REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE : Science Technologie

DEPARTEMENT : Hydraulique

DOMENE : : Science Technologie

FILIERE : Hydraulique

OPTION : Hydraulique Urbaine

Mémoire présenté pour l'obtention

Du diplôme de Master Académique

Intitulé

Etude de comparative des méthodes de comblement de lacune dans les enregistrements de précipitations

Présenté par :

- ✓ **Mlle Merrouche Ikhlass**
- ✓ **Mlle Boudjelel Dounia**

Encadré par: Mr. Djerbouai

2019/ 2020

Remerciement

Tout d'abord, nous tenons à remercier Dieu de nous avoir donné

Le courage, la force et la volonté pour terminer ce travail.

Tout notre respect et notre reconnaissance pour la source d'espoir et de motivation pour nos familles surtout les parents.

Tous nos remerciements et notre appréciation à notre respecté encadreur «Salim Djerbouai » pour son soutien et son intérêt tout au long de la période de travail, et nous le chérissons également pour sa moralité et son humilité.

Notre respect pour les membres du jury qui superviseront ce travail.

Nous tenons à présenter cette occasion pour montrer notre respect à tous enseignants et le personnel administratif de l'université de Mohamed Boudiaf, en particulier le département hydraulique.

A tous ceux qui nous ont aidées de près ou de loin, à tous ceux qui nous ont encadré, à tous ceux qui nous encouragé, à tous ceux qui nous ont accordé confiance, à tous ceux qui ont montré leur intérêt

à vous tous MERCI !!

Dédicace

Je dédie ce modeste travail à ceux qui m'ont tout donné sans rien demandé et à qui je dois

énormément et qui je ne remercierais jamais assez :

À ma famille qui a tout donné pour que je sois à ce niveau et qui m'a inculqué un esprit de

combativité et de persévérance et qui m'a toujours poussé et motivé dans mes études.

À ma meilleure amie ma sœur « Manal khalfallah » pour son aide, son temps, son encouragement, son assistance

et son soutien.

À tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

IKHLASS

dédicace

cadeau à ce que j'espère que let tout-puissant Mawla conservera dans les hauteurs des cieux, ma chère mère au soutien et au soutien indéfectibles pour moi.

Mon cher père, que dieu bénisse sa vie à mes chers frères et sœurs, à mes professeurs et compagnons de vie et à tous les collègues de spécialisation je dédie le fruit de cet humble effort, en espérant du tout-puissant, le tout-puissant il l'accepte de moi au maximum.

BOUDJELLeL DOUNIA.

Résumé

Dans la pratique, les enregistrements de précipitations sont liés au problème des données manquantes dû aux défauts de fonctionnement dans les stations pluviométriques.

L'évaluation des valeurs manquantes des précipitations a été un sujet difficile en hydrologie due à la variabilité spatiotemporelle des précipitations et de la complexité des processus physiques impliqués.

A cet égard nous avons mené une étude comparative entre les méthodes d'estimation des données manquantes dans les enregistrements des précipitations suivantes : Méthodes classiques ; CCWM, ACP, IDWM, Méthode basée sur les algorithmes génétiques : FFSGAM, dans le but de juger quelles sont les méthodes qui permettent de mieux estimer les données manquantes des précipitations.

L'application de ces méthodes a été faite en utilisant les données pluviométriques de cinq stations pluviométriques situées dans le bassin d'oued k'sob, portant le code (09) de l'A.N.R.H.

Ces méthodes ont été testé sur le pas de temps mensuel, en utilisant les critères de comparaison les plus recommandés.

A la fin, nous avons constaté que :

Toutes les méthodes d'estimation utilisées ont donné de bons résultats d'estimation.

La méthode ACP (analyse composante principale) ont donné des résultats plus performants que toutes les autres méthodes.

Mots clés : Données de précipitations manquantes, algorithmes génétiques, corrélations, méthodes de pondération, ACP.

ملخص

إن عملية تسجيل كميه تساقط الأمطار تواجه مشكلا أساسيا ألا و هو نقص أو ضياع بعض التسجيلات و ذلك راجع إلى خلل في عمل محطة التسجيل.

عملية تقدير القيم الضائعة كانت دائما مشكلا أساسيا في الهيدرولوجية نضرا إلى التباين الزماني و المكاني لظاهرة هطول الأمطار, وتعقيدها من الناحية الفيزيائية.

لقد قمنا في هذا العمل بإجراء مقارنة بين طرق تقدير المعلومات الضائعة حول تسجيل كمية الأمطار المتساقطة, التالية :الطرق الكلاسيكية ,CCWM, ACP,IDWM, وطريقة تعتمد على الخوارزميات الوراثية FFSGAM , بغرض معرفة أفضل في عملية تقدير كمية الأمطار غير المسجلة.

الدراسة أجريت باستخدام قيم الأمطار المسجلة في خمس محطات تسجيل واقعة في حوض واد قصب, الذي يحمل الرقم (09) حسب ترقيم الوكالة لوطنية للموارد المائية.

قمنا بإجراء عمليات تجريبية لكل الطرق على التسجيلات الشهرية و ذلك باستعمال أفضل معايير المقارنة و في الأخير استنتجنا أن:

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

الكلمات دالة :

قيم الأمطار المفقودة, الخوارزميات الوراثية, طرق الترجيح, الترابط, ت م أ.

Abstract

In the practice, the precipitations records are linked to the problem of missing data caused by fault in the rain gagging station.

In hydrology, estimating missing precipitation data is a crucial task due to the spatiotemporal variability of precipitations, also the complexity of physical processes involved.

We have done a comparative study between missing precipitations data estimation methods as next: classical methods: CCWM, ACP, Method based on genetic algorithms :FFSGAM, as a target to judge, which methods are better to assess missing precipitation data.

The application of these methods has been done using data of five rain gagging stations situated in the d'oued k'sob watershed, having the code (09) of A.N.R.H.

We have tested the methods using the most recommended criterions of comparison.

With the end we have noted that:

All the methods used, gave good results of estimate. And ACP model gave result more powerful than all the other methods.

Key words**:** Missing precipitation data, Genetic algorithms, Correlation, Weighing methods, ACP.

Table des matières

CHAPITRE III

Méthodes classiques d'estimation des données manquantes

CHAPITRE IV

Méthode basée sur l'intelligence artificielle

CHAPITRE V

Application sur des séries pluviométriques du bassin versant d'oued k'sob

CHAPITRE VI

Résultats et interprétations

Liste des tableaux

Liste des figures

Introduction générale

Introduction générale

Avec l'augmentation des besoins en eau liés à la poussée démographique, la croissance urbaine, les besoins industriels et agricoles, la gestion de l'eau est devenue une préoccupation majeure du pays.

