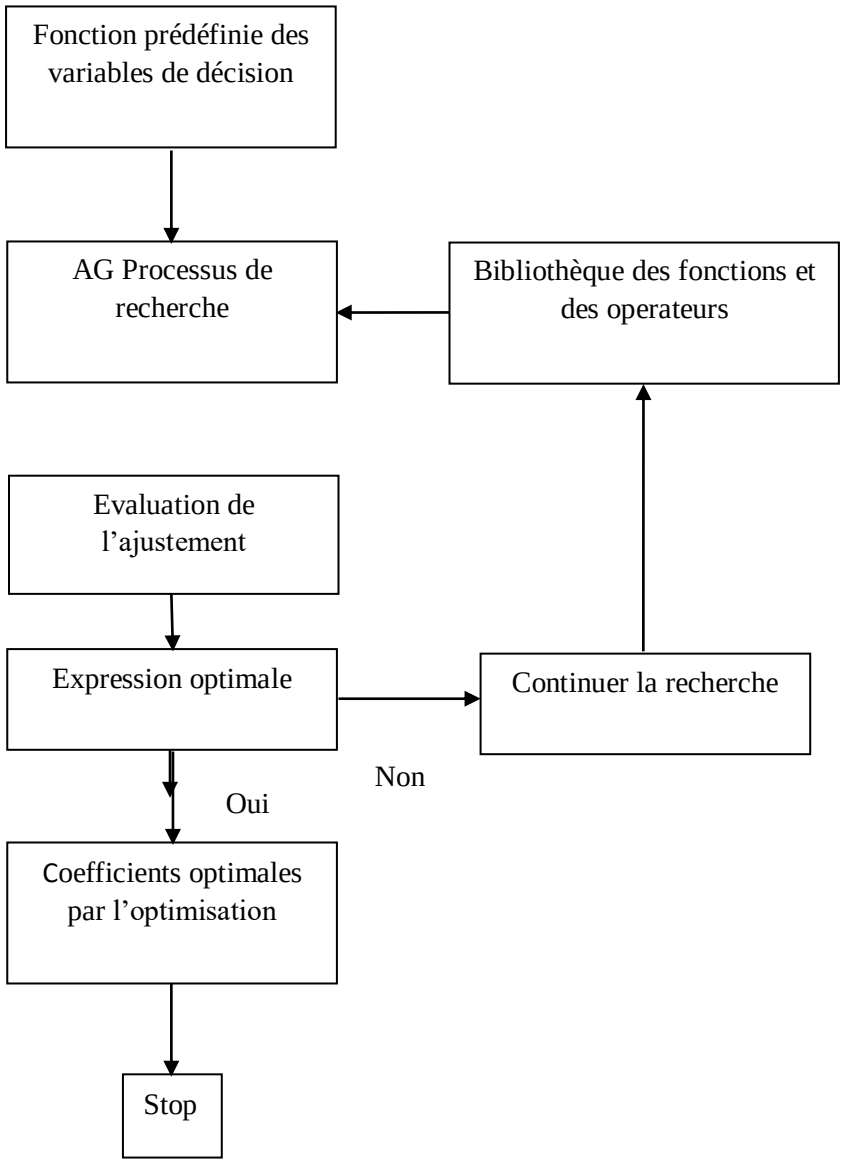


Figure IV.1:Organigramme du modèle FFSGAM

IV.1.2.2.FFSGA modèle d'estimations des donnés manquantes de précipitations

Le modèle FFSGAM d'estimation des données manquantes de précipitations, est établit en utilisant comme variables de décision:

- La distance entre chaque station i et celle de base (dont on veut compléter l'information) ;
- Le coefficient de corrélation entre chaque station i et celle de base.

La fonction prédéfinie est donnée par l'équation IV.2.

Fonction prédéfinie= $\{C_1[\text{fonction-1}(R_{mi})\text{operateur-1fonction-2}(R_{mi})]\}\text{operateur-3}\{C_2[\text{fonction-3}(d_{mi})\text{operateur-1fonction-4}(d_{mi})]\}$ IV.2

Où:

- Les coefficients C_1 C_2 sont des nombre réels ;
- Les operateurs qui peuvent être utilisés dans l'équation IV.2 sont données par le tableau donnés IV.1.

Tableau IV.1: Tableau des operateurs

Operateur #	Operateur
1	+
2	-
3	*
4	/
5	^

- Les fonctions élémentaires qui peuvent être utilisées pour les variables de décision (équation IV.2) sont données par le tableau IV.2.

Tableau IV.2: Tableau des fonctions élémentaires

Fonction #	Fonction f (d_{mi}) ou Fonction f (R_{mi})
1	1
2	d_{mi} ou R_{mi} ou Sqrt(d_{mi}) ou Sqrt(R_{mi})
3	1/ d_{mi} ou 1/R_{mi}
4	Exp (d_{mi}) ou Exp (R_{mi})
5	Log_e (d_{mi}) ou Log_e (R_{mi})
6	Log_{10} (d_{mi}) ou Log_{10} (R_{mi})
7	Exp (1/d_{mi}) ou Exp (1/R_{mi})
8	Log_e (1/d_{mi}) ou Log_e (1/R_{mi})
9	Log_{10} (1/d_{mi}) ou Log_{10} (1/R_{mi})
10	d_{mi}*Exp(d_{mi}) ou R_{mi}*Exp(R_{mi})
11	d_{mi}*Log_e (d_{mi}) ou R_{mi}*Log_e (R_{mi})
12	d_{mi}*Log_{10} (d_{mi}) ou R_{mi}*Log_{10} (R_{mi})
13	(1/d_{mi})*Exp(d_{mi}) ou (1/R_{mi})*Exp (R_{mi})
14	(1/d_{mi})* Log_e (d_{mi}) ou (1/R_{mi})*Log_e (R_{mi})
15	(1/d_{mi})*Log_{10} (d_{mi}) ou (1/R_{mi})*Log_{10} (R_{mi})

Une fois la fonction est obtenue par FFSGAM et par l'optimisation, la précipitation dans la station de base m peut être exprimée par:

$$P_m = \frac{\sum_{i=1}^{n} P_i * (\text{FFSGAMfonction})_i}{\sum_{i=1}^{n} (\text{FFSGAMfonction})_i} \qquad \text{IV.3}$$

En plus des coefficients numériques donnés dans l'équation IV.2, l'équation d'estimation de précipitations donnée par l'équation IV.3 peut contenir des coefficients (pour chaque station), ces coefficients sont estimés par la minimisation de l'erreur moyenne quadratique, par conséquent l'équation d'estimation devient :

$$P_m = \frac{\sum_{i=1}^{n} P_i * c_i * (FFSGAMfonction)_i}{\sum_{i=1}^{n} c_i * (FFSGAMfonction)_i} \quad \text{IV.4}$$

IV.1.2.3.Evaluation des coefficients optimaux

Les coefficients optimaux sont obtenus en minimisant l'erreur moyenne quadratique par la formule suivante:

$$\frac{1}{N}[\sum_{i=1}^{n}(P_i - \widehat{P}_i)^2] \quad \text{IV.5}$$

Où :

$\widehat{P}_i, P_i$: Sont respectivement les valeurs de précipitations estimée et observée dans la station

De base m.

N : Nombre de jour, de mois ou d'années.

Les Quatre fonctions suivantes sont obtenues en utilisant la méthode FFSGAM:

$$P_m = \frac{\sum_{i=1}^{n} P_i C_i [R_{mi} \log_{10}\left(\frac{1}{R_{mi}}\right) - \left(\frac{1}{R_{mi}}\right)\log_{10} R_{mi}][\frac{\log_{10}\left(\frac{1}{d_{mi}}\right)}{\log_{10}(d_{mi})}]}{\sum_{i=1}^{n} C_i [R_{mi} \log_{10}\left(\frac{1}{R_{mi}}\right) - \left(\frac{1}{R_{mi}}\right)\log_{10} R_{mi}][\frac{\log_{10}\left(\frac{1}{d_{mi}}\right)}{\log_{10}(d_{mi})}]} \quad \text{IV.6}$$

$$P_m = \frac{\sum_{i=1}^{n} P_i C_i [\left[\frac{\exp(R_{mi})}{\left(\frac{1}{R_{mi}}\right)\log_{10}\left(\frac{1}{R_{mi}}\right)}\right] + [\sqrt{d_{mi}} \log_{10}\left(\frac{1}{d_{mi}}\right)]]}{\sum_{i=1}^{n} C_i [\left[\frac{\exp(R_{mi})}{\left(\frac{1}{R_{mi}}\right)\log_{10}\left(\frac{1}{R_{mi}}\right)}\right] + \left[\sqrt{d_{mi}} \log_{10}\left(\frac{1}{d_{mi}}\right)\right]]} \quad \text{IV.7}$$

$$P_m = \frac{\sum_{i=1}^{n} P_i C_i [\left[\frac{(\log_{10}\left(\frac{1}{R_{mi}}\right))^2}{R_{mi}}\right][\log_{10}\left(\frac{1}{R_{mi}}\right)]]^2}{\sum_{i=1}^{n} C_i [\left[\frac{(\log_{10}\left(\frac{1}{R_{mi}}\right))^2}{R_{mi}}\right][\log_{10}\left(\frac{1}{R_{mi}}\right)]]^2} \quad \text{IV.8}$$

$$P_m = \frac{\sum_{i=1}^{n} P_i C_i \left[\frac{R_{mi}\log_{10}\left(\frac{1}{R_{mi}}\right)\log_{10}(R_{mi})}{\left[\frac{\left(\frac{1}{R_{mi}}\right)\ln(R_{mi})}{\left(\frac{1}{d_{mi}}\right)\ln(d_{mi})}\right]}\right]}{\sum_{i=1}^{n} C_i \left[\frac{R_{mi}\log_{10}\left(\frac{1}{R_{mi}}\right)\log_{10}(R_{mi})}{\left[\frac{\left(\frac{1}{R_{mi}}\right)\ln(R_{mi})}{\left(\frac{1}{d_{mi}}\right)\ln(d_{mi})}\right]}\right]} \qquad \text{IV.9}$$

IV.2.Les réseaux de neurones

IV.2.1.Historique sur les réseaux de neurones

Sous le terme réseaux de neurones, on regroupe aujourd'hui un certain nombre de modèles dont l'intention est d'imiter certaines des fonctions du cerveau humain en reproduisant certaines de ses structures de base. Historiquement, les origines de cette discipline sont très diversifiées.

En 1890 : W. James, célèbre psychologue américain introduit le concept de mémoire associative, et propose ce qui deviendra une loi de fonctionnement pour l'apprentissage sur les réseaux de neurones connue plus tard sous le nom de loi de Hebb.

En 1943 : J. Mc Culloch et W. Pitts laissent leurs noms à une modélisation du neurone biologique (un neurone au comportement binaire). Ceux sont les premiers à montrer que des réseaux de neurones formels simples peuvent réaliser des fonctions logiques, arithmétiques et symboliques complexes (tout au moins au niveau théorique).

En 1949, D. Hebb initie dans son ouvrage "The Organization of Behavior", la notion d'apprentissage. Deux neurones entrant en activité simultanément vont être associés (c'est-à dire que leurs contacts synaptiques vont être renforcés). On parle de loi de Hebb et d'associationnisme.

IV.2.1.1.Les premiers succès

En 1958, F. Rosenblatt développe le modèle du Perceptron. Il construit le premier neuro-ordinateur basé sur ce modèle et l'applique au domaine de la reconnaissance de formes. Notons qu'à cet époque les moyens à sa disposition sont limités et c'est une prouesse technologique que de réussir à faire fonctionner correctement cette machine plus de quelques minutes.

En 1960, B. Widrow, un automaticien, développe le modèle Adaline (Adaptative LinearElement). Dans sa structure, le modèle ressemble au Perceptron, cependant la loi d'apprentissage est différente. Celle-ci est à l'origine de l'algorithme de rétro-propagation de gradient très utilisé aujourd'hui avec les Perceptrons multicouches.

En 1969, Minsky et Papert publient un ouvrage qui met en exergue les limitations théoriques du perceptron. Limitations alors connues, notamment concernant l'impossibilité de traiter par ce modèle des problèmes non linéaires. Ils étendent implicitement ces limitations à tous modèles de réseaux de neurones artificiels. Après leur ouvrage, il y a abandon financier des recherches dans le domaine (surtout aux U.S.A.), les chercheurs se tournent principalement vers l'intelligence artificielle et les systèmes à bases de règles.

IV.2.1.2.L'ombre

Entre 1967 et 1982, Toutes les recherches ne sont, bien sûr, pas interrompues. Elles se poursuivent, mais déguisées, sous le couvert de divers domaines comme : le traitement adaptatif du signal, la reconnaissance de formes, la modélisation en neurobiologie, etc. De grands noms travaillent durant cette période, tels : S. Grossberg, T. Kohonen, ...

IV.2.1.3. Le renouveau

En 1982, J. J. Hopfield est un physicien, Présente une théorie du fonctionnement et des possibilités des réseaux de neurones. Il explique notamment dans un ouvrage la structure et loi d'apprentissage d'un réseau de neurones correspondant à un résultat escompté. Ce modèle est encore très utilisé aujourd'hui pour des problèmes d'optimisation.

Bien que les limitations du Perceptron mise en avant par M. Minsky ne soient pas levées par le modèle d'Hopfield, les recherches sont relancées.

IV.2.1.4.La levée des limitations

En 1983, La Machine de Boltzmann est le premier modèle connu apte à traiter de manière satisfaisante les limitations recensées dans le cas du perceptron. Mais l'utilisation pratique

s'avère difficile, la convergence de l'algorithme étant extrêmement longue (les temps de calcul sont considérables).

En 1985, La rétropropagation de gradient apparaît. C'est un algorithme d'apprentissage adapté aux réseaux de neurones multicouches (aussi appelés Perceptrons multicouches).Sa découverte a été réalisée par trois groupes de chercheurs indépendants.

IV.2.1.5.Actuellement

De nos jours, l'utilisation des réseaux de neurones dans divers domaines ne cesse de croître. Les applications en sont multiples et variées.

IV.2.2.Fondements biologique (Davalo et Naïm, 1993)

a) Neurone biologique

Les cellules nerveuses, appelées neurones, sont les éléments de base du système nerveux central. Celui-ci en posséderait environ cent milliards. Les neurones possèdent de nombreux points communs dans leur organisation générale et leur système biochimique avec les autres cellules. Ils présentent cependant des caractéristiques qui leur sont propres et se retrouvent au niveau des cinq fonctions spécialisées qu'ils assurent :

- Recevoir des signaux en provenance de neurones voisins ;
- Intégrer ces signaux;
- Engendrer un influx nerveux ;
- Le conduire ;
- Le transmettre à un autre neurone capable de recevoir.

b) Structure du neurone

neurone est constitué de trois parties :

- Le corps cellulaire (soma ou péricaryon) ;
- Les dendrites ;
- L'axone.

❖ ***Le corps cellulaire***

Il contient le noyau du neurone et effectue les transformations biochimiques nécessaires à la synthèse des enzymes et des autres molécules qui assurent la vie du neurone.

Sa forme est pyramidales ou sphérique dans la plus part des cas. Elle dépond souvent de sa position dans le cerveau. Ce corps cellulaire fait quelques microns de diamètre.

- ***Les dendrites***

Chaque neurone possède une " chevelure" de dendrites. Celles-ci sont de fines extensions tubulaires, de quelques dixièmes de microns de diamètre et d'une longueur de quelques dizaines de microns. Elles se ramifient, ce qui les amène à former une espèce d'arborescence autour du corps cellulaire. Elles sont les récepteurs principaux du neurone pour capter les signaux qui lui parviennent.

- ***L'axone***

L'axone, qui est proprement parler la fibre nerveuse, sert de moyen de transport pour les signaux émis par le neurone. Il se distingue des dendrites par sa forme et par les propriétés de sa membrane externe. En effet, il est généralement plus long (sa longueur varie d'un millimètre à plus d'un mètre) que les dendrites, et se ramifie à son extrémité, là où il communique avec d'autres neurones, alors que les ramifications des dendrites se produisent plutôt près du corps cellulaire.

Pour former le système nerveux, les neurones sont connectés les uns aux autres suivant des répartitions spatiales complexes. Les connexions entre neurones se font dans des endroits appelés synapses où ils sont séparés par un petit espace synaptique de l'ordre d'un centième de microns.

c) Fonctionnement du neurone

Les fonctions spécifiques réalisées par un neurone dépendent essentiellement des propriétés de sa membrane externe.

- ***La membrane externe***

La membrane externe d'un neurone remplit cinq fonctions principales :

- Elle sert à propager des impulsions électriques tout au long de l'axone et des dendrites;
- Elle libère des médiateurs à l'extrémité de l'axone;
- Elle réagit à ces médiateurs au niveau des dendrites;

- Elle réagit au niveau du corps cellulaire aux impulsions électriques que lui transmettent les dendrites pour générer ou non une nouvelle impulsion;
- Enfin, elle permet au neurone de reconnaître les autres neurones afin qu'il puisse se situer au cours de la formation du cerveau et trouver les cellules auxquelles il doit être connecté.

❖ ***Le corps cellulaire comme sommateur à seuil***

D'une façon simple, on peut dire que le soma du neurone traite les courants électriques qui lui proviennent de ses dendrites, et qu'il transmet le courant électrique résultant de ce traitement aux neurones auxquels il est connecté par l'intermédiaire de son axone.

Le schéma classique présenté par les biologistes est celui d'un soma effectuant une sommation des influx nerveux transmis par ses dendrites.

Si la sommation dépasse un seuil, le neurone répond par un influx nerveux ou potentiel d'action qui se propage le long de son axone. Si la sommation est inférieure à ce seuil, le neurone reste inactif.

L'influx nerveux qui se propage entre différents neurones est, au niveau de ces neurones, un phénomène électrique.

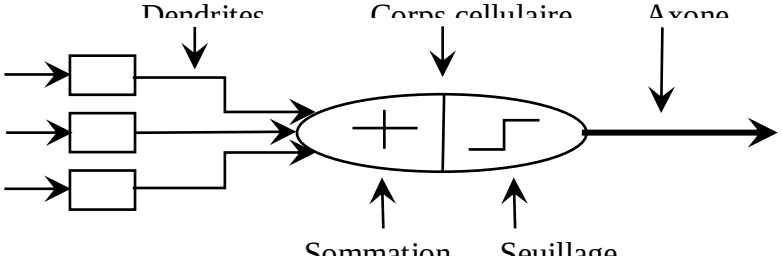


Figure IV.2: Schéma classique du neurone présenté par les biologistes

IV.2.3. Réseaux de neurones artificiels

Les réseaux de neurones biologiques réalisent facilement un certain nombre d'applications telles que la reconnaissance de formes, le traitement du signal, l'apprentissage par l'exemple, la mémorisation, la généralisation. Ces applications sont pourtant, malgré tous les efforts déployés en algorithmique et en intelligence artificielle, à la limite des possibilité

actuelles. C'est à partir de l'hypothèse que le comportement intelligent émerge de la structure et du comportement des éléments de base du cerveau que les réseaux de neurones artificiels se

sont développés. Les réseaux de neurones artificiels sont des modèles, à ce titre ils peuvent être décrits par leurs composants, leurs variables descriptives et les interactions des composants.

IV.2.3.1.Le neurone formel

L'étude biologique du système nerveux, a permis le passage des observations neurophysiologiques au neurone formel (Lippmann, 1987), ce dernier a été proposé par McCulloch et Pitts en 1943 qui est analogue au neurone biologique fondé sur une structure complexe (tableau IV.3), le modèle mathématique de neurone proposé a été repris par Rosenblatt pour définir le premier réseau de neurones artificiels, le Perceptron (Rosenblatt, 1962) .

Tableau IV.3:Analogie entre le neurone biologique et le neurone formel

Neurone artificiel	Neurone biologique
Poids de connexion	Synapses
Signal de sortie	Axones
Signal d'entrée	Dendrites
Fonction d'activation	Soma

Le neurone formel est un modèle mathématique simplifié du neurone biologique, il fait une somme pondérée des potentiels d'action qui lui parviennent, (chacun de ces potentiels est une valeur numérique qui représente l'état du neurone qui l'a émis), puis s'active suivant la valeur de cette sommation pondérée. Si cette somme dépasse un certain seuil, le neurone est activé et transmet une réponse (sous forme de potentiel d'action) dont la valeur est celle de son activation, si le neurone n'est pas activé, il ne transmet rien (Davalo et Naïm, 1993).

Mathématiquement, on peut définir un neurone formel comme étant une fonction non linéaire, paramétrée, à valeurs bornées (Dreyfus, 2004).

Les variables sur lesquelles opère le neurone sont habituellement désignées sous le terme entrées du neurone, et la valeur de la fonction sous celui de la sortie ; il est commode de représenter graphiquement un neurone comme indiqué sur la figure 1.Cette représentation est

le reflet de l'inspiration biologique qui a été à l'origine de la première vague d'intérêt pour les neurones formels, dans les années 1940 à 1970.

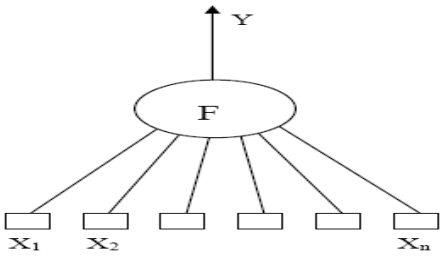
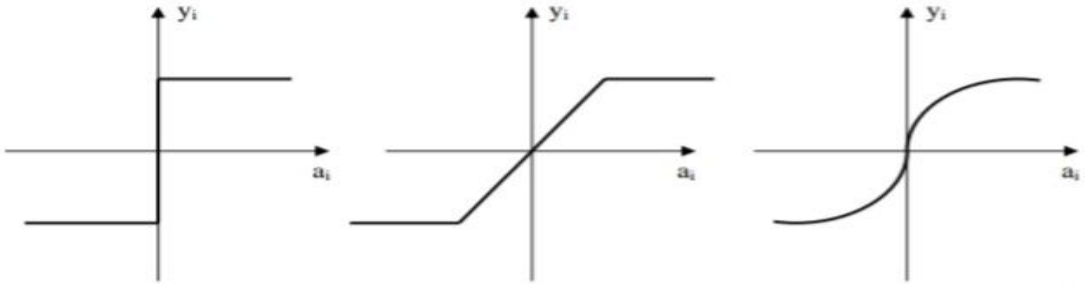


Figure IV.3:Un neurone réalise une fonction non linéaire bornée

$y = f(x_1, \dots, x_n; w_1, \dots, w_p)$où les $x_i$sont les v variables et $w_j$sont les paramètres.

Le neurone réalise alors trois opérations sur ses entrées :

- ✓ Pondération : multiplication de chaque entrée par un paramètre appelé poids de connexion,
- ✓ Sommation : une sommation des entrées pondérées est effectuée
- ✓ Activation : passage de cette somme dans une fonction, appelée fonction d'activation. La valeur calculée est la sortie du neurone qui est transmise aux neurones suivants.



Fonction à seuil Fonction linéaire par morceaux Fonction de type sigmoïde

Figure IV.4:Différents types de fonction de transfert pour le neurone

Artificiel (Ammar, 2007).

La fonction f est appelée fonction d'activation (figure IV.4). Elle peut être une fonction à seuil, une fonction linéaire ou non linéaire. La fonction sigmoïde se présente comme une approximation continûment dérivable de la fonction d'activation linéaire par morceaux ou de

la fonction seuil. Elle présente l'avantage d'être régulière, monotone, continûment dérivable, et bornée entre 0 et 1 (Ammar, 2007).

La fonction f peut être paramétrée de manière quelconque. Deux types de paramétrages sont fréquemment utilisés

a) Les paramètres sont attachés aux entrées du neurone : La sortie du neurone est une fonction non linéaire d'une combinaison des entrées $\{x_i\}$ pondérées par les paramètres $\{w_i\}$, qui sont alors souvent désignés par le nom de « poids » ou, en raison de l'inspiration biologique des réseaux de neurones,« poids synaptiques ». Conformément à l'usage (également inspiré par la biologie), cette combinaison linéaire est appelée « potentiel ». Le potentiel ν le plus fréquemment utilisé est la somme pondérée, à laquelle s'ajoute un terme constant ou « biais» :

$$\nu = \sum_{i=1}^{n} w_i x_i + w_0 \qquad \text{IV.11}$$

La sortie d'un neurone a pour équation :

$$y = f\left[\sum_{i=1}^{n} w_i x_i + w_0\right] \qquad \text{IV.12}$$

La fonction *f* est appelée fonction d'activation (ou de transfert). Elle sert à calculer la valeur de l'état du neurone. Il existe de nombreuses formes possibles pour la fonction de transfert.

La fonction non linéaire sigmoïde est fréquemment utilisée dans les ANN, particulièrement dans les réseaux utilisant l'algorithme de rétropropagation (Vander Baan et Jutten, 2000) car contrairement à la fonction sigmoïde, les autres fonctions donnent seulement une sortie binaire ce qui rend plus difficile à estimer les poids optimaux.

Le biais w_0 joue un rôle de seuil, quand le résultat de la somme pondérée dépasse ce seuil, l'argument de la fonction de transfert devient positif ou nul; dans le cas contraire, il est considéré négatif. Finalement si le résultat de la somme pondérée est:

1. En dessous du seuil, le neurone est considéré comme non-actif ;
2. Aux alentours du seuil, le neurone est considéré en phase de transition ;
3. Au-dessus du seuil, le neurone est considéré comme actif.

Remarque : Le neurone crée par McCulloch et Pitts était un automate booléen, c'est-à-dire que ses entrées et sa sortie étaient booléennes.

b) Les paramètres sont attachés à la non-linéarité de neurone : ils interviennent directement dans la fonction *f* ; cette dernière peut être une fonction radiale ou RBF (en anglais Radial Basis Function).

Par exemple, la sortie d'un neurone RBF à non-linéarité gaussienne a pour équation :

$$y = \exp\left[-\frac{\sum_{i=1}^{n}(x_i - w_i)^2}{2w_{n+1}^2}\right] \qquad \text{IV.13}$$

Où les paramètresw_i, i=1 à n sont les coordonnées du centre de la gaussienne, et w_{n+1}^2 est son écart type.tanh(x) (Krajnc, 2021).

IV.2.3.2.Architecture des réseaux de neurones

L'architecture est un concept très important qui joue un rôle déterminant dans la classification des ANN. Dans la littérature on utilise souvent le mot structure comme synonyme d'architecture (Maren et al. 1990; Hertz et al. 1991). Chaque architecture a sa propre organisation qui est adaptée à des applications bien spécifiques (Sarle1994;Haykin 1994).On distingue deux structures de réseau,

- Les réseaux «Feedforward» (ou non bouclés, ou statiques, ou acycliques).
- Les réseaux «Feedback» (ou Récurrents ou bouclés, ou dynamiques, ou cycliques).

IV.2.3.2.1.Les réseaux de neurones feed-forwarded

Dans ce type de réseaux (figure IV.5), l'information se propage dans un sens unique, sans aucune rétroaction (des entrées vers les sorties). Si l'on représente le réseau comme un graphe dont les nœuds sont les neurones et les arêtes les « connexions » entre ceux-ci, le graphe est acyclique. Ce genre de réseaux utilise un apprentissage supervisé, par correction des erreurs (Lippmann,1987) ou le signal d'erreur est rétropropagé vers les entrées afin de mettre à jour les poids des neurones.

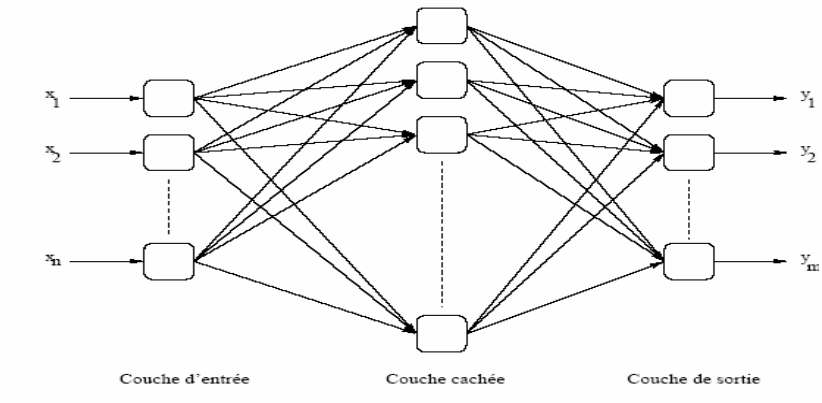


Figure IV.5 :Réseau Feedforward à trois couches

Le terme de « connexions » doit être pris dans un sens métaphorique, car le réseau de neurones n'est pas donc, en général on objet physique tel un circuit électrique, et les «connexions » n'ont pas de réalité matérielle ; néanmoins le terme de connexion, issu des origines biologiques des réseaux de neurones, et passé dans l'usage, car il est commode quoique trompeur. Il a même donné naissance au terme de connexionnisme (Dreyfus, 2004).
On distingue dans cette catégorie :

- ***Le perceptron multicouche MLP***

Les perceptrons multicouches sont une amélioration du perceptron. Ils sont, en effet, les plus employés et les plus étudiés. Une abréviation anglaise est utilisée dans la littérature pour les nommer : MLP pour Multi Layer Perceptrons(Bishop, 1995; Haykin, 1994).

Ils disposent d'une ou plusieurs couches cachées. Les neurones y sont arrangés en couches successives : la première couche qui forme le vecteur des données d'entrée est appelée couche d'entrée tandis que la dernière couche qui produit les résultats est appelée couche de sortie. Toutes les autres couches qui se trouve au milieu sont appelées couches cachées (Lippmann 1987, Hagan et al. 1996).

Chaque couche est composée d'un certain nombre de neurones. Les connexions sont établies entre les neurones appartenant à des couches successives mais les neurones d'une même couche ne peuvent pas communiquer entre eux.

Contrairement au perceptron monocouche la présence d'une couche cachée dans le perceptron multicouche facilite la modélisation des relations non linéaires entre les entrées et la sortie.

Le choix du nombre de couches cachées dépend généralement de la complexité du problème à résoudre, en théorie une seule couche cachée peut être suffisante pour résoudre un problème donné mais il se peut que le fait de disposer de plusieurs couches cachées permette de résoudre plus facilement un problème complexe.

Construire un réseau de neurone à couche (perceptron multicouche) pour un problème quelconque revient à faire un choix judicieux de la taille du réseau, du nombre total de couches et de neurones, distribution des données et des fonctions de transfert (Baum et Haussier 1989). Le choix de ces paramètres dépend de l'utilisateur. Il n'existe pas dans la littérature pour le moment des données suffisantes qui peuvent déterminer clairement les paramètres à adopter pour résoudre un problème donné (Coulibaly et al 1998).

- ***Réseaux à RBF (fonction radiale de base)***

Les réseaux RBF sont très semblables aux perceptrons multicouches. Ils sont utilisés dans les mêmes genres de problèmes que les perceptrons multicouches à savoir, en classification et en prédiction, à travers une combinaison linéaire de fonctions non linéaires à

base radiale. Parmi ces fonctions, la fonction gaussienne, (présentée plus haut), qui est la plus utilisée.

IV.2.3.2.2.Réseaux récurrents

Ce genre de réseaux est caractérisé par le pouvoir de laisser 1'information circulé récursivement d'une manière partielle ou bien total (Kasabov, 1996;Elman,1990).

L'architecture la plus générale pour un réseau de neurones est le « réseau récurrent», dont le graphe des connexions est cyclique : lorsqu'on se déplace dans le réseau en suivant le sens des connexions, il est possible de trouver au moins un chemin qui revient à son point de départ (un tel chemin est désigné sous le terme de « cycle »). La sortie d'un neurone du réseau peut donc être fonction d'elle-même; cela n'est évidemment concevable que si la notion de temps est explicitement prise en considération.

Ainsi, à chaque connexion d'un réseau de neurones bouclé (ou à chaque arête de son graphe) est attaché, outre un poids comme pour les réseaux non bouclés, un retard, multiple entier (éventuellement nul) de l'unité de temps choisie. Une grandeur, à un instant donné, ne pouvant pas être fonction de sa propre valeur au même instant, tout cycle du graphe du réseau doit avoir un retard non nul (Dreyfus, 2004).

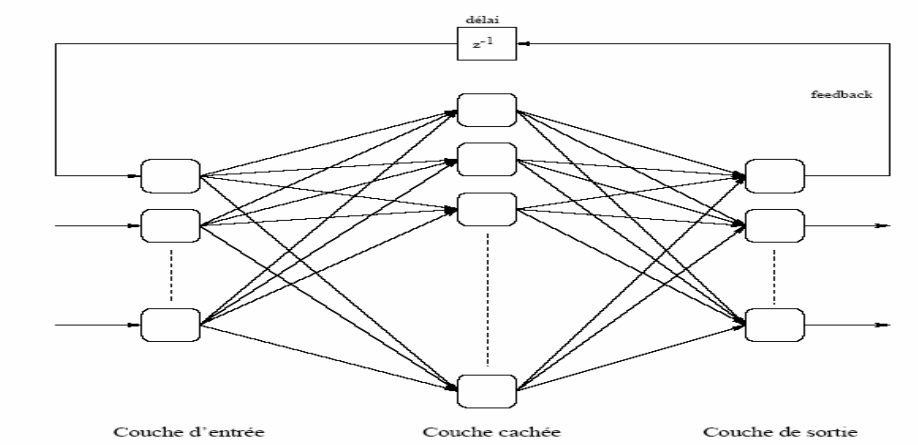


Figure IV.6: réseau feedback simple

IV.2.3.3.Apprentissage des réseaux de neurones

L'apprentissage est une étape très importante du développement d'un réseau de neurones durant laquelle le comportement du réseau est modifié itérativement jusqu'à l'obtention du comportement désiré, et ce par l'ajustement des poids (connexion ou synapse) des neurones à une source d'information bien définit (Hebb 1949;Grossberg 1982; Rumelhart et al. 1986).
Dans la majorité des algorithmes actuels, les variables modifiées pendant l'apprentissage sont les poids des connexions. L'apprentissage est la modification des poids du réseau dans l'optique d'accorder la réponse du réseau aux exemples et à l'expérience. Les poids sont initialisés avec des valeurs aléatoires. Puis des exemples expérimentaux représentatifs du fonctionnement du procédé dans un domaine donné, sont présentés au réseau de neurones. Ces exemples sont constitués de couples expérimentaux de vecteurs d'entrée et de sortie.

Une méthode d'optimisation modifie les poids au fur et à mesure des itérations pendant lesquelles on présente la totalité des exemples, afin de minimiser l'écart entre les sorties calculées et les sorties expérimentales.

On distingue trois types d'apprentissages :

a) L'apprentissage supervisé ;
b) L'apprentissage non supervisé ;
c) L'apprentissage par renforcé.

a) L'apprentissage supervisé

Dans ce type d'apprentissage, on cherche à imposer au réseau un fonctionnement donné en forçant les sorties du réseau à prendre des valeurs bien spécifiques désirées (choisies par l'opérateur) et ce en modifiant les poids synaptiques.

Le réseau se comporte alors comme un filtre dont les paramètres de transfert sont ajustés à partir des couples entrée-sortie présentés (Hassoum, 1995).

L'adaptation des paramètres du réseau s'effectue à partir d'un algorithme d'optimisation, l'initiation des poids synaptiques étant le plus souvent aléatoire.

L'apprentissage supervisé est mise en œuvre essentiellement pour une modélisation statique ou une régression.

b) L'apprentissage non supervisé

L'apprentissage non supervisé consiste à ajuster les poids à partir d'un ensemble d'apprentissage formé uniquement d'entrées. Aucun résultat désiré n'est fourni au réseau.
Qu'est-ce que le réseau apprend exactement dans ce cas ? L'apprentissage consiste à détecter les similarités et les différences dans l'ensemble d'apprentissage. Les poids et les sorties du réseau convergent, en théorie, vers les représentations qui capturent les régularités statistiques des données (Fukushima, 1988).Ce type d'apprentissage est également dit compétitif et (ou) coopératif (Grossberg, 1988). L'avantage de ce type d'apprentissage réside dans sa grande capacité d'adaptation reconnue comme une auto-organisation, « self-organizing» (Kohonen, 1988). L'apprentissage non supervisé est surtout utilisé pour le traitement du signal et l'analyse factorielle.

c) L'apprentissage renforcé

L'apprentissage renforcé est une technique similaire à l'apprentissage supervisé à la différence qu'au lieu de fournir des résultats désirés au réseau (coulibaly et al, 1999), on lui

accorde plutôt un grade (ou score) qui est une mesure du degré de performance du réseau après quelques itérations.

Les algorithmes utilisant la procédure d'apprentissage renforcé sont surtout utilisés dans le domaine des systèmes de contrôle (White et Sofge, 1992; Sutton, 1992).

IV.2.3.3.1.Algorithme d'apprentissage

L'algorithme d'apprentissage est la méthode mathématique qui va modifier les poids de connexions afin de converger vers une solution qui permettra au réseau d'accomplir la tâche désirée. L'apprentissage est une méthode d'identification paramétrique qui permet d'optimiser les valeurs des poids du réseau.

- ***Fonction du coût***

L'apprentissage consiste minimiser une fonction de coût, représentative des différences entre les sorties effectives du réseau de neurones et les sorties désirées ou cibles.

La fonction la plus couramment utilisée est la fonction des moindres carrés :

$$J(w) = \frac{1}{2}\sum_{k=1}^{N}\left[y_p(x^k) - g(x^k, w)\right]^2 \qquad \text{IV.14}$$

Où x^k désigne le vecteur des valeurs des variables pour l'exemple k, $y_p(x^k)$ la valeur de mesure correspondante, w désigne le vecteur des poids du réseau de neurones, et $g(x^k, w)$est la valeur calculée par le réseau de neurone muni des poids w pour le vecteur x^k de variables. La fonction du coût est donc une fonction de tous les paramètres ajustables w de tous les neurones et toutes connexions du réseau. L'apprentissage consiste donc à trouver des paramètres w qui qui rendent $J(w)$ « minimum ». L'apprentissage est donc un problème numérique d'optimisation.