Les données pluviométriques mesurées directement sur le terrain par les services météorologiques nationaux ont l'avantage de fournir souvent de longues séries d'observations indispensables à la détection des changements climatiques, mais elles présentent en contrepartie certaines limites comme par exemple souvent la présence de valeurs manquantes. Ces lacunes peuvent être 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 et/ou définitif de la mesure.

Les observations pluviométriques manquantes sont généralement estimées par des méthodes d'interpolation spatiale allant des techniques de pondération conceptuellement simples aux méthodes utilisant des techniques dépendant de la variance stochastique.

L'objectif attendu de ce travail est d'étudier certaines des méthodes d'estimation des données manquantes et de faire une comparaison pour trouver la meilleure méthode.

1-Méthodes classiques ;

2-Méthode basée sur l'intelligence artificielle (basée sur les algorithmes génétiques).

C'est donc pour répondre à ces objectifs, que ce mémoire a été structuré comme suit :

- 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 pour contrôlée de la qualité des données pluviométriques ;
- Le troisième chapitre présentée les méthodes classiques utilisées dans l'estimation des données de précipitations manquantes ;

- Le quatrième chapitre présente la méthode d'estimation des données manquantes basée sur l'intelligence artificielle FFSGAM ;
- Le cinquième chapitre est consacré à l'application des différentes méthodes d'estimations des données manquantes présentées dans les deux chapitres qui le précédent ;
- Les résultats et leurs interprétations sont présentés dans le sixième chapitre ;

En fin, ce travail est clôturé par une conclusion et des recommandations.

CHAPITRE I

Synthèse bibliographique

Synthèse bibliographique

Introduction

Le traitement des données pluviométriques avec observations manquantes est un problème concret et toujours embarrassant, 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.

1-Méthodes utilisées dans le comblement de lacune

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

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 ordinaire a 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.

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 modéles 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) and 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).

Il ya 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 algorithm method) 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 algorithm method), la méthode de pondération par la surface et 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.

CHAPITRE II

Contrôle de la qualité des données pluviométriques

Contrôle de la qualité des données pluviométriques

Introduction

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.

Les méthodes statistiques d'analyse de ces séries exigent de celles-ci une homogénéité de leurs composantes. En d'autres termes, on ne peut faire une analyse statistique d'un échantillon composé de n réalisations d'une variable climatique ou hydrologique, que si certaines de ces n réalisations ne présentent pas d'erreurs systématiques rendant l'échantillon hétérogène.

En particulier, les données pluviométriques sont généralement des relevés journaliers effectués par un pluviomètre. Cet appareil est relativement facile à mettre en place et à utiliser ; cela explique, sans le justifier, que l'on a souvent changé l'emplacement d'un pluviomètre, ou bien qu'on l'a confié successivement à de nombreux observateurs plus au moins qualifiés. Il en résulte que les séries de données présentent des lacunes fréquentes, et que l'on n'est jamais parfaitement sûr qu'elles présentent, comme disent les statisticiens, un échantillon d'une seule population.

Il est donc nécessaire, avant toute utilisation des variables pluviométriques, de contrôler leur qualité afin de réduire les erreurs systématiques qui pourraient les affecter. Des échantillons homogènes sortent de ce contrôle de qualité Laborde,(2003).

1- Erreurs possibles dans les mesures de précipitations (Sari, 2002)

Les erreurs qui peuvent avoir lieu dans les mesures de précipitations sont :

1.1- Erreurs d'observation

- Lecteur peu consciencieux : depuis celui qui lit le pluviomètre tous les cinq à six jours, jusqu'à celui qui invente purement et simplement les résultats en passant par celui , inconscient, qui arrose ses plantes avec l'eau du pluviomètre ;
- Erreurs fortuite de lecture de l'éprouvette;
- Erreurs dues à l'évaporation;
- Débordement du pluviomètre quant les pluies sont très intenses ;
- Pluviomètre percé ;

- Pertes d'eau pendant le transvasement de l'éprouvette dans le sceau;
- Pluviomètre sous un arbre, etc.

1.2- Erreurs systématiques

Parmi les erreurs systématiques, on peut citer:

- La graduation de l'éprouvette ne correspond pas à l'ouverture du pluviomètre ;
- Un changement dans l'exploitation du pluviomètre dû à ;
  - Un déplacement du pluviomètre ;
  - Modification de l'environnement pluviomètre ;
  - Un changement de l'observateur;
  - Une éprouvette cassée remplacée par une autre qui ne convient pas.

1.3 - Erreurs de transcription et de calcul

On peut rencontrer, s'il s'agit d'une copie, des erreurs supplémentaires :

Des chiffes peu lisibles ont pu être mal interprétés par le copieur, la virgule a pu être omise, l'ordre chronologique des feuilles mensuelles a pu être mal reproduit, etc. Egalement, on peut rencontrer des erreurs la sommation des relevés.

2 -Homogénéisation des données pluviométriques (Touaibia, 2004)

Qu'est ce que l'homogénéisation des données ? Pour répondre à cette question qui n'est pas aussi simple que l'on ne croit, il faut saisir et mesurer l'importance des dégâts que l'on peut avoir suite à une information fausse impliquée par un ingénieur pour dimensionner un ouvrage hydrotechnique. A ce moment là, il faut revenir et se poser la question : au fait à partir de quelle information de base suis-je parti pour faire mes calculs ? Est-elle fiable ? Tout le problème est là.

L'homogénéisation des données est une analyse statistique de l'information aidant à une prise de discision conséquente.

Elle consiste en:

- La détection des anomalies par des méthodes appropriées et d'en chercher la cause ;
- La correction des anomalies par des méthodes appropriées ;
- L'extension des séries hydrologiques courtes à partir à partir de séries de base homogènes, soit l'estimation d'une ou plusieurs observations d'un échantillon à partir d'autres observations prises dans des endroits différents.

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 :

3.1- Méthodes graphiques

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).

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).

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.

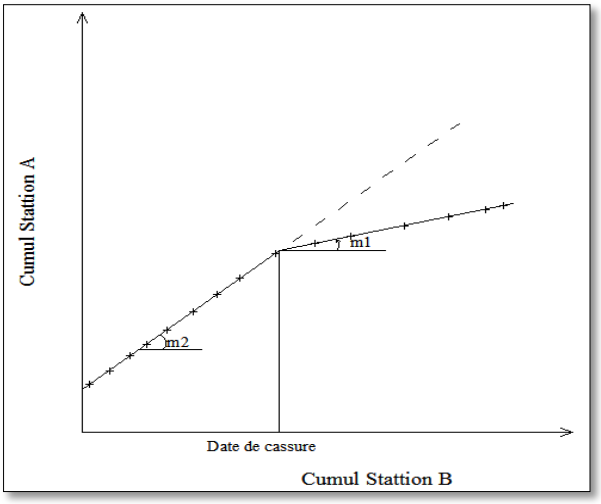


Figure II.1 : Méthode des doubles masses

- 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 (Touibia, 2004).

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.

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 :

$$Y_i = a_0 + a_1 x_i + \varepsilon_i \quad \text{ou encore} \quad \varepsilon_i = y_i - (a_0 + a_1 x_i) = y_i - \hat{y}_i \qquad \text{II.1}$$

De la théorie de la régression il découle que la somme des résidus est nulle et que leur distribution est normale, d'écart type :

$$\sigma_\varepsilon = \sigma_y \sqrt{1 - r^2} \qquad \text{II.2}$$

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:

$$E_0 = 0; \quad E_j = \sum_{i=1}^{j} \varepsilon_i \qquad \text{II.3}$$

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'inhomogénéités.

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 :

$$z_{1-\frac{\alpha}{2}} \sigma_\varepsilon \frac{n}{\sqrt{n-1}} \qquad \text{II.4}$$

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

3.2 - Méthodes numériques

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

Nous étudierons ici les tests suivants.

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.

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

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

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

$$W_{min} = \frac{(N_1+N_2+1)N_1 -1}{2} - u_{1-\alpha/2}\sqrt{\frac{N_1N_2(N_1+N_2+1)}{12}} \qquad \text{II.6}$$

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

$$w_{max} = (N_1 + N_2 + 1)N_1 - W_{min} \qquad \text{II.7}$$

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

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).

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_1 N_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 :

$$K = L - \frac{N_1(N_1+1)}{2} \qquad \text{II.8}$$

$$S = N_1N_2 - K \qquad \text{II.9}$$

Avec:

$L = \sum_{i=1}^{N_1} R(X_i)$: 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 :

$$\overline{K} = \overline{S} = \frac{N_1N_2}{2} \qquad \text{II.10}$$

$$S_k = S_s = \frac{N_1N_2}{2}(N_1 + N_2 + 1) \qquad \text{II.11}$$

- 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 :

$$T = \left|\frac{K-\overline{K}}{S_K}\right| \qquad \text{II.12}$$

Avec la variable centré réduite ayant une probabilité au dépassement α/2.

Si $T < Z_{\alpha/2}$, on accepte H_0

3.2.3 - Test de la médiane (ou test de Mood)

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

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 :

$$N_s > \frac{1}{2}(N+1-u_{1-\frac{\alpha}{2}}\sqrt{N+1}) \qquad \text{II.13}$$

$$T_s < 3.3(\log_{10}(N)+1) \qquad \text{I1.14}$$

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

Méthodes classiques d'estimation

des données manquantes

Introduction

Afin de ne pas perdre la continuité de l'information dans les enregistrements de données concernant les précipitations, il est préférable d'estimer les données manquantes au moyen de l'information obtenue dans les stations voisines.

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 nous voulons 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 :

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

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 X.

2- Méthode IDWM (Inverse distance weightingméthod)

La méthode de pondération par la distance inverse (reciprocal-distance), (Wei and McGuinness, 1973) est très utilisée pour l'estimation des données de précipitations manquantes. Cette méthode estime la valeur d'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.

3-Méthode CCWM (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 .

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 : Observation dans la station i ;

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

4 - Méthode basée sur la corrélation

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 théorique n'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).

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 n'entreprend une corrélation que lorsque la dépendance entre les variables peut être expliquée.

- **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.

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 $Y=A+BX+CX^2$

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.

L'expérience montre que sous tout climat à pluviosité abondante et (ou) peu variable (climat tempéré, équatorial, tropical, entre autres), la hauteur annuelle de précipitations est une variable normale et indépendante et que la régression entre deux variable est linéaire.

4.5-Régression linéaire simple

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 :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{III.6}$$

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).

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 :

$$\hat{y} = b_0 + b_1 x_i \qquad \text{III.7}$$

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 $X=x_i$.

$ei= y_i- \hat{y} = y_i- b_0 - b_1x_i,$ pour $i=1,\ldots.n.$

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

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

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:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i y_i)+n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = r\frac{S_x}{S_y} \qquad \text{III.8}$$

$$b_0 = \bar{y} - b_1\bar{x} \qquad \text{III.9}$$

Où :

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

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 .

Soient :$\bar{x}_k$; $\bar{y}_k$; ${}_k\sigma_x$; ${}_k\sigma_y$ 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:

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

$$k < j \leq n$$

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 étendu par :

$$\hat{\sigma}_{y\,n}^2 = {}_k\sigma_y^2 + r_k^2 \frac{{}^k\sigma_y^2}{{}^k\sigma_x^2}\left({}_n\sigma_x^2 - {}_k\sigma_x^2\right) \qquad \text{III.12}$$

4.5.4 - Moyen d'appréciation du gain obtenu par l'extension (Sari,2002)

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:

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

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.

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 plurimensuelles, 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 (III.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 plurimensuelles de la station Y chacune desquelles composées de kvaleurs 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 plurimensuelles, 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.

4.6 -Régression double (Laborde,2003)

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 :

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

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

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

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

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

On a alors, tout calcul fait :

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

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 écartsε_i. On peut donc construire de même un coefficient de corrélation multiple R qui mesurera la dispersion des ε_i:

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

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

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

$$R^2 = \frac{r_1^2 + r_2^2 - 2pr_1r_2}{1-p^2}$$

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

$$R^2 = r_1^2 + r_2^2 \qquad \text{III.21}$$

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

$$\sigma_{z_{xy}} = \sigma_{\varepsilon i} = \sigma_z\sqrt{1-R^2} \qquad \text{III.22}$$

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 :

$$r_{zx_y}{}^2 = \frac{R^2 - r_2^2}{1 - r_2^2} \qquad \text{III.23}$$

$$r_{zy_x}{}^2 = \frac{R^2 - r_1^2}{1 - r_1^2} \qquad \text{III.24}$$

4.7-Régressions linéaires multiples (Laborde, 2003)

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] :

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

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

$$\hat{y}_{xi} = a_0 + a_1 x_1 + \cdots + a_i x_i + \cdots a_k x_k \qquad \text{III.25}$$

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 $r_{yxi_{x1,x2,x3\ldots..xk}}$

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 : $\frac{\partial \varepsilon^2}{a_0} = -2\sum(y - a_0 - a_1x_1 \dots \dots . - a_kx_k) = 0$

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

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

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 :

$$a_i = \frac{\sigma_y}{\sigma_{xi}} \frac{\Delta_{0i}}{\Delta_{00}} \text{ (k fois)} \qquad \text{III.26}$$

$$a_0 = \bar{y} - \sum_{i=1}^{k} \bar{x}_i \qquad \text{III.27}$$

Enfin le coefficient de corrélation multiple R est donné par :

$$R^2 = 1 - \frac{\Delta}{\Delta_{00}} \qquad \text{III.28}$$

4.7.3 -Seuils de signification

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

$$F = \frac{n-(k+1)}{K} \frac{R^2}{1-R^2} \qquad \text{III.29}$$

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.