- ✓ ***Optimisation***

Comme, la sortie du réseau n'est pas linéaire par rapport aux paramètres, on doit résoudre un problème d'optimisation non linéaire multi-variable. Il n'est donc pas possible d'utiliser la méthode des moindres carrés ordinaire pour minimiser cette fonction. Les méthodes utilisées à cet effet sont des techniques itératives, qui à partir d'un réseau muni de

poids dont les valeurs sont aléatoires, modifient ces paramètres jusqu'à ce qu'un minimum de la fonction du coût soit atteint, ou qu'un critère d'arrêt soit satisfait.

Ces techniques sont toutes des méthodes de gradient : elles sont fondées sur le calcul, à chaque itération, du gradient de la fonction du coût par rapport aux paramètres, gradient qui est ensuite utilisé pour calculer une modification de ceux-ci. Le calcul du gradient peut être effectué de diverses manières ; il en est une, appelée la rétropropagation, qui est généralement plus économes que les autres en termes de nombres d'opération arithmétiques à effectuer pour évaluer le gradient (Gérard Dreyfus, 2002). .

La procédure suivie lors de l'apprentissage est alors la suivante :

1- Initialisation des paramètresw_i;
2- Calcul du gradient de fonction du coût $\nabla J(w)$ par l'algorithme de rétropropagation ;
3- Modification des paramètres par une méthode de minimisation (la méthode du gradient simple ou par une des méthodes de gradient de second ordre) ;
4- Reprise de la procédure à l'étape 2 jusqu'au minimum de la fonction du coût.

❖ *Evaluation du gradient de la fonction du coût par la rétro propagation*

Considérons un réseau de neurones non bouclé avec neurones cachés et un neurone de sortie. L'extension à un réseau qui possède plusieurs neurones de sortie est triviale.

Le neurone i calcule une grandeur y_i qui est une fonction non linéaire de son potentiel v_i ; potentiel $v_i$est une somme pondérée des entréesx_j, la valeur de l'entrée x_j étant pondérée par un paramètre w_{ij} :

$$y_i = f\left(\sum_{j=1}^{n_i} w_{ij} x_j\right) = f(v_i) \qquad \text{IV.15}$$

Les entrées n_i du neurone i peuvent être soit des sorties d'autres neurones, soit les entrées du réseau. Dans toutes la suite, x_j désignera donc indifféremment soit la sortie $y_j$du neurone j, soit l'entrée j du réseau.

La fonction du coût dont on cherche à évaluer le gradient est de la forme.

$$J(w) = \frac{1}{2}\sum_{k=1}^{N}\left[y_p^k - g(x^k, w)\right]^2 = \sum_{k=1}^{N} J^k(w) \qquad \text{IV.16}$$

Pour évaluer son gradient, il suffit donc d'évaluer le gradient du coût partiel$J^k(w)$, relatif à l l'observation k et faire ensuite la somme sur tous les exemples.

L'algorithme de rétro propagation consiste essentiellement à l'application répétée de la règle des dérivées composées. On remarque tout d'abord que la fonction du coût partielle ne dépend pas du paramètre W_{ij} que par l'intermédiaire de la valeur de la sortie du neurone i, qui est elle-même fonction uniquement du potentiel du neurone i ; on peut donc écrire :

$$(\frac{\partial \mathrm{J^k}}{\partial \mathrm{w_{ij}}})_\mathrm{k} = \left(\frac{\partial \mathrm{J^k}}{\partial \mathrm{v_i}}\right)_\mathrm{k} (\frac{\partial \mathrm{v_i}}{\partial \mathrm{w_{ij}}})_\mathrm{k} = \delta_\mathrm{i}^\mathrm{k} \mathrm{x_j^k} \qquad \text{IV.17}$$

Où :

- $\left(\frac{\partial \mathrm{J^k}}{\partial \mathrm{v_i}}\right)_\mathrm{k}$ Désigne la valeur du gradient du coût partiel par rapport au potentiel du neurone i lorsque les entrées du réseau sont celles qui correspondent à l'exemple k ,
- $(\frac{\partial \mathrm{v_i}}{\partial \mathrm{w_{ij}}})_\mathrm{k}$Désigne la dérivée partielle au potentiel du neurone i par rapport au paramètre $\mathrm{w_{ij}}$ lorsque les entrées du réseau sont celles qui correspondent à l'exemple k ,
- $\mathrm{x_j^k}$est la valeur de l'entrée j du neurone i lorsque les entrées du réseau sont celles qui correspondent à l'exemple k.

Il reste donc à évaluer les quantitésδ_i^k . Nous allons voir que ces quantités peuvent avantageusement calculées d'une manière récursive en menant les calculs depuis la ou la (ou les) sortie(s) du réseau vers ses entrées.

- En effet pour le neurone i :

$$\delta_\mathrm{i}^\mathrm{k} = \left(\frac{\partial \mathrm{J^k}}{\partial \mathrm{v_i}}\right)_\mathrm{k} = (\frac{\partial}{\partial \mathrm{v_i}}[(\mathrm{y_p^k} - \mathrm{g(x^k, w)})^2])_\mathrm{k} = -2\mathrm{e}(\mathrm{x^k, w})\left(\frac{\partial \mathrm{g(x, w)}}{\partial \mathrm{v_i}}\right)_\mathrm{k} \qquad \text{IV.18}$$

Où $\mathrm{e}(\mathrm{x^k, w}) = \mathrm{y_p^k} - \mathrm{g(x^k, w)}$ est l'erreur de modélisation commise par le réseau, muni du vecteur paramètres w, pour l'exemple $\mathrm{x^k}$.

Or, la sortie du modèle est la sortie $\mathrm{y_i}$du neurone de sortie ; cette relation s'écrit donc :

$$\delta_\mathrm{i}^\mathrm{k} = -2\mathrm{e}(\mathrm{x^k, w})\mathrm{f}'(\mathrm{v_i^k}) \qquad \text{IV.19}$$

Où $f'(v_i^k)$ désigne la dérivée de la fonction d'activation du neurone de sortie lorsque les entrées du réseau sont celles qui correspondent à l'exemple k. Si, comme c'est le cas lorsque le réseau est utilisé en modélisation, le neurone de sortie est linéaire, l'expression se réduit à :

$$\delta_i^k = -2e(x^k, w) \qquad \text{IV.20}$$

Pour un neurone caché i : la fonction du coût ne dépend du potentiel du neurone i que par l'intermédiaire des potentiels des neurones m qui reçoivent la valeur de la sortie du neurone i, c'est à dire, de tous les neurones qui, dans le graphe des connexions du réseau, sont adjacents au neurone i, entre ce neurone et la sortie. La relation s'écrit :

$$\delta_i^k = \sum_m \delta_m^k w_{mi} f'(v_i^k) = f'(v_i^k) \sum_m \delta_m^k w_{mi} \qquad \text{IV.21}$$

Ainsi, les quantités δ_i^k peuvent être calculées récursivement, en parcourant le graphe des connexions « à l'envers » depuis la (les) sortie (s) vers les entrées du réseau (ce qui explique le terme rétropropagation). Une fois les gradients des coûts partiels ont été calculés, il suffit d'en faire la somme pour obtenir le gradient de la fonction du coût totale.

L'algorithme de rétro propagation comporte deux phases pour chaque exemple k :

- Une phase de propagation, au cours de laquelle les entrées correspondant à l'exemple k sont utilisées pour calculer les sorties et les potentiels de tous les neurones.
- Une phase de rétro propagation, au cours de laquelle sont calculées les quantités δ_i^k.

Une fois que ces quantités sont calculées, on calcule les gradients des coûts partiels par la relation

$$\left(\frac{\partial J^k}{\partial w_{ij}}\right)_k = \delta_i^k x_j^k \qquad \text{IV.22}$$

Puis le gradient du coût total

$$\left(\frac{\partial J}{\partial w_{ij}}\right)_k = \sum_k \left(\frac{\partial J^k}{\partial w_{ij}}\right)_k \qquad \text{IV.23}$$

❖ *Modification des paramètres en fonction du gradient de la fonction du coût*

a) Méthodes du premier ordre

Méthodes du premier ordre (ou de gradient simple) consistent à modifier les paramètres par la formule suivante, à l'itération i de l'apprentissage :

$$w(i) = w(i-1) - \mu_i \nabla J\big(w(i-1)\big) \text{ avec } \mu_i > 0 \qquad IV.24$$

La direction de descente est donc simplement opposée à celle du gradient : c'est en effet la direction suivant laquelle la fonction du coût diminue le plus rapidement. La quantité μ_i est appelée pas de gradient ou pas d'apprentissage.

b) Méthodes du second ordre

Ces méthodes sont reconnues comme étant les plus rapides à converger, c'est pourquoi c'est ce type de méthode qui a été utilisé dans ce travail.

Toutes les méthodes de second ordre sont dérivées de la méthode Newton, dont nous présentons le principe ci-après.

Le développement de Taylor d'une fonction $J(w)$ d'une seule variable w au voisinage d'un minimum w^* est donné par la relation :

$$J(w) = j(w^*) + \frac{1}{2}(w - w^*)^2 \left(\frac{d^2J}{dw^2}\right)_{w=w^*} + O(w^3) \qquad \text{IV.25}$$

Car le gradient de la fonction du coût est nul au voisinage d'un minimum. Une approximation du gradient de la fonction du coût au voisinage d'un minimum est obtenue aisément en dérivant la relation précédente par rapport à w :

$$\frac{dJ}{dw} = (w - w^*)\left(\frac{d^2J}{dw^2}\right)_{w=w^*} \qquad \text{IV.26}$$

Par conséquent, lorsque la variable w est au voisinage dew^*, on pourrait atteindre ce minimum en une seule itération si l'on connaissait la dérivée seconde de la fonction à son minimum : il suffirait pour cela de modifier la variable w de la quantité :

$$\Delta w = \frac{(dJ/dw)}{(d^2J/dw^2)_{w=w^*}} \qquad \text{IV.27}$$

Le même raisonnement s'applique à une fonction de plusieurs variables, la dérivée seconde étant remplacé par la matrice hessienne *H(w)* de la fonction à optimiser, le terme générale $\frac{\partial^2 J}{\partial w_i \partial w_j}$:pour atteindre le minimum de la fonction du coût en une itération, il suffirait d'appliquer au vecteur des poids la modification suivante (sous réserve que la matrice hessienne soit inversible) :

$$\Delta w = -H(w^*)^{-1}\nabla J(w)I \qquad \text{IV.28}$$

Cette dernière formule n'est évidemment pas applicable, puisque le vecteur n'est pas w^* connu. Néanmoins, elle suggère plusieurs techniques qui mettent en œuvre une approximation itérative de la matrice hessienne. Dans ce travail nous avons choisi l'algorithme de Levenberg-Marquardt.

- ✓ ***Algorithme de Levenberg-Maquardt***

L'Algorithme de Levenberg-Maquardt (Levenberg et al, 1944) (Maquardt et al 1963) consiste à modifier les paramètres par la formule:

$$w(i) = w(i-1) - [H(w(i-1) + \mu_i I)]^{-1} \nabla J\big(w(i-1)\big) \qquad \text{IV. 29}$$

IV.2.3.4.Propriété fondamentale des réseaux de neurones non bouclés

IV.2.3.4.1.L'approximation universelle

Toute fonction bornée suffisamment régulière peut être approchée uniformément, avec une précision arbitraire, dans un domaine fini de l'espace de ses variables, par un réseau de neurones comportant une couche de neurones cachés en nombre fini, possédant tous la même fonction d'activation, et un neurone de sortie linéaire (Hornik et al., 1989 ; Hornik et al., 1990;Hornik, 1991).

IV2.3.5.Surapprentissage

L'objectif de l'apprentissage est de trouver un modèle qui a la complexité suffisante pour rendre compte de la relation déterministe entre les facteurs sélectionnés et la sortie du processus, mais qui n'est pas trop complexe, de sorte qu'il ne s'ajuste pas au bruit présent dans les données d'apprentissages. En d'autres termes, il faut trouver un modèle qui réalise le meilleur compromis entre les capacités d'apprentissage et les capacités de généralisation: Si le réseau apprend «trop bien», il s'ajuste au bruit, et donc a de mauvaises performances de généralisation (Dreyfus ,2002).

Dans ce travail pour éviter le surapprentissage nous avons utilisé la technique dite d'«arrêt prématuré» (early stopping).

IV.2.3.5.1.Arrêt prématuré : principe

Cette technique permet d'arrêter prématurément c'est à dire avant la convergence complète de l'algorithme. Ainsi, le modèle ne s'ajuste pas trop finement aux données d'apprentissage : le sur-ajustement est alors limité. La difficulté réside alors évidemment dans la détermination du moment où arrêter l'apprentissage. La méthode d'«arrêt prématuré» (early stopping) (Bishop, 1995) consiste à suivre l'évolution de la fonction du coût sur une base de validation, et arrêter les itérations lorsque le coût calculé sur cette base commence à croitre.

Pour ce faire, La variation de de la fonction du coût MSE en fonction des itérations pour les ensembles d'apprentissage et de validation durant l'apprentissage est montré dans la figure, durant les premières itérations la MSE pour les données de l'apprentissage et de la validation diminue. Apres un certain nombre d'itération, la MSE des données d'apprentissage continue de diminuer mais la MSE des données utilisées pour la validation commence à augmenter (figure..). Cela est indicatif d'un surapprentissage. Quand cela se produit, le processus d'apprentissage est stoppé (i.e. La méthode d'«arrêt prématuré» a été appliquée), et les poids et les biais de cette itération sont considérées comme valeurs optimales.

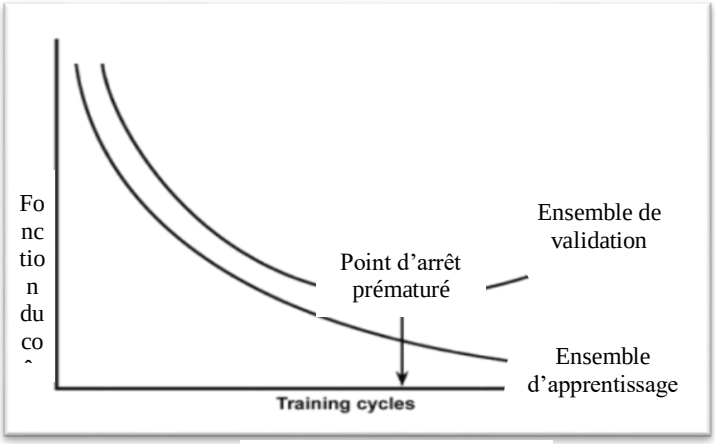


Figure IV.7: L'Arrêt prématuré (early stopping)

IV.2.3.6.Comment mettre en œuvre un réseau de neurone ?

Les réseaux de neurones réalisent des fonctions non linéaires paramétrées. La mise en œuvre d'un réseau de neurones nécessite donc :

- De déterminer les entrées pertinentes, c'est à dire les grandeurs quiont une influence significative sur le phénomène que l'on cherche à modéliser;
- La collecte des données nécessaires à l'apprentissage et à l'évaluation des performances du réseau de neurones;
- La détermination du nombre de neurones cachés nécessaires pour obtenir une approximation satisfaisante;
- La réalisation de l'apprentissage;
- L'évaluation des performances du réseau de neurones à l'issue de l'apprentissage ;
- Utilisation du réseau pour la prévision.

IV.3.Apprentissageprofond (Deep learning)

IV.3.1.Introduction sur Deep learning

Depuis 2006, le domaine a fait de grandes avancées avec l'apprentissage de réseaux profonds (Deep learning). Cette approche permet, à partir de données d'entrée, d'en extraire une représentation plus riche. Les réseaux profonds sont composes de plusieurs couches de neurones. Chaque couche est une étape qui représente les données de façonun peu plus complexe (abstraite) en se basant sur ce qui a été appris dans la couche précédente. Cette approche est à l'image de l'apprentissage humain qui commence par apprendre des concepts simples, comme l'addition et la soustraction en mathématiques, pour ensuite se baser sur ces concepts afin d'en apprendre des plus complexes, comme la multiplication et le principe de fonction (Alouache et Chia, 2019).

IV.3.2.Définition

Le Deep Learning est basé sur l'idée des réseaux de neurones artificiels et il est taillé pour gérer de larges quantités de données en ajoutant des couches au réseau.

Un modèle de Deep learning a la capacité d'extraire des caractéristiques à partir des données brutes grâce aux multiples couches de traitement composé de multiples transformations linéaires et non linéaires et apprendre sur ces caractéristiques petit à petit à travers chaque couche avec une intervention humaine minime (Foued, 2019).

Le Deep Learning est un nouveau domaine de recherche du ML, qui a été introduit dans le but de rapprocher le ML de son objectif principal : l'intelligence artificielle. Il

concerne les algorithmes inspirés par la structure et le fonctionnement du cerveau. Ils peuvent apprendre plusieurs niveaux de représentation dans le but de modéliser des relations complexes entre les données (Foued, 2019) .

IV.3.3.Histoire du Deep learning

Tableau IV.4: Les étapes majeures du Deep Learning

Année	Contributeur	Contribution
2006	Geoffrey Hinton	Introduction des deepBelief network
2009	Salakhutdinov and Hinton	Introduction des deep Boltzmann machines
2012	Alex Krizhevsky	Introduction de AlexNet qui remporta le challengeImageNet

IV.3.4.Domaines d'application de Deep learning

L'apprentissage en profondeur investit progressivement notre quotidien :

- La reconnaissance vocale ;
- Le tagging automatique de morceaux de musique ;
- La synthèse vocale avancée L'étiquetage automatique d'image ;
- La conception de nouvelles molécules pharmaceutiques ;
- La régression ;
- La prévision ;
- La classification .

Toutes ces applications mettent aujourd'hui en œuvre des techniques de Deep Learning .

IV.3.5.Apprentissage profond (Deeplearning)

IV.3.5.1.Apprentissage automatique

Chaque algorithme qui est capable d'apprendre de données est un algorithme d'apprentissage automatique. Ce dernier est dit capable d'apprendre de données si sa

performance aux tâches dans T, mesurée par la performance P, s'améliore avec l'expérience E (Labiad, 2017).

a) ***La tâche, T***

De nombreux types de tâches peuvent être résolus avec l'apprentissage automatique à titre d'exemple : la classification, la régression, et la traduction, etc. Dans la classification, l'algorithme spécifie à laquelle des catégories *k* certaines entrées appartiennent (la reconnaissance d'objet est un exemple de classification, où l'entrée est une image et la sortie est un code numérique identifiant l'objet dans l'image). La régression, prédit une valeur numérique étant donnée une entrée (la prévision du montant réclamé par une personne assurée). La traduction est une tâche qui consiste à convertir une séquence de symboles écrite dans une certaine langue à une autre langue(Labiad, 2017).

b) ***La mesure de performance, P***

Nous devons concevoir une mesure quantitative de performance pour évaluer les capacités d'un algorithme d'apprentissage automatique. Par exemple, pour la tâche classification, nous mesurons souvent la précision du modèle (une proportion d'exemples pour lesquels le modèle produit la sortie correcte). Il est souvent difficile de choisir une mesure de performance qui corresponde bien au système (Labiad, 2017)

classification, nous mesurons souvent la précision du modèle (une proportion d'exemples pour lesquels le modèle produit la sortie correcte). Il est souvent difficile de choisir une mesure de performance qui corresponde bien au système (Labiad, 2017).

c) ***L'expérience, E***

Les algorithmes d'apprentissage automatique peuvent être classés en deux catégories : non supervisés ou supervisés. Les algorithmes d'apprentissage supervisés expérimentent une base de données d'apprentissage contenant des exemples prélassés.Tandis que, les algorithmes d'apprentissage supervisés expérimentent un ensemble de données pour apprennent la structure de ces données sans utilisation d'une base de données d'apprentissage(Labiad, 2017).