4.8-Régression non linéaire

Vue sa complexité, dans de nombreux cas on pourra se sortir d'affaire en linéarisant 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 : $y = ax^b \rightarrow \ln y = \ln a + b \ln x$.

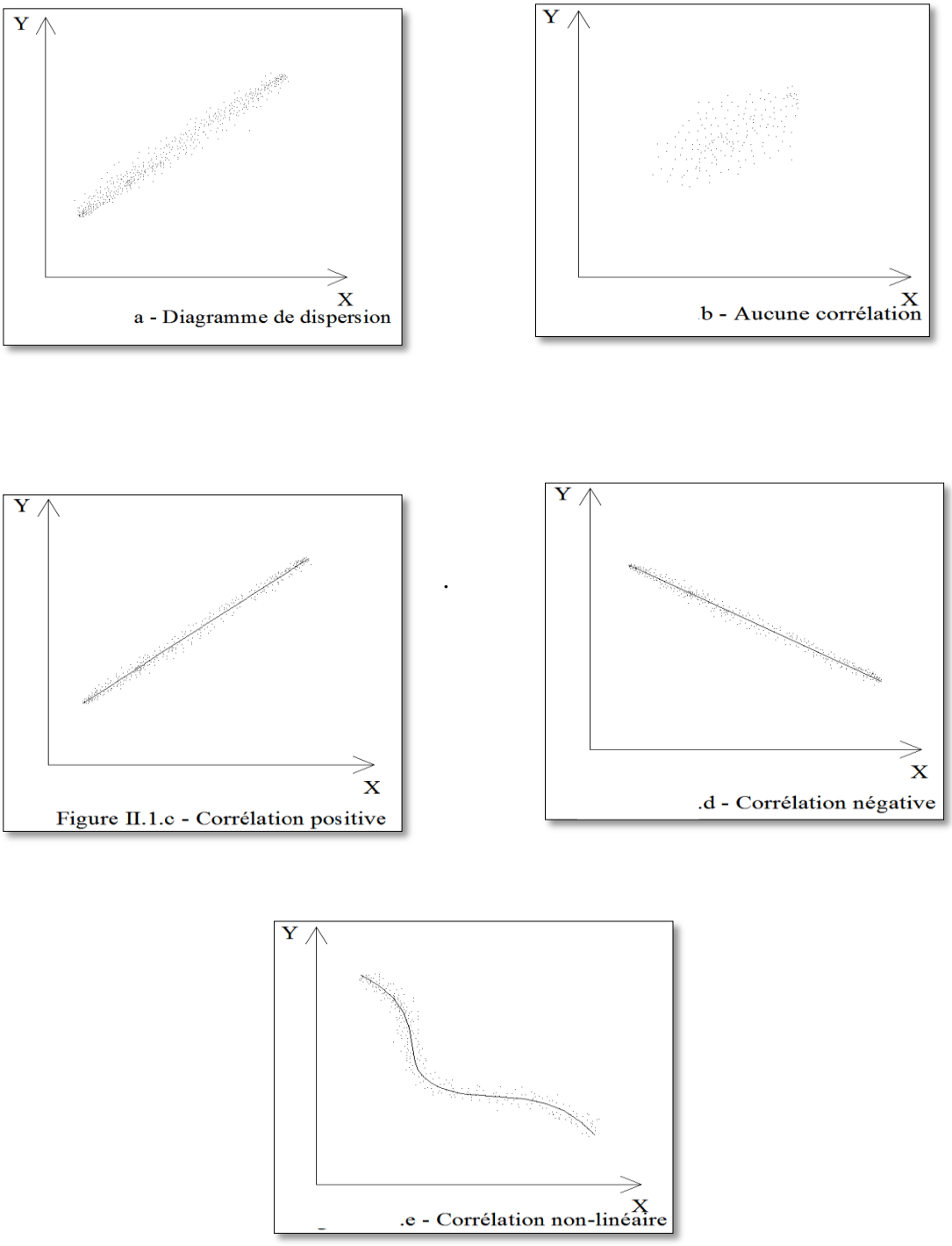


Figure III.1 : schémas explicatifs

5-Méthode d'analyse en composante principale

L'Analyse en Composantes principales (ACP) fait partie du groupe des méthodes descriptives multidimensionnelles appelées méthodes factorielles. Ces méthodes qui sont apparues au début des années 30 ont été surtout développées en France dans les années 60, en particulier par Jean-Paul Benzécri qui a beaucoup exploité les aspects géométriques et les représentations graphiques. Dans la mesure où ce sont des méthodes descriptives, elles ne s'appuient pas sur un modèle probabiliste, mais elles dépendent d'un modèle géométrique.

L'ACP propose, à partir d'un tableau rectangulaire de données comportant les valeurs de p variables quantitatives pour n unités (appelées aussi individus), des représentations géométriques de ces unités et de ces variables. Ces données peuvent être issues d'une procédure d'échantillonnage ou bien de l'observation d'une population toute entière. Les représentations des unités permettent de voir s'il existe une structure, non connue a priori, sur cet ensemble d'unités. De façon analogue, les représentations des variables permettent d'étudier les structures de liaisons linéaires sur l'ensemble des variables considérées. Ainsi, on cherchera si l'on peut distinguer des groupes dans l'ensemble des unités en regardant quelles sont les unités qui se ressemblent, celles qui se distinguent des autres, etc. Pour les variables, on cherchera quelles sont celles qui sont très corrélées entre elles, celles qui, au contraire ne sont pas corrélées aux autres, etc (Duby,2006)

5.1- Procédé de calcul d'une ACP (Laborde,2003)

Soit une série de NECH observations sur NVAR variables. Ces variables peuvent être plus ou moins liées entre elles, et selon l'intensité de leur liaison, on peut réduire la taille de cet ensemble d'informations en se contentant d'un nombre inférieur à NVAR de variables qui permettent cependant de conserver la quasi-totalité de la variance de l'ensemble.

Soit X la matrice de départ :

$$X = \begin{vmatrix} X_{11} & X_{12} & X_{1NECH} \\ X_{21} & & \\ & Xij & \\ X_{NVAR1} & & X_{NVAR\ NECH} \end{vmatrix}$$

On construit la matrice $\dot{X}$ des variables centrées réduites. Chaque élément $\dot{x}_{ij}$ se déduit de x_{ij} par :

$$\dot{x}_{ij} = \frac{x_i - \bar{x}_i}{\sigma_{xi}} \qquad \text{III.30}$$

Où :

$$\bar{x}_i = \frac{1}{NECH}\sum x_{ij} \text{ et } \sigma_{xi}{}^2 = \frac{1}{NECH}\sum (x_{ij} - \bar{x}_i)^2$$

On cherche alors s'il existe un changement de base M qui transforme la matrice X en une matrice Y dont les composantes soient indépendantes. Ceci se traduit par les équations suivantes :

$yy^t = M\dot{X}(M\dot{X})^t = M\dot{X}\dot{X}^tM^t = D$

$y = M\dot{X}$ Changement de base

$yy^t = D$ (Matrice diagonale, indépendance des composantes)

on peut montrer que la matrice de changement de base M existe et est la matrice modale de $\dot{X}\dot{X}^t$.

Chaque ligne de cette matrice représente les vecteurs propres normés de$\dot{X}\dot{X}^t$;$\dot{X}\dot{X}^t$ est la matrice des covariances qui représentes à $\frac{1}{NECH}$ prés de la matrice des coefficients de corrélations totales.

Quant à D est la matrice spectrale $\dot{X}\dot{X}^t$, c'est-à-dire la matrice diagonale dont les termes sont les valeurs propres de $\dot{X}\dot{X}^t$.

$$M = \begin{vmatrix} a_{11} & a_{12} & a_{1NECH} \\ a_{21} & & \\ & aij & \\ a_{NVAR1} & & a_{NVAR\ NECH} \end{vmatrix}$$

$$\begin{vmatrix} \lambda_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

$$\begin{vmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_{\text{NVAR NECH}} \end{vmatrix}$$

D=

$y = M\dot{X}$ d'où $M^{-1}y = \dot{X} = M^t y$ car $M^{-1} = M^t$

$$\dot{X}\dot{X}^t = M^t y(M^{-1}y)^t = M^t y y^t M = M^t D M$$

Cette dernière formule permet d'évaluer d'une autre façon la variance des $\dot{x}_i$ égale à 1 dans notre cas où les variables sont centrées réduites.

$$1 = {\sigma_{xi}}^2 = \frac{1}{\text{NCEH}} \sum_{j=1}^{\text{NVAR}} a_{ij}^t \lambda_j a_{ji} = \frac{1}{\text{NCEH}} \sum_{j=1}^{\text{NVAR}} a_{ij}^2 \lambda_j$$

La somme des variances des NVAR variables est égale à NVAR:

$$\text{NVAR} = \sum_{i=1}^{\text{NVAR}} \frac{1}{\text{NCEH}} \sum_{j=1}^{\text{NVAR}} a_{ij}^2 \lambda_j = \frac{1}{\text{NCEH}} \sum_{j=1}^{\text{NVAR}} \lambda_j \sum_{j=1}^{\text{NVAR}} a_{ij}^2$$

Or, les vecteurs propres étant normés, on doit avoir :

$\sum_{j=1}^{\text{NVAR}} a_{ij}^2 = 1$ et donc $\text{NVAR} = \frac{1}{\text{NCEH}} \sum_{j=1}^{\text{NVAR}} \lambda_j$

Dans le cas particulier fréquent où au lieu de travailler sur la matrice des covariances $\dot{X}\dot{X}^t$ on travaille sur la matrice R des coefficients de corrélation, on a :

$\dot{X}\dot{X}^t = \text{NECH}.\,R$ d'où $\text{NECH} = \sum_{j=1}^{\text{NVAR}} \lambda_j$

Comme on vient de le constater, la somme des valeurs propres est proportionnelle à la variance totale des observations.

Si on range les valeurs propres par ordre décroissant : λ_1 ,λ_2, λ_3 ... λ_{NVAR}, on constate que les valeurs propres deviennent très vite très petites. Il paraît donc justifié de ne retenir qu'un nombre ń inférieur à NVAR, de vecteur propre qui expliqueront un pourcentage p de la variance totale :

$$P = \frac{\sum_{i=1}^{\text{NVAR }\grave{}}}{\sum_{i=1}^{\acute{n}}} \lambda_1 * 100 \text{ (en \%)} \qquad \text{III.31}$$

Nous pourrons donc condenser l'information contenue dans le tableau de dimension NVAR* NECH, par l'information contenue dans le tableau Y' de dimension n' NECH. Y est le tableau correspondant aux n' premières lignes de Y.

Dans le cas où on travaille sur la matrice des coefficients de corrélation:

- chaque ligne i de Y correspond aux valeurs prises par une des composantes ; ces valeurs ont une moyenne nulle et une variance égale à l ;
- chaque terme aij de la matrice M de changement de base est tel que :

$$pc_j x_i = \sqrt{\lambda_j}\, a_{ji} \qquad \text{III.32}$$

$pc_j x_i$ Etant le coefficient de corrélation totale entre les valeurs prises pour la $j^{ème}$ composante et la $i^{ème}$ variable.

En conclusion, on pourra retenir que l'analyse en composante principale permet de réduire les observations sur NVAR variables liées entre elles en une série d'observations sur n' < NVAR variables indépendantes, chacune d'entre elles expliquant une part décroissante de la variance totale initiale.

5.2-Comblement des lacunes par L'ACP

Soient nv postes pluviométriques, durant une période d'observation de no années les nv postes pluviométriques n'ont pas tous été observés. La matrice des données se présente donc généralement ainsi :

$$[X]= \begin{vmatrix} x(1,1) & inconnu & \dots & x(1,j) & x(1,nv) \\ x(2,1) & x(2,2) & \dots & x(2,j) & inconnu \\ \dots & & \dots & \dots & \dots \\ x(i,1) & x(i,2) & \dots & x(i,j) & x(i,nv) \\ inconnu & & \dots & inconnu & \dots \\ \dots & inconnu & \dots & \dots & \dots \\ x(no,1) & x(no,2) & \dots & x(no,j) & x(no,nv) \end{vmatrix}$$

Si le nombre de poste nv est grand, il est quasiment impossible de "boucher" les trous un par un par des techniques de régression. Le choix des variables explicatives est vite inextricable et de plus les erreurs accidentelles ou systématiques qui ne manquent pas

d'affecter certaines observations, sont ainsi répétées. Nous proposons donc une méthode qui permet de combler rapidement et simplement les lacunes avec la partie la plus fiable de ce tableau de données.