IV.3.5.2.La catégorisation de l'apprentissage profond

Selon la façon dont les architectures et les techniques sont destinées à être utilisées, on peut classer globalement l'apprentissage profond en trois grandes catégorie :

- ***Les réseaux profonds pour l'apprentissage non supervisé***

Ils sont destinés à capturer une relation élevée des données observées pour l'analyse de motifs ou la synthèse quand aucune information sur les étiquettes de sorties n'est disponible(Labiad, 2017).

- ***Les réseaux profonds pour l'apprentissage supervisé***

Ils sont destinés pour fournir directement une puissance discriminative pour la classification des motifs, souvent en caractérisant les distributions postérieures des classes conditionnées sur les données visibles(Labiad, 2017).

- ***Les réseaux profonds hybrides***

Dans cette catégorie, l'objectif est la discrimination qui est assistée par les résultats des réseaux profonds non supervisés à titre d'exemple les autoencoders. Autrement dit, les autoencoders sont utilisés pour l'apprentissage de DN (Labiad, 2017).

IV.3.6.Avantages des réseaux profonds

Les réseaux profonds représentent des avantages tels que :

- Les réseaux profonds sont capables d'apprendre des fonctions complexes.
- Ils possèdent de bonnes capacités de généralisation (Labiad, 2017).

IV.3.7.Inconvénients des réseaux profonds

Les réseaux profonds ont aussi des inconvénients tels que :

- Les réseaux profonds nécessitent une grande quantité de données.
- Ils sont extrêmement coûteux en apprentissage (Labiad, 2017).

Conclusion

Dans ce chapitre, nous avons présenté un ensemble de méthodes basées sur l'intelligence artificielle qui seront appliquées dans le prochain chapitre.

Chapitre V Application sur des séries pluviométriques du bassin versant d'oued Soummam

Chapitre V : Application sur des séries pluviométriques du bassin versant d'oued *Soummam*

Introduction

Le présent chapitre est composé de deux parties principales. La première partie est consacrée à la présentation du bassin versant de la Soummam, qui donne un aperçu général sur les caractéristiques du bassin et mets en valeurs les différentes données qui seront utilisées dans la deuxième partie du chapitre.

Dans la deuxième partie, une étude comparative des différentes méthodes d'estimation de données manquantes a été menée .L'application a été faite sur le pas de temps mensuel. Deux classes de méthodes sont appliquées :

- Méthodes classiques : CCWM, IDWM
- Méthodes basées sur l'intelligence artificielle basée sur les algorithmes génétiques FFSGAM (FFSGAM[1], FFSGAM[2], FFSGAM[3], FFSGAM[4]), et réseaux de neurones artificiels (apprentissage en profondeur)

Afin de pouvoir juger de ces méthodes, une étude comparative est menée entre les résultats fournis par chacune de ces méthodes.

V.1.Méthodologie

L'application des méthodes d'estimations a été faite dans le bassin versant d'oued Soummam sur 5 stations choisies sur la base de la disponibilité des données pluviométriques et fournies par l'agence national des Ressources hydraulique (A.N.R.H), ces stations sont situées dans le bassin versant du d'oued Soummam , qui porte le code (15) .

V.1.1.Présentation de la région d'étude

V.1.1.1.Situation géographique

D'une superficie de 9125Km², le bassin versant de la Soummam est situé dans la partie Nord-est de l'Algérie à mi-chemin entre la ville d'Alger et la ville de Constantine, exactement entre les méridiens 3° 38' et 5° 38' et les parallèles 35° et 36° 45'. Il est limité au

Nord par les chaînes montagneuses du Djurdjura et ses contreforts qui s'étendent jusqu'à la mer Méditerranée, au Sud par les contreforts des monts du Hodna, à l'Est par les chaînes des Babors et le plateau de Sétif et à l'Ouest par le plateau de Bouira (Tebbani, 2016).

Le bassin versant de la Soummam porte selon la codification de l'ANRH le **N°15**. Il est limitrophe de plusieurs bassins versants (Figure V.1). Limité au Nord, par le bassin de l'Oued Sebaou(code 02b) et par celui du côtier Algérois (code 02a), au Nord-est par le bassin versant de l'Oued Kébir Rhumel (code 10) et par le bassin versant du côtier Constantinois (code 03), à l'Est par les basses hauts plateaux Constantinois (code 07) ; au Sud par le bassin de Chott El Hodna (code 05), et à l'Ouest par le bassin de l'Oued Isser (code 09)(Tebbani, 2016).

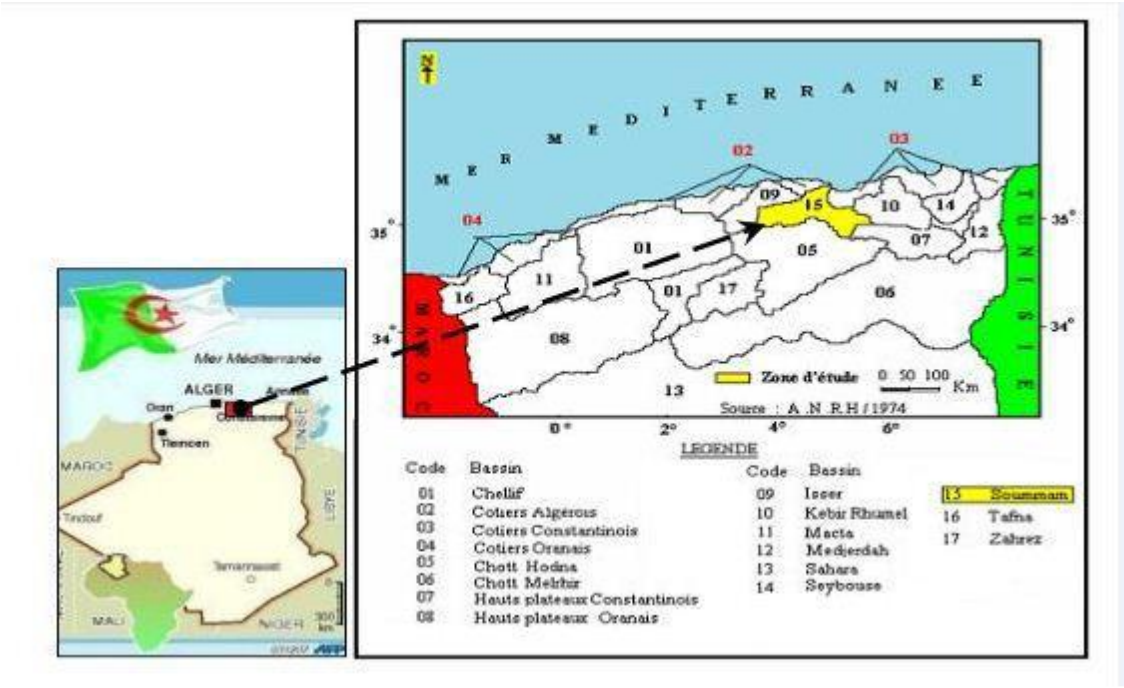


Figure V.1 : La situation de la zone d'étude par rapport au bassin hydrographique De l'Algérie du nord (Ramdini, 2016)

Le bassin versant de l'Oued Soummam est dans ses grandes lignes constitué, sur la rive Gauche , par de l'oligocènetra vers épar des formations du crétacé inférieur ;du miocène inférieur apparaît dans la partie aval ,en bordure de l'Oued de terrasses alluviales importantes tapissent en générale pied des pentes sauf dans la région de Sidi-Aich où le crétacé apparaît jusque dans le lit. Le versant rive droite est en majeur par ti formé de crétacé inférieur moyen et supérieur ;les terrasses alluviales sont beaucoup plus restreintes. Les terrains rencontrés

sont généralement imperméables : argilo grès de l'oligocène, schistes de l'alboaptien, marno calcaires du crétacé (Tebbani, 2016).

V.1.1.2.Le réseau hydrographique

Le bassin versant de la Soummam est très développé hydro graphiquement par ses divers cours d'eau. La majorité de ses eaux sont drainées par l'Oued de Boussellam et Oued Sahel qui se rencontrent près d'Akbou formant l'Oued Soummam (Figure. V.2).

En ce qui concerne les eaux souterraines, dans la grande partie de bassin versant de la Soummam, elles sont confinées dans des couches profondes et ne peuvent remonter en surface que par des courants ascendants (Tebbani,2016) .

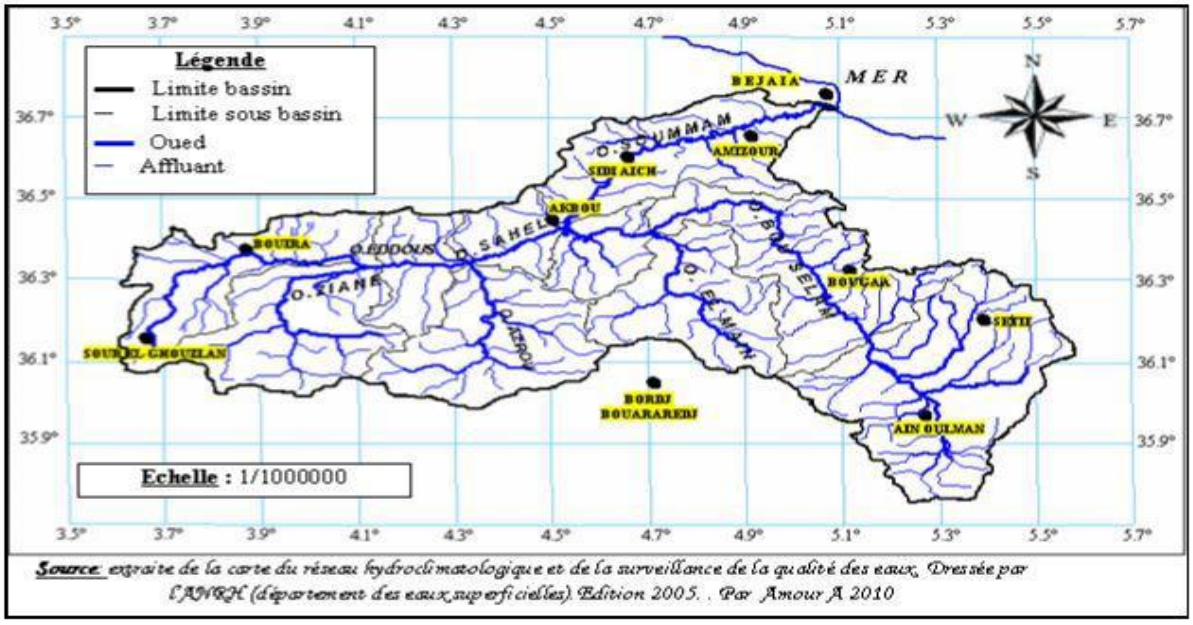


Figure V.2: Réseau hydrographique du bassin versant de la Soummam (Ramdini, 2016)

V.1.1.3.Situation climatique

Climat peut être défini comme étant l'ensemble des phénomènes météorologiques qui caractérisent l'état de l'atmosphère et de son évolution en un lieu donné. Les principales composantes du climat sont : les précipitations, la température et le vent (Ramdini, 2016).

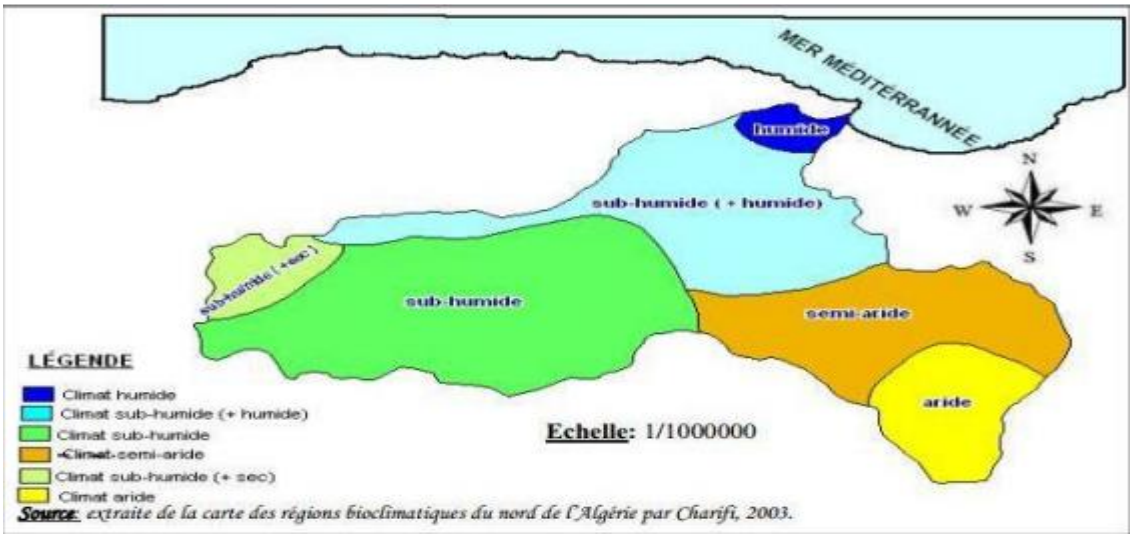


Figure V.3: Etages bioclimatiques du bassin Soummam (Ramdini, 2016).

- **Les températures**

Le bassin de la Soummam est caractérisé généralement par deux saisons :

- Une saison chaude allant du mois de Juin au mois de Septembre.
- Une saison froide nettement plus longue, allant du mois d'Octobre au mois de mai .

Les valeurs de la température moyenne mensuelle enregistrées aux niveaux de 04 stations climatiques représentatives dans le bassin sont mentionnées dans le tableau V.1. La période d'observation est de 1993 à 2003(Ramdini, 2016).

TableauV.1: Températures moyennes mensuelles

Station	Sept	Oct	Nov	Déc	Jan	Fév	Mar	Avr	Mai	Juin	Juil	Aout
Bouira	22.3	18.1	12.6	9.2	8.3	8.8	11.7	13.7	18.6	28.9	37.1	37.5
Béjaia	23.3	20.1	16.0	13.2	11.4	11.5	12.9	15.2	18.9	29.4	33.9	36.3
B.B.A	22.1	16.6	10.6	7.0	6.3	7.2	10.4	12.8	17.9	27.0	36.6	37.2
Sétif	20.3	15.7	9.8	6.4	5.4	6.2	9.4	11.6	17.7	29.8	35.4	36.1
Moy Mensuelle	22.0	17.6	12.2	9.0	7.8	8.4	11.1	13.3	18.3	28.8	34.5	36.7

D'après le tableau V.1 on peut constater que les mois les plus froids sont Décembre, Janvier et Février tandis que les mois les plus chauds sont Juillet et Août (Ramdini, 2016).

❖ **Les précipitations**

La pluviométrie dans le bassin de la Soummam est déterminée grâce à l'existence de 41 stations pluviométriques représentatives prises en compte par l'ANRH dans le cadre du projet PNUD/ALG/88/021, pour l'étude de la pluviométrie de l'Algérie du Nord.

Le régime de la pluviométrie moyenne annuelle est connu grâce à l'interprétation de la carte pluviométrique à l'échelle 1/500 000 éditée par l'ANRH en 1993 dans le cadre du projet suscité.

Il en ressort d'après cette carte que la pluviométrie moyenne annuelle pour l'ensemble du bassin oscillant entre 300 et 1000 mm en augmentant d'Ouest vers l'Est (Ramdini, 2016).

❖ **Le vent**

Le phénomène est habituellement accompagné d'une part d'une évaporation accentuée de la surface du sol et de la végétation et d'autre part du dessèchement du sol et de la couverture végétale. Des vents chauds et secs brûlent les champs de blé par suite d'une réduction d'humidité dans les plantes et de l'évaporation de la couche superficielle du sol.

Dans le bassin versant de la Soummam est canalisé par les massifs montagneux voisins sa direction prédominante est nord-est et sud-ouest(ONM)(Ramdini, 2016).

V.1.1.4.Géologie générale

La répartition des formations géologiques de bassin versant de l'oued Soummam est comme suit:

- Sur la rive gauche par l'oligocène traversée par des formations du crétacé inférieur; du miocène inférieur apparait dans la partie aval, à coté de l'oued de terrasses

alluviales qui recouvrent en général le pied des pentes sauf dans la région de Sidi Aich ou le crétacé apparait dans le lit.

- Sur la rive droite est formée par le crétacé inférieur moyen et supérieur, les terrasses alluviales sont beaucoup plus restreinte (Bouchellah et Mazoua, 2020).

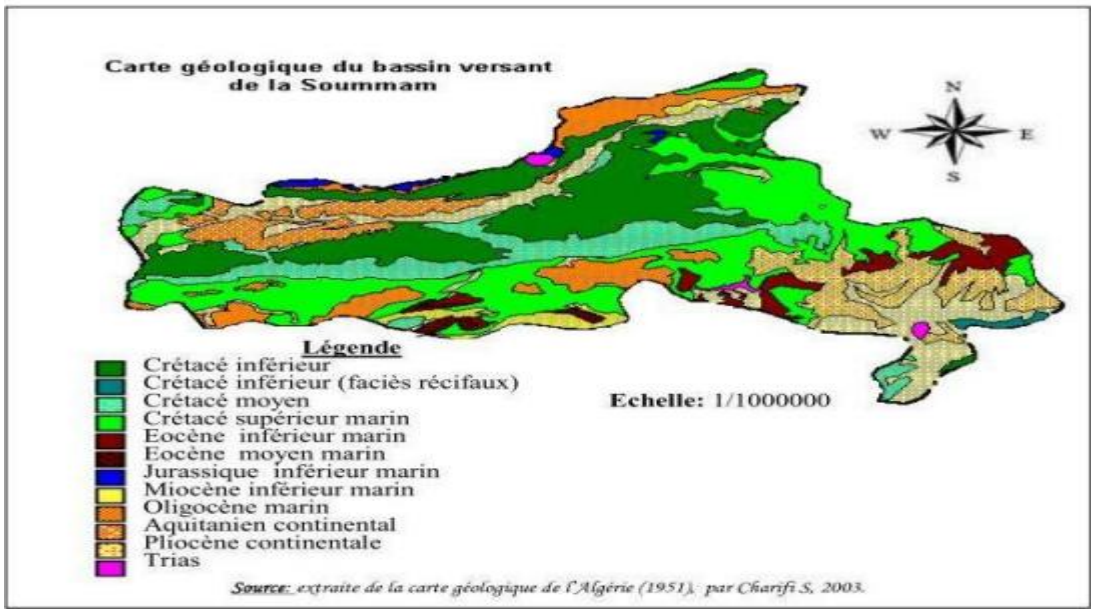


Figure V.4: Carte géologique du bassin versant de la Soummam (Ramdini, 2016)

V.1.2.Donnés pluviométriques utilisées

Les données pluviométriques mensuelles de 5 stations son utilisées dans le présent travail, dont les caractéristiques sont présentées dans le tableau suivant :

Tableau V.2: caractéristiques des cinq stations pluviométriques

Nom de la station	Code (A.N.R.H)	Période de fonctionnement
Sour El Ghozlane	150101	1980 -2006
El EsnamSh	150204	1980 -2006
Ben Daoud	150402	1980 -2006
TenietEnasr	150807	1980 -2006
SidiYahia	150904	1980 -2006

Les 5 stations utilisées sont schématisées sur la carte V.5.