Nous calculerons tout d'abord les moyennes expérimentales Mxo(j) et les écart-types expérimentaux σxo(j) des précipitations pour chaque station j et sur les seules années réellement observées :

$$[Mxo] = \begin{vmatrix} Mxo(1) & Mxo(2) & \dots & Mxo(j) & \dots & Mxo(nv) \end{vmatrix}$$

$$[\sigma xo] = \begin{vmatrix} \sigma xo(1) & \sigma xo(2) & \dots & \sigma xo(j) & \dots & \sigma xo(nv) \end{vmatrix}$$

On peut alors évaluer la matrice des valeurs centrées réduites :

$$u(i,j) = \frac{x(i,j) - Mxo(j)}{\sigma xo} \qquad \text{III.33}$$

et calculer pour chaque observation la moyenne des valeurs centrées réduites disponibles $\bar{u}(i)$:

$$[u] = \begin{vmatrix} u(1,1) & \text{inconnu} & \dots & u(1,j) & u(1,nv) & \bar{u}(1) \\ u(2,1) & u(2,2) & \dots & u(2,j) & \text{inconnu} & \bar{u}(2) \\ \dots & & \dots & \dots & \dots & \dots \\ u(i,1) & u(i,2) & \dots & u(i,j) & u(i,nv) & \bar{u}(i) \\ \text{inconnu} & & \dots & \text{inconnu} & \dots & \dots \\ \dots & \text{inconnu} & \dots & \dots & \dots & \dots \\ u(n0,1) & u(n0,2) & \dots & u(n0,j) & u(n0,nv) & \bar{u}(no) \end{vmatrix}$$

On peut alors dans une première étape remplacer chaque valeur inconnue x(i,j) par une valeur :

$$xe_0(i,j) = \frac{\bar{u}(i) - Mxo(j)}{\sigma xo} \qquad \text{III.34}$$

correspondant à la variable réduite moyenne pour cette observation. On obtient alors une première matrice complète [xo] :

$$[xo] = \begin{vmatrix} x(1,1) & xe_0(1,2) & \dots & x(1,j) & x(1,nv) \\ x(2,1) & x(2,2) & \dots & x(2,j) & xe_0(2,nv) \\ \dots & & \dots & \dots & \dots \\ x(i,1) & x(i,2) & \dots & x(i,j) & x(i,nv) \end{vmatrix}$$

inconnu		…	$xe_0(i+1,j)$	…
…	$xe_0(no-1,2)$	…	…	…
x(no,1)	x(no,2)	…	x(no,j)	x(no,nv)

Nous effectuons alors une A.C.P. sur cette matrice [xo] et obtenons les projections des variables [a_1] et des observations [c_1] sur les k seules premières composantes principales pouvant avoir une signification physique :

$$\mathbf{[a_1]=}\begin{vmatrix} a_1(1,1) & a_1(1,2) & \dots & a_1(1,j) & \dots & a_1(1,nv) \\ a_1(2,1) & a_1(2,2) & \dots & a_1(2,j) & \dots & a_1(2,nv) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_1(k,1) & a_1(k,2) & \dots & a_1(k,j) & \dots & a_1(k,nv) \end{vmatrix}$$

$$\mathbf{[c_1]=}\begin{vmatrix} c_1(1,1) & c_1(1,2) & \dots & c_1(1,k) \\ c_1(2,1) & c_1(2,2) & \dots & c_1(2,k) \\ \dots & \dots & \dots & \dots \\ c_1(i,1) & c_1(i,2) & \dots & c_1(i,k) \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ c_1(no,1) & c_1(no,2) & \dots & c_1(no,k) \end{vmatrix}$$

Ainsi que les matrices des moyennes et écarts-types :

$$[Mx_1]=\begin{vmatrix} Mx_1(1) & Mx_1(2) & \dots & Mx_1(j) & \dots & Mx_1(nv) \end{vmatrix}$$

$$[Sx_1]=\begin{vmatrix} Sx_1(1) & Sx_1(2) & \dots & Sx_1(j) & \dots & Sx_1(nv) \end{vmatrix}$$

Il est alors possible de reconstituer chaque observation manquante x(i,j) par une nouvelle valeur estimée $xe_1(i,j)$:

$$xe_1(i,j) = Mx_1(j) + Sx_1(j) * \{ a_1(1,j)\, c_1(i,1) + a_1(2,j)\, c_1(i,2) + \dots + a_1(k,j)\, c_1(i,k) \} \quad \text{III.35}$$

Cette estimation n'est pas très correcte puisque l'A.C.P. a été effectuée sur une matrice "bouchée" à partir de moyennes interannuelles, cependant $xe_1(i,j)$ est une meilleure estimation

que $xe_0(i, j)$ puisqu'elle tient compte des observations aux autres stations pour cette année j. On peut donc réitérer le processus en remplaçant dans la matrice [xo]
chaque $xe_0(i, j)$ par les $xe_1(i,j)$ adaptés. On obtient ainsi une nouvelle matrice [x1].

$$[x1]=\begin{vmatrix} x(1,1) & xe_1(1,2) & \dots & x(1,j) & x(1,nv) \\ x(2,1) & x(2,2) & \dots & x(2,j) & xe_1(2,nv) \\ \dots & \dots & \dots & \dots & \dots \\ x(i,1) & x(i,2) & \dots & x(i,j) & x(i,nv) \\ inconnu & \dots & \dots & xe_1(i+1,j) & \dots \\ \dots & xe_1(no-1,2) & \dots & \dots & \dots \\ x(no,1) & x(no,2) & \dots & x(no,j) & x(no,nv) \end{vmatrix}$$

Onrecommence alors une A.C.P. sur la matrice [x1] permettant d'obtenir de nouvelles matrices $[c_2]$, $[a_2]$,[Mx2] et [Sx2] d'où l'on tirera de nouvelles estimations $xe_2(i,j)$.
A chaque itération nous modifions les estimations pour les observations manquantes.