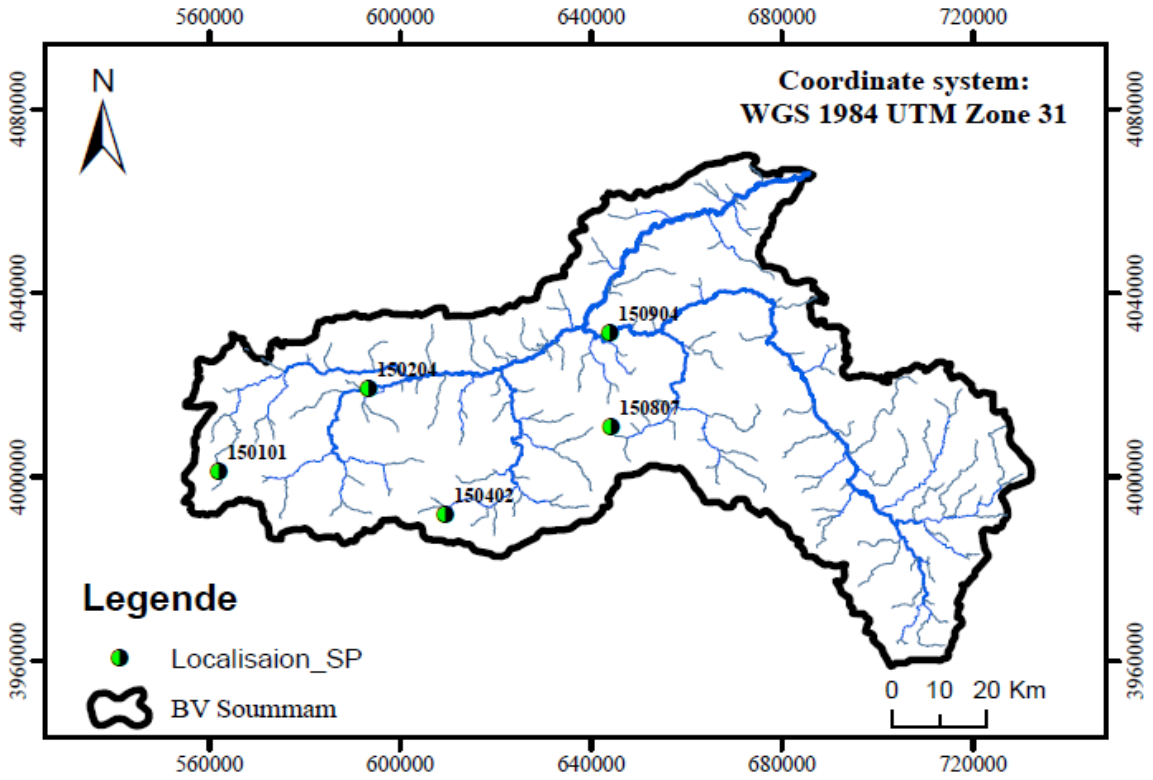


Figure V.5: Localisation des 5 stations pluviométriques sur un extrait de la carte du réseau hydro-climatologique

- **Les stations sont schématisées sur la carte sont**

- ✓ Station de Sour El Ghozlane .
- ✓ Station de El Esnam Sh(Station de base).
- ✓ Station de Ben Daoud.
- ✓ Station de TenietEnasr.
- ✓ Station de SidiYahia.

V.1.3.Répartition des données

Toutes les méthodes sont utilisées pour estimer les données de la station de base ***m*** de El Esnam Sh, supposées manquantes dans le but de tester la qualité de l'estimation.

Les données mensuelles des cinq stations sont divisées en deux parties, la première partie pour le calage avec 70% (12 années), la deuxième partie pour la validation avec 30% (5années).

- La partie calage est constituée des données mensuelles des 12 années suivantes : 80-81 à 85-87 ; 92 ; 94-97 ; 2000-2001 ;.
- La partie validation constituée des données mensuelles des 6 années suivantes : 2001-2006.

V.1.4. Critères de comparaison

Les performances des différentes méthodes d'estimations sont comparées en utilisant deux types de critères de comparaison :

a) Critères graphiques : L'analyse graphique est indispensable et primordial, cela est obtenu

en portant sur un graphique les valeurs estimées par les différentes méthodes d'estimation, en fonction de celles observées.

c) Critères numériques : Les critères numériques de comparaison les plus recommandés (Kanevski and Maignan, 2004; Chang, 2004; Ahrens, 2006;Legates and McCabe,1999) sont :

- Racine de l'erreur moyenne quadratique (RMSE);
- Erreur moyenne absolue (MAE);
- Le critère de Nash (NSE) ;
- Erreur moyenne relative (MRE) .

Les expressions des différents critères sont données par les équations

$$\textbf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{P_i} - P_i)^2} \qquad \text{V.1}$$

$$\textbf{MRE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(\widehat{P_i} - P_i)}{P_i}\right| \qquad \text{V.2}$$

$$\textbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|(\widehat{P_i} - P_i)\right| \qquad \text{V.3}$$

$$\textbf{NSE} = 1 - \frac{\sum_{i=1}^{n}(\widehat{P_i} - P_i)^2}{\sum_{i=1}^{n}(\widehat{P_i} - P_i)^2} \qquad \text{V.4}$$

Où

$\widehat{P_i}$: Valeur estimée ;

P_i: Valeur observée ;

n : Nombre total d'observation.

V.1.5.Evaluation des coefficients optimaux

Pour améliorer la précision des quatre fonctions FFSGAM[1], FFSGAM[2], FFSGAM[3],FFSGAM[4], Nous avons estimé les coefficients C_i en minimisant l'erreur moyenne quadratique donnée par l'expression V.5, au moyen du Microsoft Excel Solveur qui utilise des algorithmes d'optimisation non linéaire GRG (generalizedreduced gradient).

$$\left[\frac{1}{N}\sum_{i=1}^{n}\left(\left(P_i - \widehat{P}_i\right)^2\right]\right. \qquad \text{V.5}$$

Où :

$\widehat{P_i}, P_i$: Sont respectivement les valeurs des précipitations estimée et observée dans la station de base m ;

N : Nombre de mois.

V.1.6.Application des méthodes d'estimations

Toutes les méthodes sont utilisées pour estimer les données de la station de base mde El Esnam Sh supposées manquantes dans le but de tester l'estimation.

V.1.6.1.Méthode IDWM

a) Evaluation des facteurs de pondération d_{mi}

Pour IDWM les distance d_{mi},sont mesurées sur la carte du réseau hydroclimatologique et de surveillance de la qualité des eaux à l'échelle 1 : 500 000.

Les distances d_{mi}, entre la station de base (station de El Esnam Sh) et les autres stations i sont portées dans le tableau V.3

Tableau V.3: Facteur de pondération d_{mi}

Station i	**150101**	**150402**	**150807**	**150904**
d_{mi}(m)	34291.86	26314.58	50408.11	51173.04

V.1.6.2.Méthode CCWM

L'application de cette méthode nécessite uniquement les coefficients de corrélation R_{mi} entre chaque station i et celle de base que nous allons donner dans chaque cas d'estimation.

Tableau V.4: Coefficients de corrélation entre la station de base et les autres stations i

Station pluviométrique i	**150101**	**150402**	**150807**	**150904**
Coef.de corrélation	0.87	0.57	0.83	0.86

V.1.6.3.Méthode FFSGAM

Cette méthode va être appliquée en deux cas :

a) Les coefficients locaux C_i sont évalués en minimisant l'erreur moyenne quadratique entre les valeurs estimées et celles observées, a l'aide du Microsoft Excel Solveur ;

b) Tous les coefficients locaux C_i sont pris égaux à l'unité.

L'application de cette méthode nécessite deux paramètres déjà calculés : les distances d_{mi} et les coefficients de corrélation R_{mi}, entre chaque station i et la station de base m(station de ElEsnam Sh).

Tableau V.5: Coefficients optimaux Ci

C_i / **FFSGAM**	**C_1**	**C_2**	**C_3**	**C_4**
FFSGAM[1]	0.23642	0.32374	-0.86446	6.30816
FFSGAM[2]	0.14339	1.02401	-0.54716	3.33570
FFSGAM[3]	0.73815	0.00047	-0.60286	10.94380
FFSGAM[4]	0.13596	0.388451	-0.77748	5.40144

1.6.4.Méthode Apprentissage profond (Deep learning)

L'application de la méthode de comblement des lacunes par Apprentissage profond (Deep Learning) a été faite à l'aide du logiciel MATLAB version 2020.

1.6.4.1. Présentation du logiciel MATLAB

Matlab (MATrixLABoratory) est un logiciel pour effectuer des calculs numériques. Il a été conçu initialement pour faciliter le traitement des matrices mais il est maintenant utilisé dans tous les domaines des sciences qui nécessite de faire des calculs contient également une interface graphique, ainsi qu'une grande variété d'algorithmes scientifiques.

1.6.4.2.Le rôle de MATLAB

Les cas d'usage de Matlab sont nombreux. Le langage de programmation est notamment utilisé dans:

- Les systèmes de contrôle ;
- Le machine et le deep learning ;
- La maintenance prédictive ;
- Le traitement du signal et les séries temporelles ;
- L'automatisation des tests ;
- Les systèmes de télécommunication, la robotique.

Conclusion

La première partie de ce chapitre est consacrée à la collecte de toutes les données Caractérisation de la zone d'étude et traitement des précipitations. Dans la deuxième partie les différentes étapes de l'application des méthodes de cambrement de lacune , les résultats de ces méthodes et leurs interprétations seront présentés dans le prochain chapitre.

Chapitre VI Résultats et interprétations

Chapitre VI : Résultats et interprétations

Introduction

Ce chapitre est le fruit du travail puisqu'il présente l'application des méthodes évoquées dans le chapitre précédent et montre les différents résultats obtenus.Il fournit également des interprétations et les conclusions qui en résultent.

VI.1.Estimation avec des coefficients globaux égaux à l'unité

Pour des calculs rapides, sans passer par l'optimisation des coefficients locaux (un coefficient pour chaque station pluviométrique) Ci, on se propose de tester les quatre modèles de la FFSGAM en prenant ces coefficients égaux à l'unité.

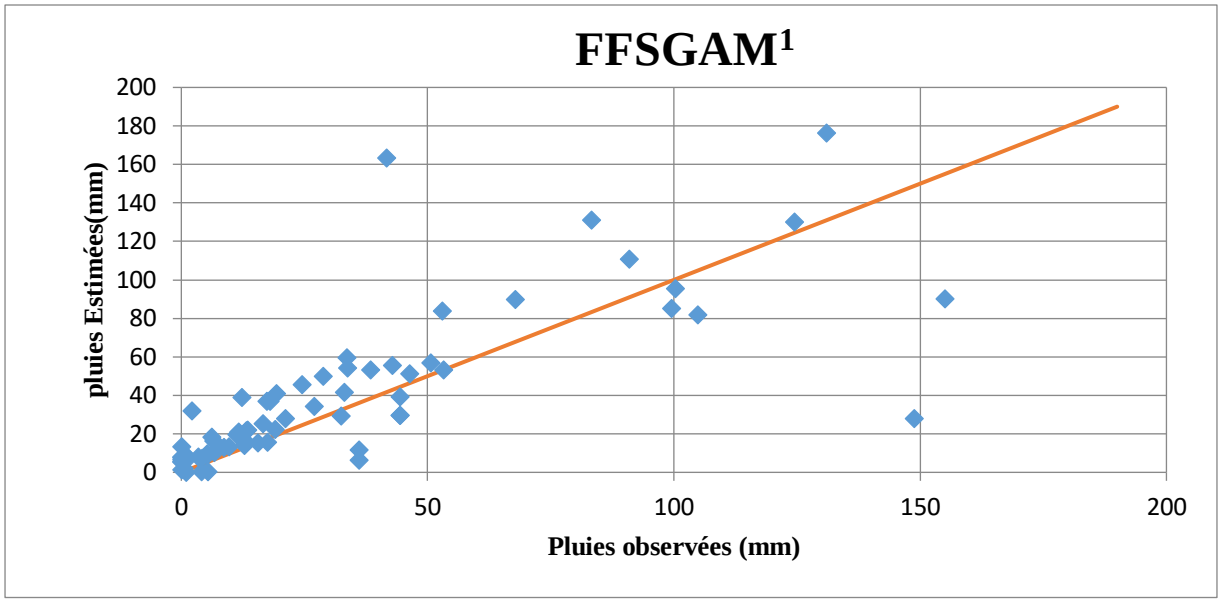
VI.1.1.Estimation avec Ci=1

- Les critères de comparaison entre les différentes méthodes sont résumés dans le Tableau VI.1.

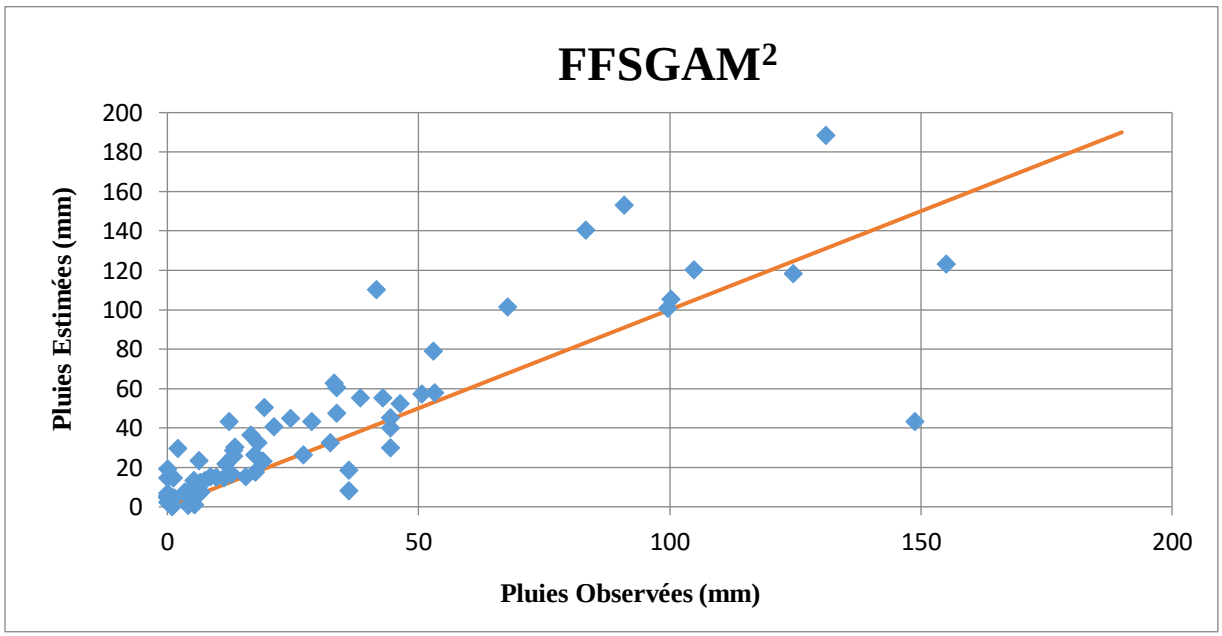
Tableau VI.1:Critère de comparaison (Estimation Ci=1)

FFSGAM / Critère	FFSGAM^1	FFSGAM^2	FFSGAM^3	FFSGAM^4
RMSE	26.94	24.19	35.27	25.71
RME	1.04	1.10	1.09	1.05
MAE	15.62	15.54	18.73	15.31
NSE	0.48	0.58	0.11	0.58

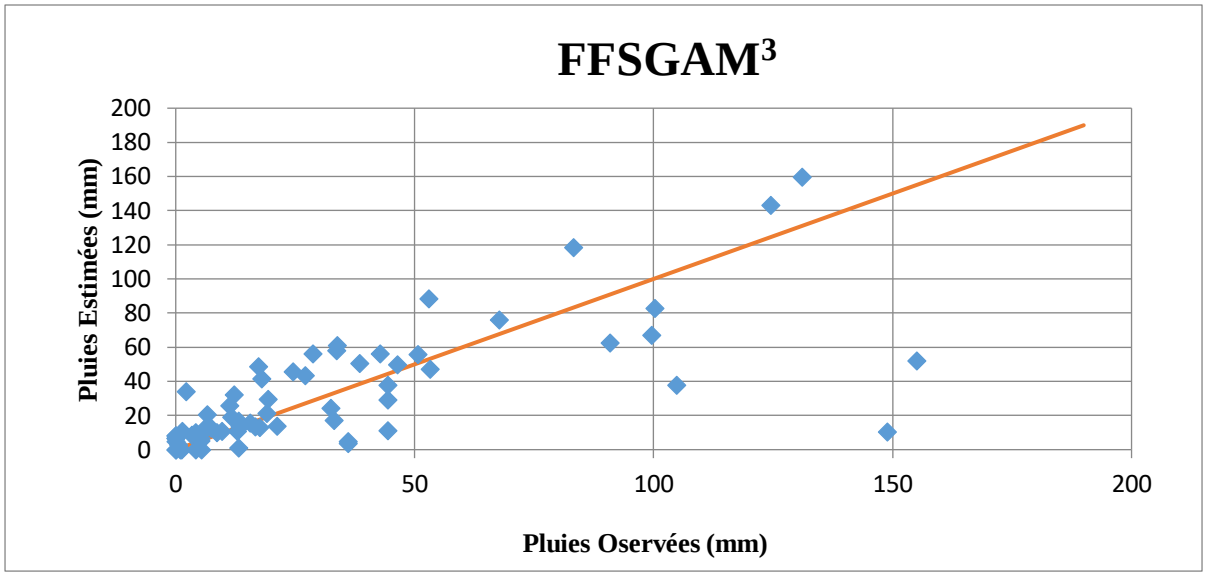
Le tableau VI.1 : montre que même avec des coefficients locaux Ci égaux à l'unité tous les modèles de la FFSGAM ont donné des résultats acceptables.