Généralement, pour les données pluviométriques on travaille avec 10 à 15 itérations, et 3 à 5 composantes.

Conclusion

Dans le présent chapitre nous avons présenté les différentes méthodes classiques utilisées dans l'estimation des données de précipitations manquantes, dans le chapitre qui suit nous allons voir l'une des méthodes basées sur l'intelligence artificielle.

CHAPITRE IV

Méthode basée sur l'intelligence artificielle

Méthode basée sur l'intelligence artificielle

Introduction

Depuis plus de cinquante ans, les chercheurs s'inspirent de la biologie dans le but de réaliser des structures capables de résoudre divers problèmes complexes**.**

Récemment, plusieurs méthodes basées sur des principes de biologie ont été développées pour estimer les données manquantes dans les enregistrements précipitations, tel que les réseaux de neurones artificiels et les algorithmes génétiques.

Dans ce travail nous nous intéressons à la méthode FFSGAM (Fixedfunction set GeniticAlgorithmMethod) , elle est récrément proposée par Tufail et Ormsbee (2006).

1-Aperçu sur les algorithmes génétiques (AG)

Les algorithmes génétiques (AG), ont été initialement développés par John Holland (1975) ses collègues et ses étudiants, à l'université du Michigan dans deux buts principaux (Beasleyet et Martin1993):

1- Mettre en évidence et expliquer rigoureusement les processus d'adaptation des systèmes naturels.

2- Concevoir des systèmes artificiels qui possèdent les propriétés des systèmes naturels.

Leurs champs d'application sont très vastes. Outre l'économie, ils sont utilisés pour l'optimisation des fonctions numériques difficiles , traitement d'image (alignement de photos satellites, reconnaissance des suspects…), optimisation d'emplois du temps, optimisation de design, contrôle des systèmes industriels, apprentissage des réseaux de neurones etc.

La raison de ce grand nombre d'application est claire c'est la simplicité de leurs mécanismes, la facilité de leurs mise en application et leur efficacité même pour des problèmes complexes.

Les (AG) peuvent être utilisés pour contrôler un système évoluant dans le temps (chaîne de production, centrale nucléaire…) car la population peut s'adapter à des conditions changeantes (Nabonne, 2004).

2-Présentation de la méthode FFSGAM (Teegavarapu et al ., 2009)

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.

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 III.1 :

$$y = a_1 F(x_1) \ @ \ b_1 F(x_2) \qquad \text{IV.1}$$

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 la fonction 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 :

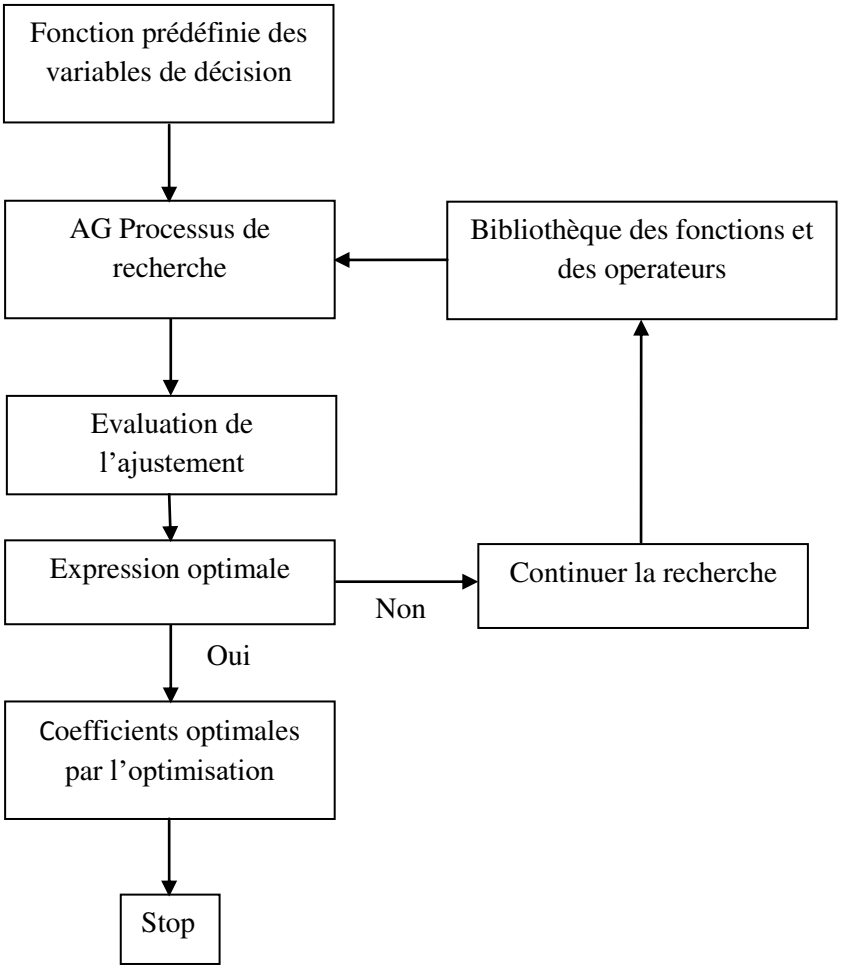


Figure IV.1: Organigramme du modèle FFSGAM

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 d_{mi} entre chaque station i et celle de base m (dont on veut compléter l'information) ;
- Le coefficient de corrélation R_{mi} 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:

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

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 locaux (pour chaque station), ces coefficients sont estimés par la minimisation de l'erreur moyenne quadratique, par conséquent l'équation d'estimation devient :

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

2.3 - Evaluation des coefficients optimaux

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

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

Où :

$\hat{P}_i, P_i$: 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:

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

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

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

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

Où

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

d_{mi} :distance entre la station i et celle de base m.

Conclusion

Parmi toutes les méthodes de pondération que nous avons présentées, seule la méthode FFSGAM utilise un facteur de pondération qui combine à la fois entre deux paramètres (la distance et le coefficient de pondération).

CHAPITRE V

Application sur des séries pluviométriques du bassin versant d'oued k'sob

Application sur des séries pluviométriques du bassin versant d'oued k'sob

Introduction

L'objectif de ce travail est d'étudier les méthodes d'estimation des données manquantes à pas de temps mensuel par les méthodes suivantes:

1-Méthodes classiques : CCWM, IDWM et l'ACP ;

2-Méthode basée sur l'intelligence artificielle (basée sur les algorithmes génétiques): FFSGAM (FFSGAM^1, FFSGAM^2, FFSGAM^3, FFSGAM^4).