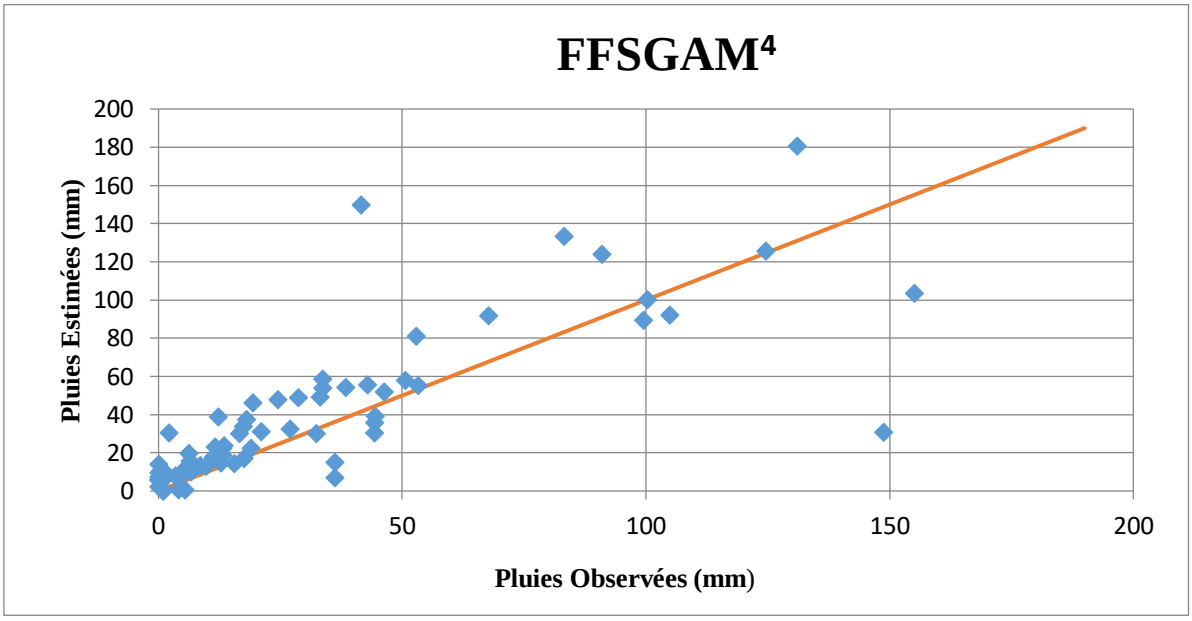
a: méthode FFGSAM[1]



b : méthode FFGSAM[2]



c: méthode FFGSAM³



d: méthode FFGSAM⁴

Figure VI.1 : Valeurs estimées en fonction de celles observées (Estimation avec Ci=1).

VI.2.Estimation des données manquantes

a) Les critères de comparaison entre les différentes méthodes sont résumés dans le tableau VI.2.

Tableau VI.2: Critère de comparaison

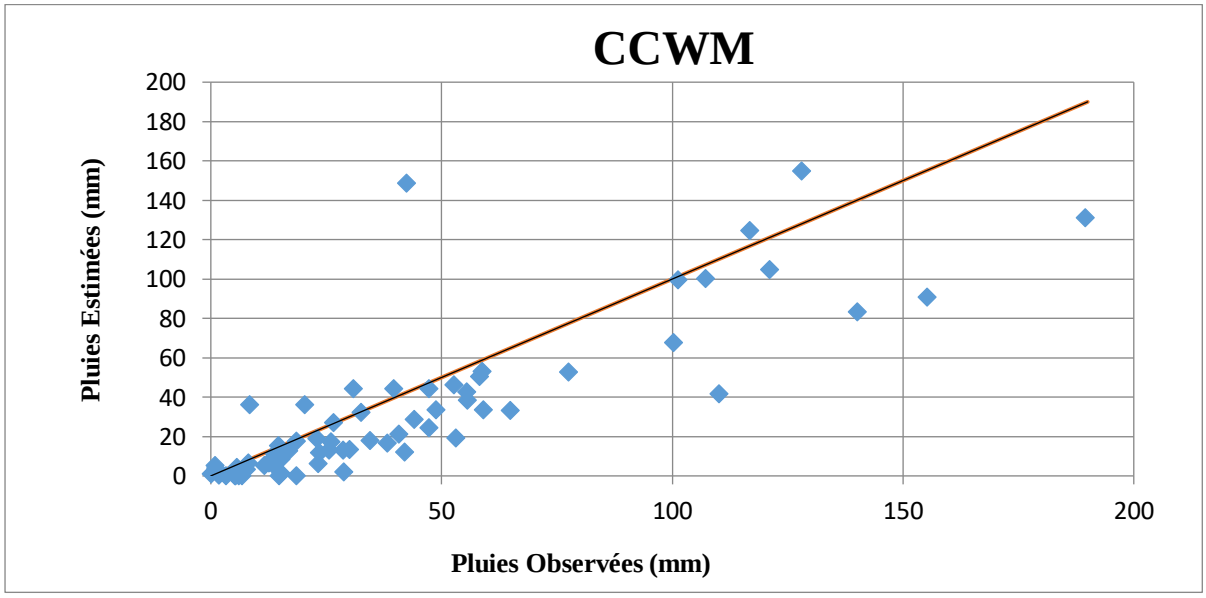
Méthode / Critère	CCWM	IDWM	FFSGAM[1]	FFSGAM[2]	FFSGAM[3]	FFSGAM[4]	ANN
RMSE	24.33	3.17	20.70	20.70	20.70	20.70	19.20
MRE	1.10	1.06	0.95	0.95	0.95	0.95	0.81
MAE	15.72	15.49	12.96	12.96	12.96	12.96	12.49
NSE	0.58	0.51	0.70	0.70	0.70	0.67	0.74

b) L'analyse des critères présentés par tableau VI.2 nous permet de tirer que :

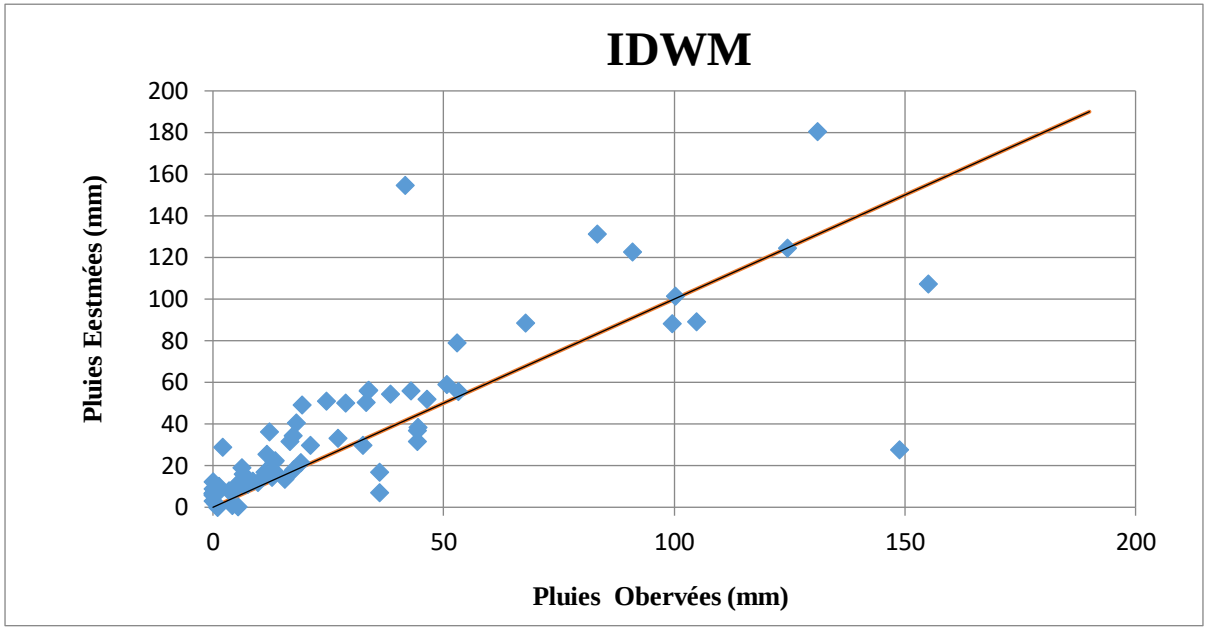
- Toutes les méthodes utilisées ont donné de très bons résultats d'estimations.

La méthode ANN a donné des résultats plus performants que toutes les autres méthodes.

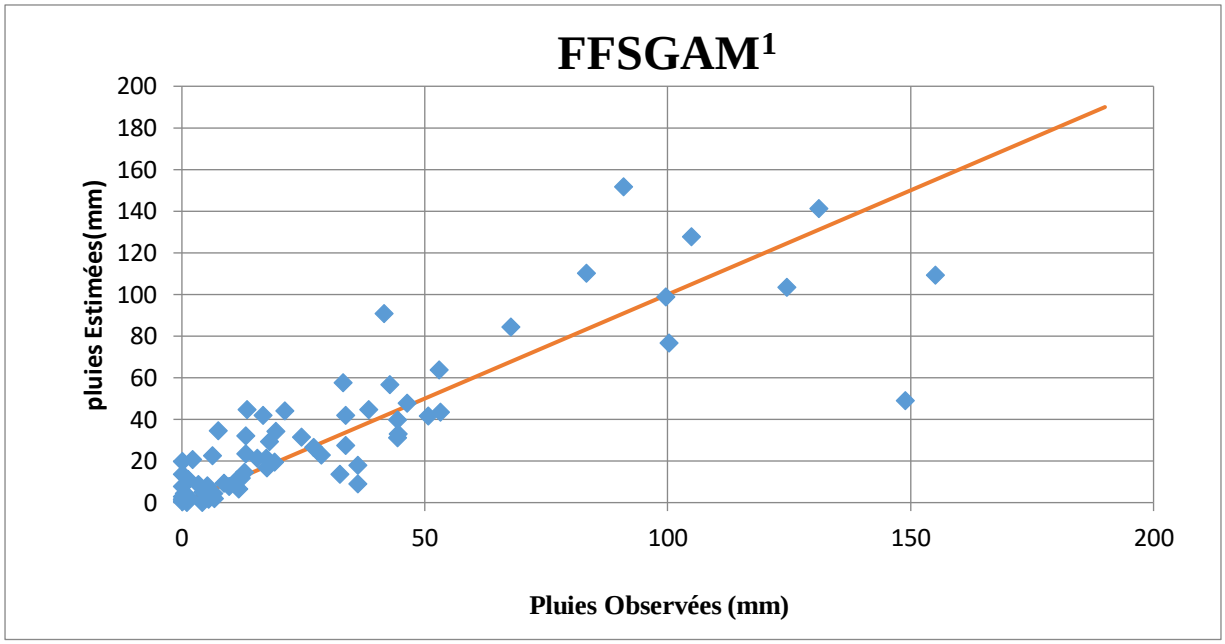
- Les quatre modèles FFSGAM convergent avec les mêmes valeurs.



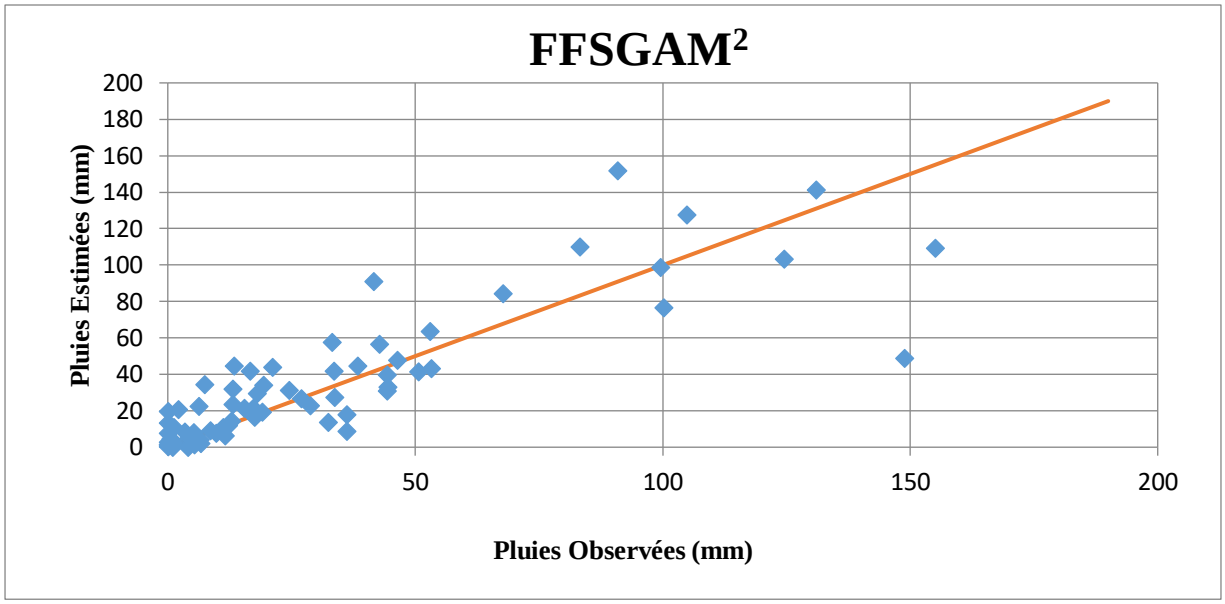
a : Méthode CCWM



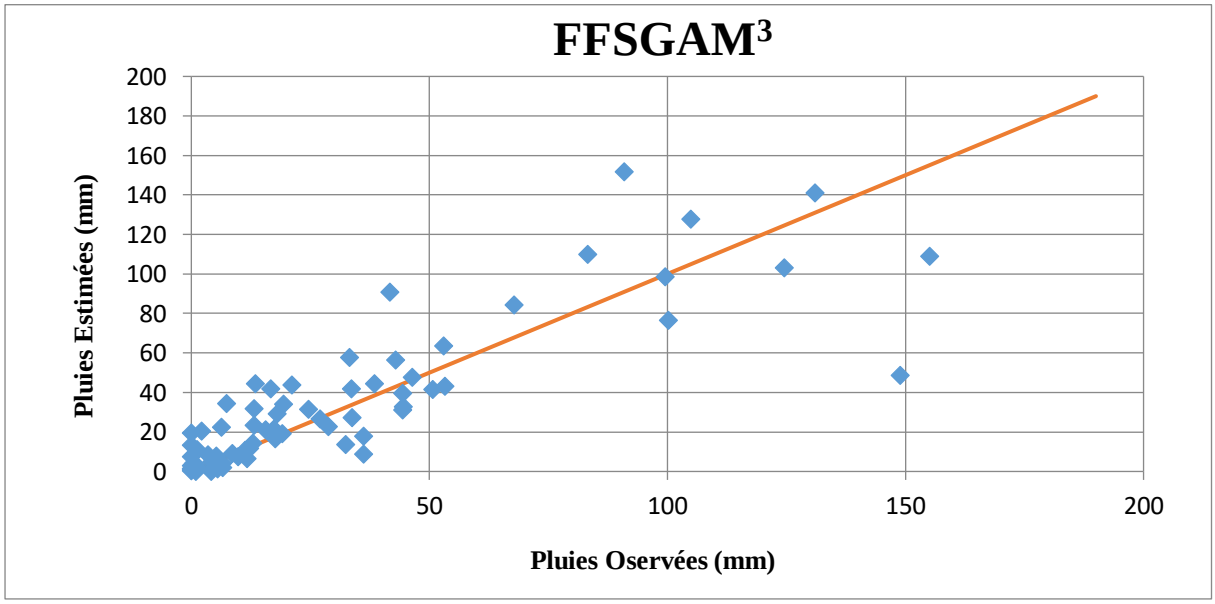
b : Méthode IDWM



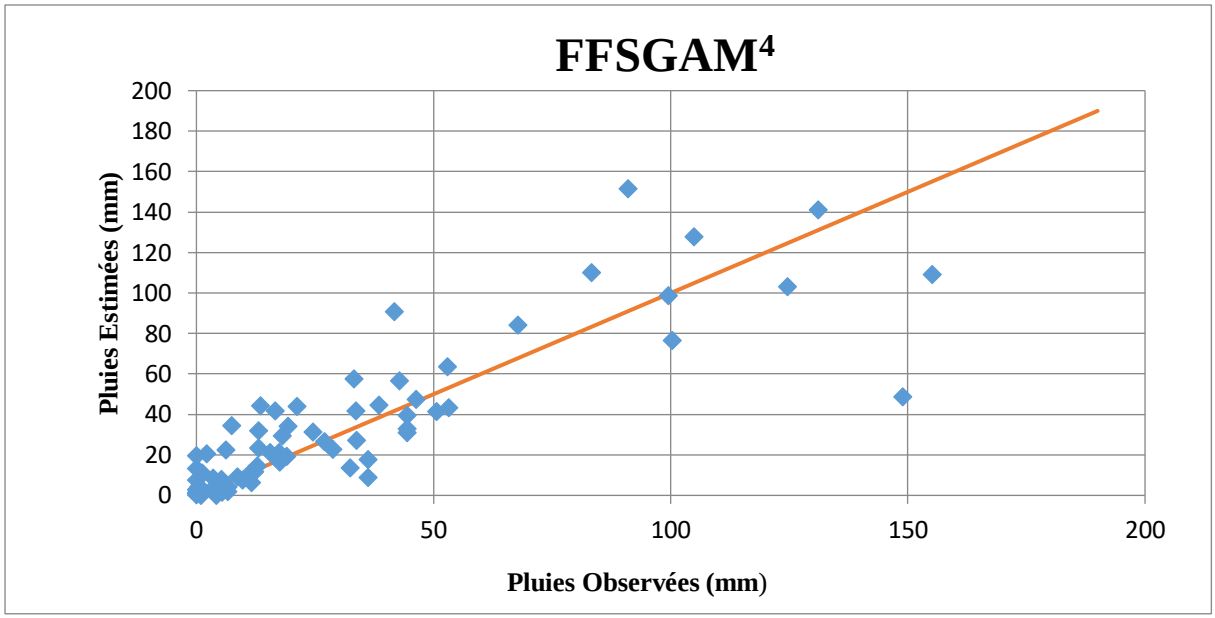
c : méthode FFGSAM¹



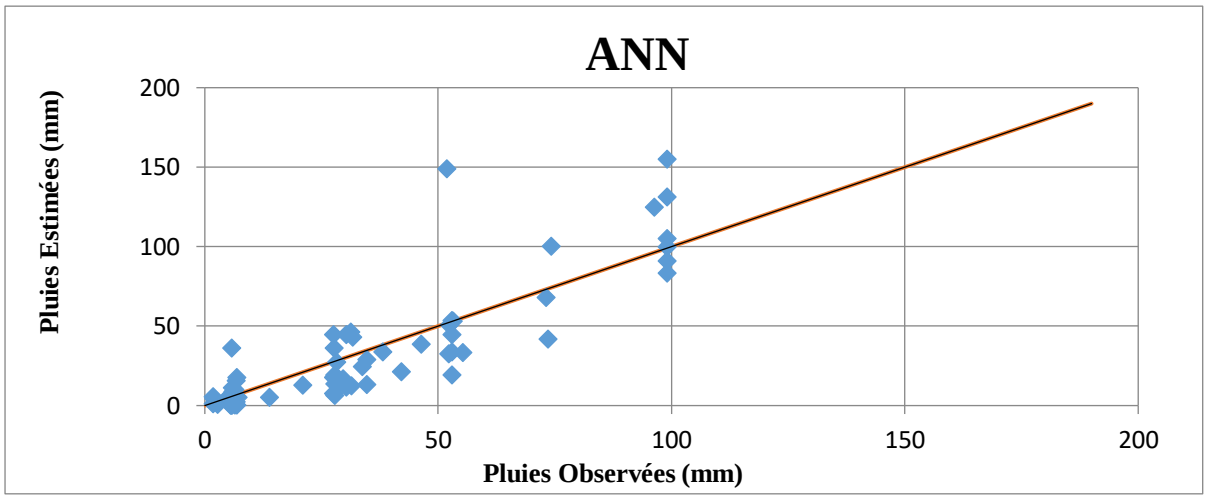
d : méthode FFGSAM²

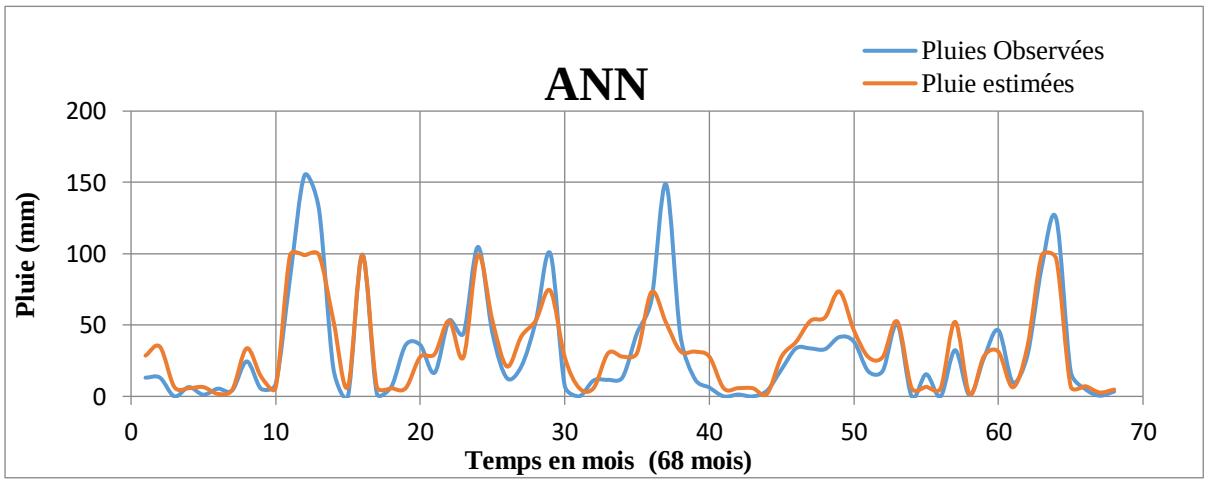


e: méthode FFGSAM³



f: méthode FFGSAM⁴





g: méthode ANN

Figure VI.2: Montre les graphes des valeurs estimées en fonctions de celles observées

VI.3.Interprétations sur les graphes

l apparait dans la figure VI.2 que :

- Pour toutes les méthodes utilisées, les nuages des points des valeurs estimées s'alignent bien sur la première bissectrice ;
- Le nuage des points estimés par la méthode CCWM et IDWM s'alignent sur la première bissectrice , on observe qu'il a une faible capacité d'estimation par rapport aux méthodes FFSGAM et ANN ;
- Le nuage des points estimés par la méthode FFSGAM s'alignent sur la première bissectrice il donne de bons résultats d'estimation bien meilleurs que les méthodes CCWM et IDWM ;
- Le nuage des points estimés par la méthode ANN s'alignent sur la première bissectrice, et le nuage de points indique que les points sont plus proches de la ligne de tendance et nous en concluons que la méthode ANN est meilleure que les autres méthodes dans l'estimation des données manquantes dans les enregistrements des précipitations.

Conclusion

Sur la base des résultats obtenus nous pouvons conclure que, toutes les méthodes utilisées ont donné de bons résultats d'estimations.

La méthode ANN a donné des résultats plus performants que toutes les autres méthodes dans l'estimation des données manquantes dans les enregistrements des précipitations.

Conclusion générale

Tous les pays du monde souffrent du problème de la perte de données. En Algérie, les services hydrologiques reconstituent généralement les observations de précipitations manquantes à l'aide de méthodes simples.

Dans d'autres pays, comme les États-Unis d'Amérique, plusieurs méthodes sont utilisées pour estimer les données manquantes, telles que les méthodes classiques, et récemment, des méthodes basées sur l'intelligence artificielle telles que les algorithmes génétiques et les réseaux de neurones artificiels ont été utilisées.

Dans ce travail, nous avons mené une étude comparative entre différentes méthodes d'estimation des données manquantes dans les relevés pluviométriques du bassin versant de d'oued Soummam en sélectionnant 5 stations en fonction de la disponibilité des données pluviométriques fournies par l'Agence Nationale des Ressources en Eau. L'étude a été menée sur un pas de temps mensuel de la période étendue. De 1980 à 2006 en utilisant les méthodes suivantes:

- méthodes classiques: (Inverse distance weightingméthod) IDWM, Méthode CCWM (Coefficient of correlation weighing method);
- Méthode basée sur l'intelligence artificielle (basée sur les algorithmes génétiques) FFSGAM et ANN (réseaux de neurones artificiels).

Les résultats des différentes méthodes ont été comparés en utilisant 6 indices différents, l'erreur absolue moyenne (MAE) et l'erreur quadratique moyenne (RMSE) et l'erreur moyenne relative (MRE) et le critère de Nash (NSE).

Et sur la base des résultats obtenus, chacune des méthodes utilisées a donné de bons résultats, et la méthode ANN a obtenu de meilleurs résultats, car elle a donné une valeur RMSE de 19,20 et une valeur MRE de 0,81 et une MAE de 12,49 et une valeur NSE de 0,74 mieux que toutes les autres méthodes.

Références bibliographiques

Références bibliographiques

Ahrens, B., 2006. Distance in spatial interpolation of daily rain gauge data. Hydrology and Earth System Sciences 10, 197–208.

ASCE, 1996. Hydrology Handbook, second ed. American Society of Civil Engineers (ASCE), New York.

ASCE, 2001a. Task committee on artificial neural networks in hydrology, artificial neural networks in hydrology. I. Preliminary concepts. Journal of HydrologicEngineering, ASCE 52, 115–123.

ASCE, 2001b. Task committee on artificial neural networks in hydrology, artificial neural networks in hydrology. II. Hydrologic Applications. Journal of Hydrologic Engineering, ASCE 52, 124–137.

Ashraf, M., Loftis, J.C., Hubbard, K.G., 1997. Application of geostatistics to evaluate partialweather station network. Agricultural Forest Meteorology 84, 255–271.

Alouache, R et Chia, R, 2019.Évaluation de la performance d'un capteur logiciel en utilisant l'apprentissage en profondeur Mémoire de Master Académique Électronique Université Mohamed Boudiaf de M'sila.

Beasley,D.,Martin,R.,1993.AnOverview of GeneticAlgorithms, Part 2. www. citeseerx.ist. /viewdoc.

Brimicombe, A., 2003. GIS, Environmental Modeling and Engineering. Taylor and Francis, London, UK.

Bouchellah, Z et Mazoua,A, 2020.Hydrodynamique des eaux souterraines de la basse Soummam Mémoire de fin d'étude master en hydraulique Université A.Mira-Bejaia .

Chang, K.-T., 2004. Introduction to Geographic Information Systems. McGraw Hill, New York.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Mathematical Control Signals Systems 2, 303–314.

Daly, C., Neilson, R.P., Phillips, D.L., 1994. A statistical topographic model for mapping climatological precipitation over mountainous terrain. Journal of Applied Meteorology 33, 140–158.

Daly, C., Gibson, W.P., Taylor, G.H., Johnson, G.H., Pasteris, P., 2002. A knowledgebasedapproach to the statistical mapping of climate. Climate Research 22, 99–113.

Djerbouai S. Missing Precipitation Data Estimation Using Long Short-Term Memory Deep Neural Networks. Journal of Ecological Engineering. 2022;23(5):216-225.

Dingman, S.L., 2002. Physical Hydrology. Prentice Hall, NJ.

Duby,C.,Robin,S.,2006.Analyse en composantes principales.Département O.M.P.I.Institut national agronomique Paris.france.www.math.univ-Lyon.fr.

French, M.N., Krajewski, W.F., Cuykendal, R.R., 1992. Rainfall forecasting in space and time using a neural network. Journal of Hydrology 137, 1–37.

Foued, N, 2019.Reconnaissance d'expression faciale à partir d'un visage réel Mémoire de Fin d'études Master en Informatique Université de 8 Mai 1945 – Guelma .

Giustolisi, O., Savic, D.A., 2004. A novel genetic programming strategy: evolutionarypolynomial regression. In: Liong, S.-Y., Phoon, K.-K., Babovic (Eds.), Proc. of the 6thInternational Conference on Hydroinformatics, vol. 1, Singapore, 21–24 June. World Scientific Publications, New Jersey, pp. 787–794.

Giustolisi, O., Savic, D.A., Doglioni, A., Laucelli, D., 2004. Knowledge discovery byevolutionary polynomial regression. In: Liong, S-Y., Phoon, K.-K., Babovic (Eds.), Proc. of the 6th International Conference on Hydroinformatics, vol. 2, Singapore, 21–24 June. World Scientific Publications, New Jersey, pp. 1647–1654.

Goldberg, D.E., 1989. Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley, New York.

Govindaraju, R.S., Rao, A.R., 2000. Neural Networks in Hydrology. Kluwer AcademicPublishers, Netherlands.

Grayson, R., Bloschl, G., 2001. Spatial Patterns in Catchment Hydrology: Observations and Modeling. Cambridge University Press.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2 (5), 359–366.

Kanevski, M., Maignan, M., 2004. Analysis and Modelling of Spatial Environmental Data. EPFL Press, Lausanne, Switzerland.

Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA.

Krajewski, W.F., 1987. Co-kriging of radar and rain gauge data. Journal of Geophysics Research 92 (D8), 9571–9580.

Laborde,J.P.,1998.Notice d'utilisation du logiciel hydrolab.CNRS.France.42.

Laborde,J.P.,1998.Logiciel Hydrolab,version 98.2 .CNRS.France.

Laborde,J.P., 2003.Hydrologie de surface. Presses de Dar El-Houda Ain M'lila Algerie.191.

Llamas,J.,1993.Hydrologie générale. Presses Gaëtan Morin C.P.180, Boucherville, Québec Canada.

Labiad, A, 2017.Sélection des mots clés basée sur la classification et l'extraction des règles d'association Mémoire présenté à L'Université du Québec à Trois-Rivières.

Le Goulven, P,1988.Hydrologue ORSTOM, Mission ORSROM, Quito, Equateur CP17-11-06596.

Meylan,P ., Musy,A .,1999. Hydrologie fréquentielle. Presses de l'office Fédérale de l'éducation et de la science, Suisse (NO 96.01). 413.

Nabonne,A.,2004.Algorithmesévolutionnairesetproblèmesinverses.www.enseignement polytechnique.fr

Pechlivanidis,I.G.Segond,M.L.,McIntyre,N.,Wheater, H.S,2005 Optimal functional forms for estimationofmissingprecipitationdata.www.3imperial.ac.uk/pls/porfallive/docs/1/35540/pdf.

Rogers, D., Hopfinger, A.J., 1994. Application of genetic function approximation toquantitative structure–activity relationships and quantitative structure– property relationships. Journal of Chemical Information and ComputerSciences 34, 854–866.

Ramdini, M, 2016.Etude de la Sécheresse cas du bassin versant de la Soummam Mémoire de Master en Hydraulique Ecole Nationale Supérieure D'Hydraulique Arbaoui Abdellah Département Hydraulique Urbaine.

Sari, A.A.,2002. Hydrologie de surface. Presses Houma. Algérie.223.

Salas, J.D.-J., 1993. Analysis and modeling of hydrological time series. In: Maidment,D.R. (Ed.), Handbook of Hydrology, vol. 19. McGraw-Hill, NY, pp. 19.1–19.72(chapter 19).

Seo, D.-J., 1996. Nonlinear estimation of spatial distribution of rainfall – an indicatorcokriging approach. Stochastic Hydrology and Hydraulics 10, 127–150.

Seo, D.-J., Krajewski, W.F., Bowles, D.S., 1990a. Stochastic interpolation of rainfall data from rain gauges and radar using cokriging – 1. Design of experiments.Water Resources Research 26 (3), 469–477.

Seo, D.-J., Krajewski, W.F., Bowles, D.S., 1990b. Stochastic interpolation of rainfalldata from rain gauges and radar using cokriging – 2. Results. Water ResourcesResearch 26 (5), 915–924.

Shi, L.M., Fan, Y., Myers, T.G., O'Connor, P.M., Paull, K.D., Friend, S.H., Weinstein, J.N.,1998. Mining the NCI anticancer drug discovery database: genetic function approximation for the QSAR study of anti-cancer ellipticine analogues. Journalof Chemical Information and Computer Sciences 38, 189–199.

Simanton, J.R., Osborn, H.B., 1980. Reciprocal-distance estimate of point rainfall.Journal ofHydraulic Engineering Division 106 (HY7), 1242–1246.

Singh, V.P., Chowdhury, K., 1986. Comparing some methods of estimating mean aerial rainfall. Water Resources Bulletin 222, 275–282.

Smith, J.A., 1993. Precipitation. In: Maidment, D.R. (Ed.), Handbook of Hydrology,vol. 3. McGraw Hill, New York (chapter 3).

Sullivan, D.O., Unwin, D.J., 2003. Geographical Information Analysis. John Wiley & Sons, Inc., NJ.

SOUAG-GAMANE, D, 2007.Développement d'outils pour la gestion des barrages réservoirs basés sur la simulation et la prévision des paramètres hydrométéorologiques Thèse Doctorat d'état en Génie Civil Université des Sciences et de la Technologie Houari Boumediene .

Teegavarapu, R.S.V., 2007. Use of universal function approximation in variance de pendentinterpolation technique: An application in hydrology. Journal of Hydrology 332, 16–29.

Teegavarapu, R.S.V., 2008. Innovative spatial interpolation methods for estimation of missing precipitation records: concepts and applications. In: Bruthans, J.,Kovar, K., Hrkal, Z.(Eds.), Proceedings of HydroPredict 2008, Prague, pp. 79–82.

Teegavarapu, R.S.V., 2009. Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules. Journalof Hydroinformatics11 (2), 133–146.

Teegavarapu,R.S.V, Tufail,M.;Ormsbee,.L.E., 2009.Optimal functional forms for estimation of missing precipitation data. Journal of hydroloy, 374, 106-115.

Teegavarapu, R.S.V., Chandramouli, V., 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missingprecipitation records. Journal of Hydrology 312, 191–206.

Tomczak, M., 1998. Spatial Interpolation and its uncertainty using automatedanisotropic inverse distance weighting (IDW)–cross-validation/jackknifeapproach. Journal of Geographic Information and Decision Analysis 2, 18–30.

Touaibia, B., 2004. Manuel Pratique d'Hydrologie. Presses Madani Frères. Blida.Algérie.166.

Tufail, M., Ormsbee, L.E., 2006. A fixed functional set genetic algorithm (FFSGAM) approach for functional approximation. IWA Journal of Hydroinformatics 3, 193–206.

Tebbani, R, 2016.Etude du transport solide à l'estuaire du bassin versant de la Soummam par le logiciel HEC-RAS Mémoire de Master en Hydraulique Université Mohamed Boudiaf de M'sila.

Vieux, B.E., 2001. Distributed Hydrologic Modeling using GIS, Water Science andTechnology Library. Kluwer Academic Publishers.

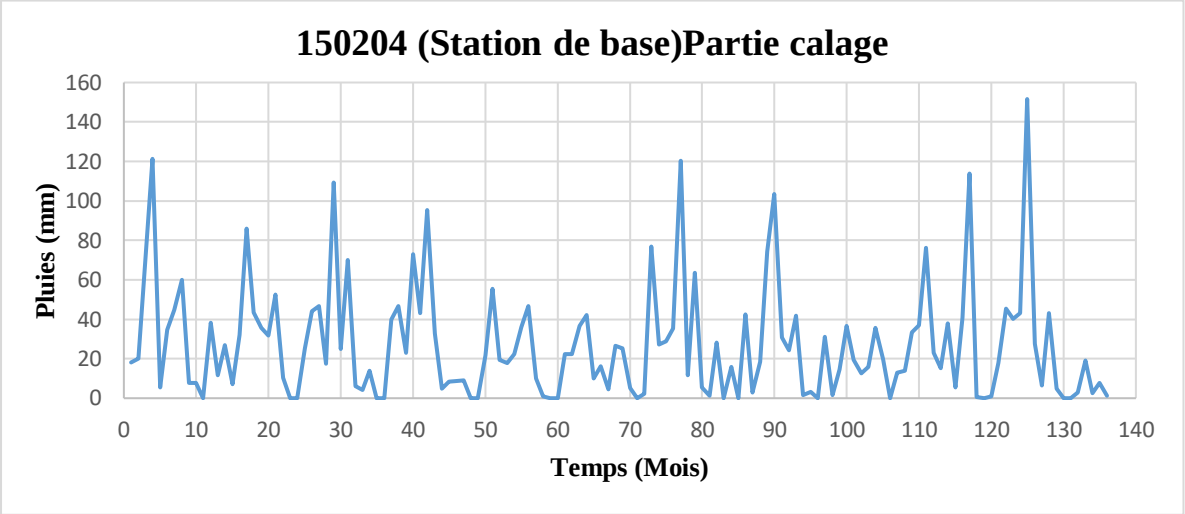
Wang, F., 2006. Quantitative Methods and Applications in GIS. CRC Press.

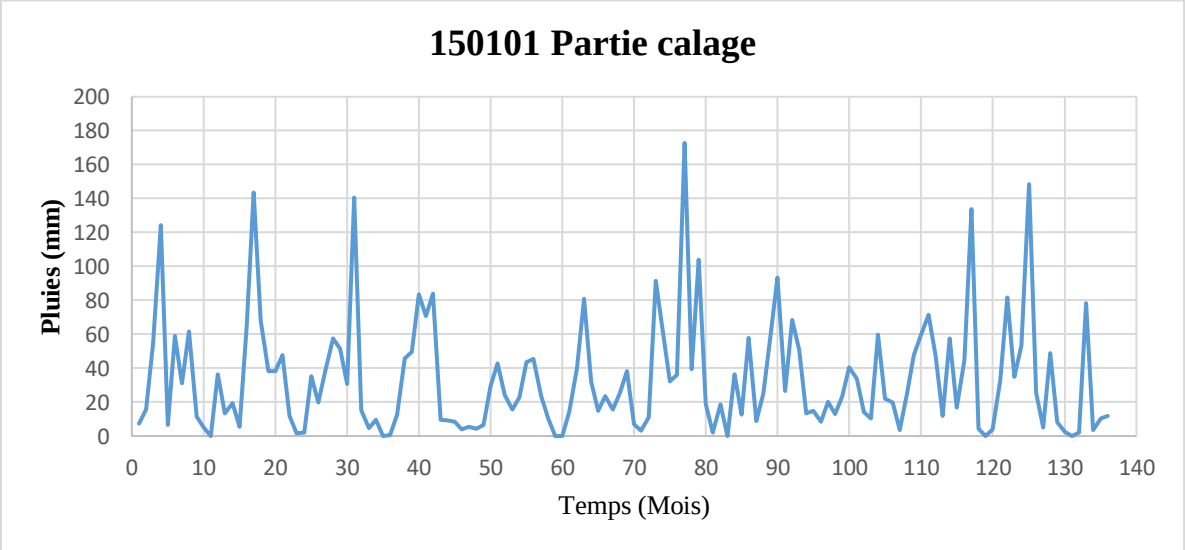
Wei, T.C., McGuinness, J.L., 1973. Reciprocal Distance Squared Method: A ComputerTechnique for Estimating Area Precipitation. Technical Report ARS-Nc8. USAgricultural Research Service, North Central Region, OH, USA.

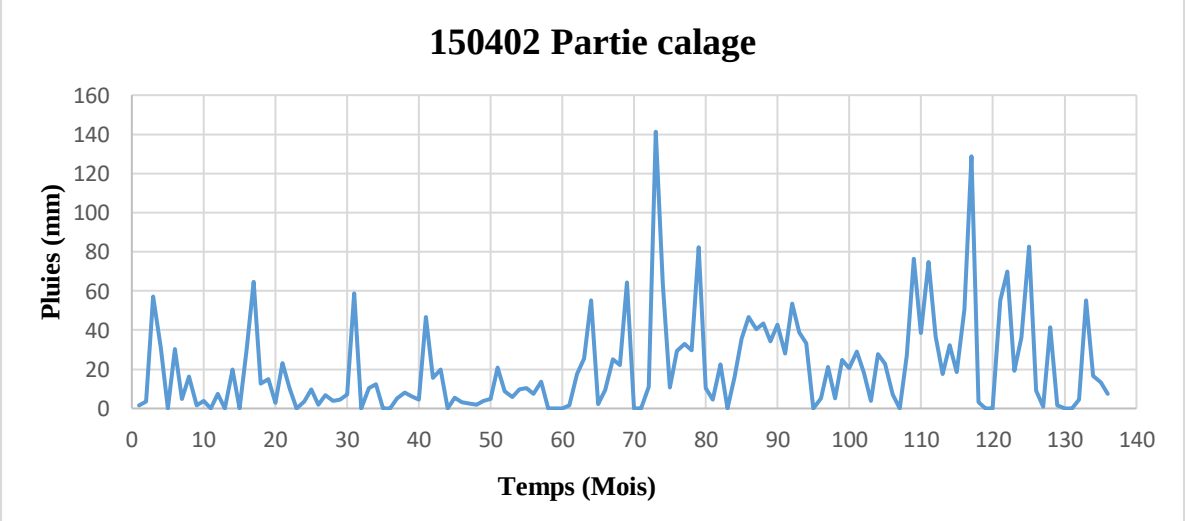
Zurada, J.M., 1992. Introduction to Artificial Neural Systems. Boston, MA, USA.www.sciencedirect.com.

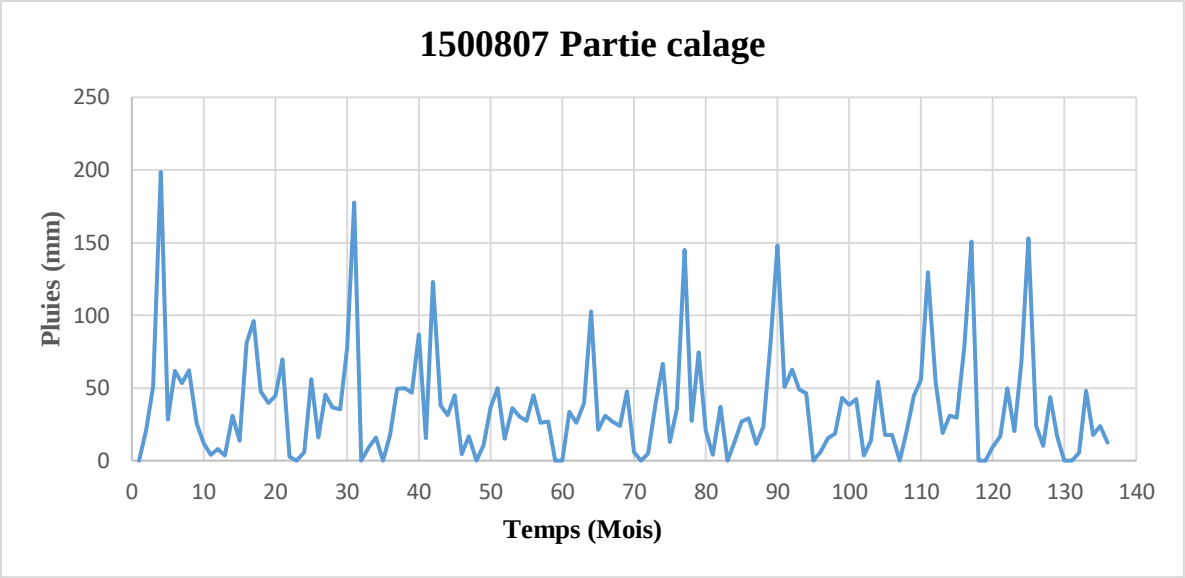
Annexe

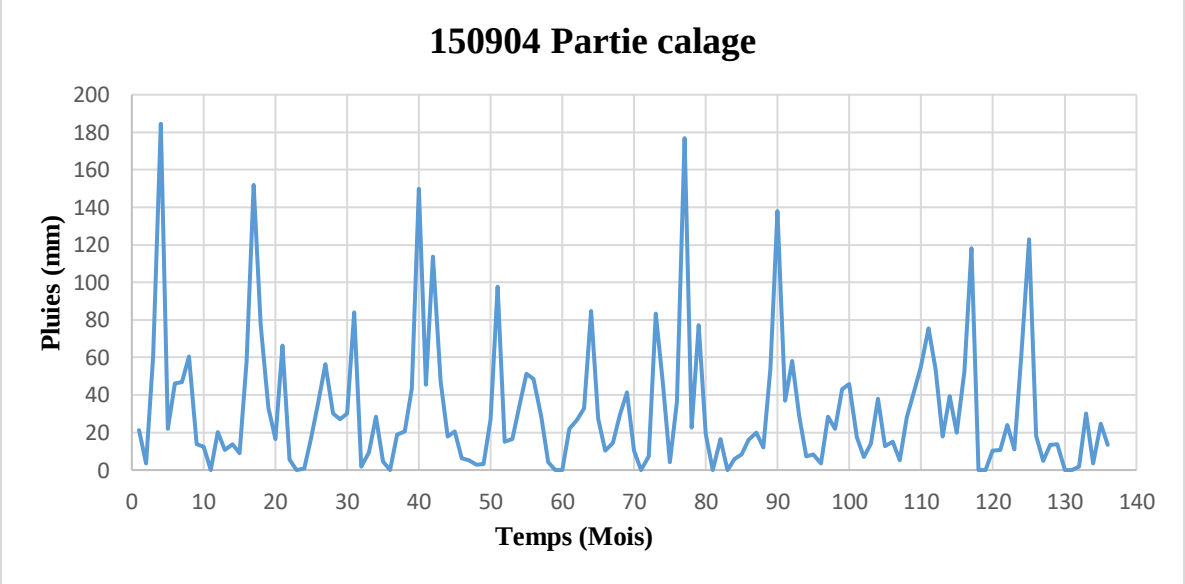
a) Données pluviométriques mensuelles utilisées pour le calage



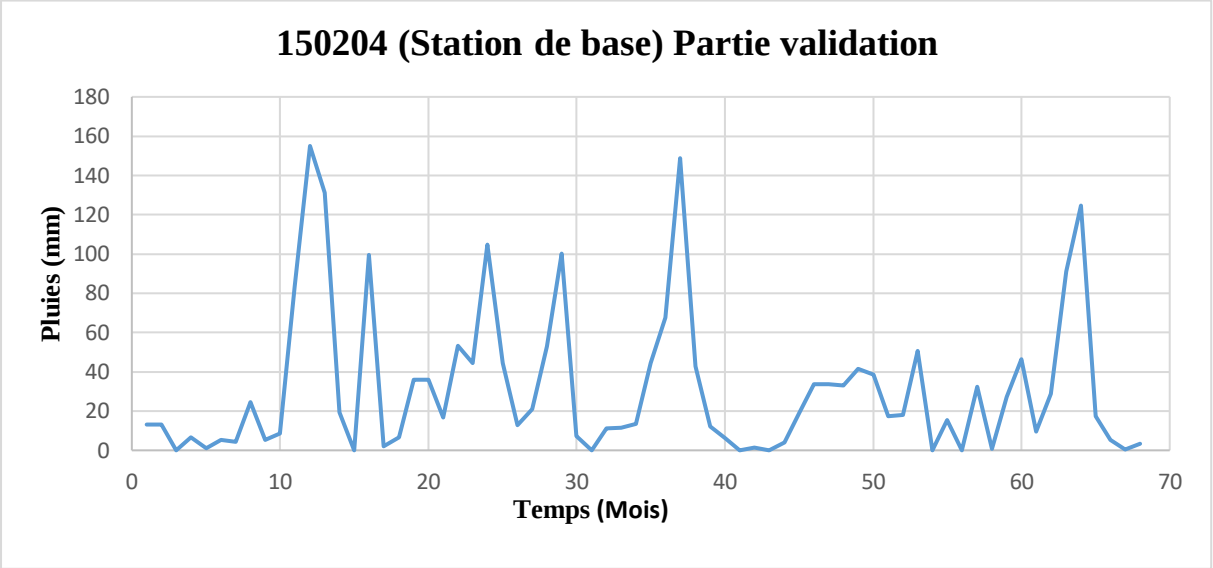


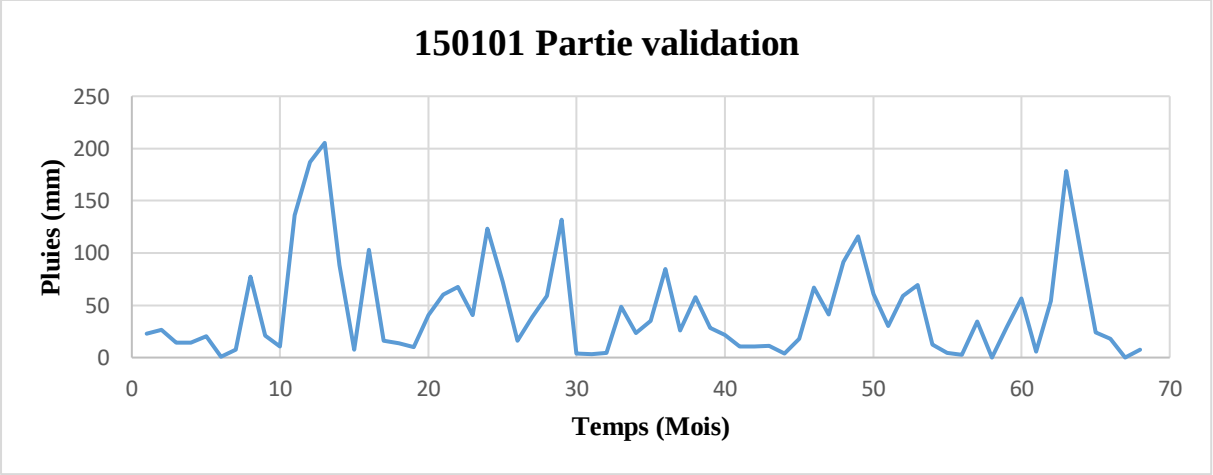


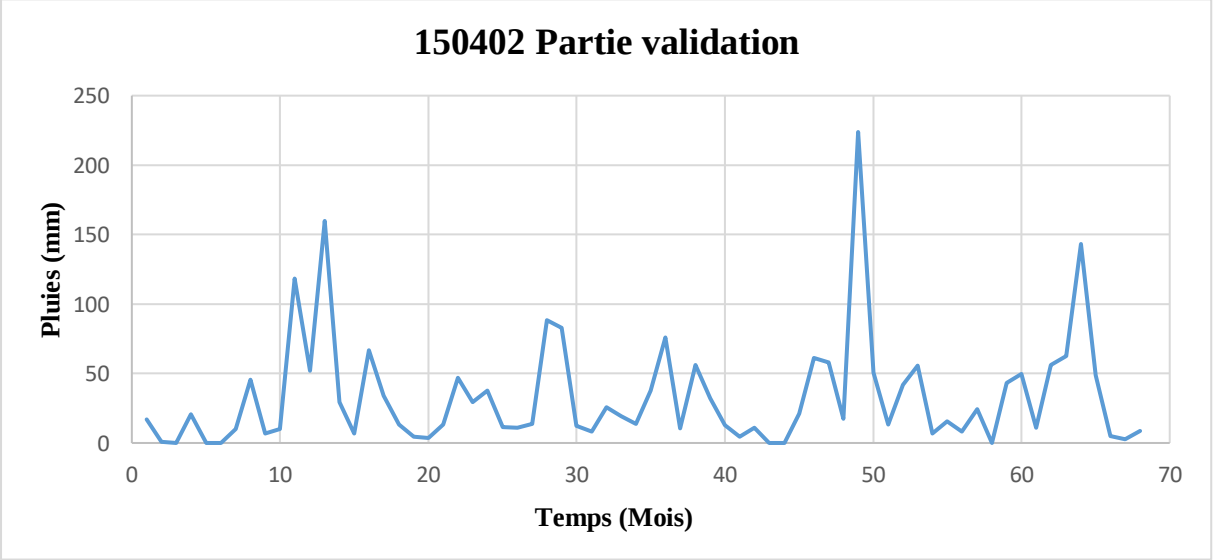

1500807 Partie calage
Pluies (mm)
250
200
150
100
50
0
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
Temps (Mois)

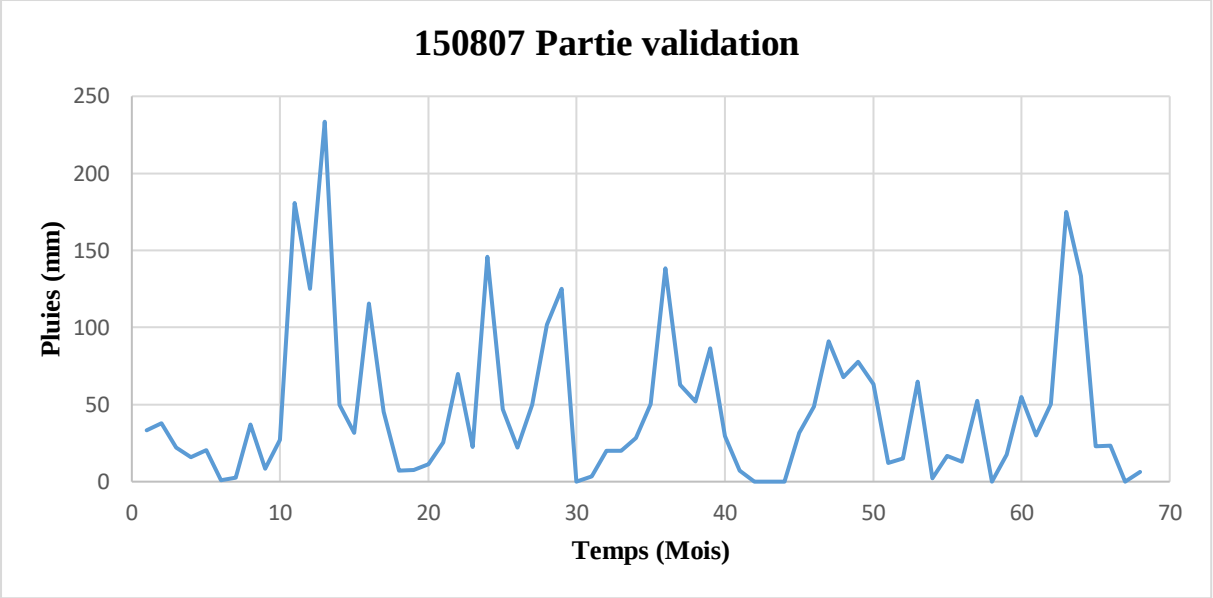

150904 Partie calage
Pluies (mm)
200
180
160
140
120
100
80
60
40
20
0
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
Temps (Mois)

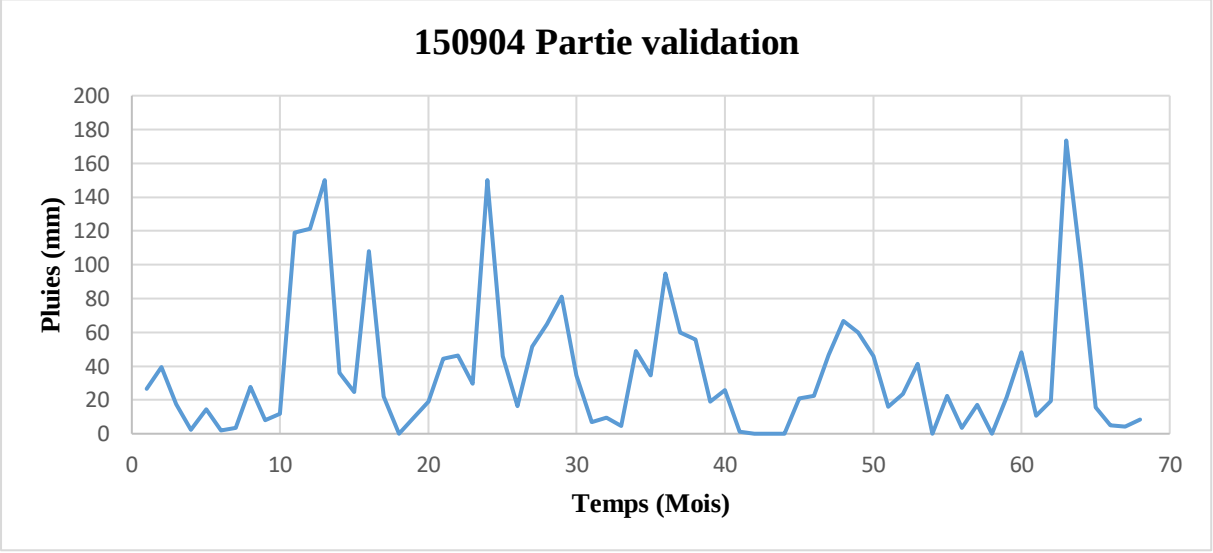
b) Données pluviométriques mensuelles pour la validation








150807 Partie validation
Pluies (mm)
250
200
150
100
50
0
0
10
20
30
40
50
60
70
Temps (Mois)


150904 Partie validation
Pluies (mm)
200
180
160
140
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
Temps (Mois)