FFSGAM^1, FFSGAM^2, FFSGAM^3, FFSGAM^4, sont les quatre fonctions d'estimations obtenues par les algorithmes génétiques présentées dans le chapitre précédent respectivement par les fonctions : IV.6-9.

Pour pouvoir juger ces méthodes, une étude comparative entre les résultats fournis par chacune de ces dernières est nécessaire. Pour ce faire nous avons consacré cette partie du mémoire pour tester ces méthodes au pas de temps mensuel puis les résultats obtenus vont être présentés et interprétés dans le chapitre qui suit.

1-Méthodologie

L'application des méthodes d'estimations a été faite dans le bassin versant d'oued k'sob 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 Hodna, qui porte le code (09).

1.1-présentation de la région d'étude

1.1.1-Situation géographique

Le bassin versant de l'oued K'sob est situé dans le grand bassin du Hodna au Nord de l'Algérie. Il est limité au Nord-Ouest par la chaine montagneuse des Bibans ; au Sud et au Sud-ouest par les monts du Hodna et à l'Est par les hautes plaines de Sétif. Il se situe aussi entre les méridiens de longitudes 5°6‟ et 4°34‟ Est et les parallèles de latitude 35°33'et 36°18'

Nord. Il s'étend sur la totalité de Bou Arreridj dont son exutoire est à la limite Nord de la wilaya de M'sila.

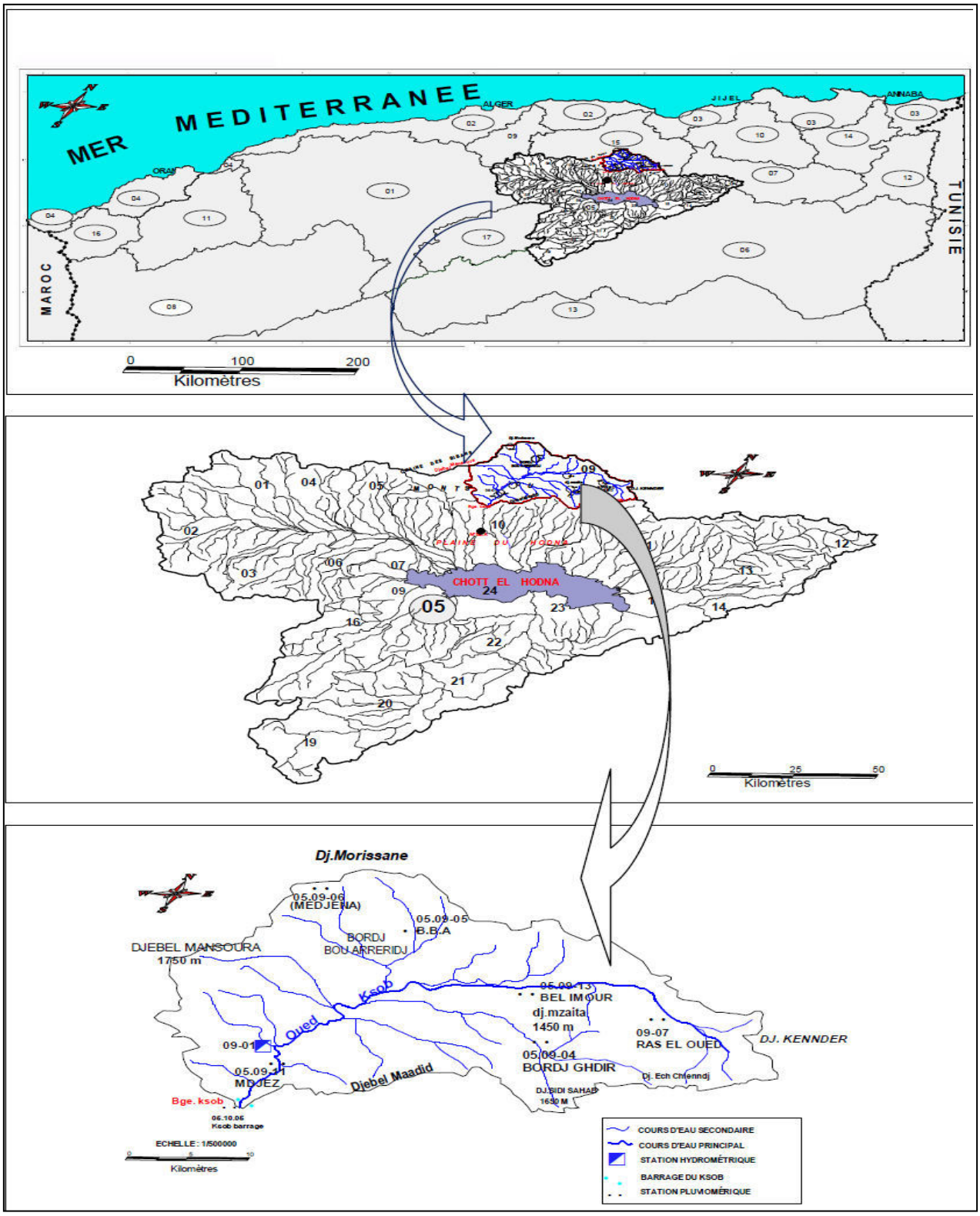


Figure V. 1 : Localisation et situation du BV d'oued K'sob

1.1.2- Situation climatique

a) Le climat

D'une manière générale, le climat de la région est de type semi-aride caractérisé par un été sec et chaud et un hiver froid.

b) La température

La température moyenne annuelle est égale à 15.16 °C. Les mois les plus chauds sont les mois d'été soit " juin ; juillet ; août et septembre "durant lesquelles les températures moyennes dépassent les 20°C. La saison froide pendant laquelle les températures sont inférieures à 10°C se prolonge de décembre à février.

c) Pluviométrie

Le paramètre climatique hydrologique essentiel dans le fonctionnement d'un bassin versant, dans la région , le régime des écoulements est fortement lié aux précipitations, Leur intensité, leur continuité et leur périodicité et même sont origine d'écoulement, de sa localisation et de sa violence. La tranche d'eau reçue varie de 500 à 600 mm environ sur les hautes plateaux hodnéenne.

La plaine, est d'une altitude voisine de 950 m à la moyenne, avec une pente très faible ne dépassant pas 1%. Elle est aménagée dans des formations essentiellement quaternaires.

1.2- Donnés pluviométriques utilisées

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

Tableau V.1 : caractéristiques des cinq stations pluviométriques

Nom de la station	Code (A.N.R.H)	Période de fonctionnement
Medjez	05 09 01	1974- 2010
Bordj Ghedir	05 09 04	1974- 2010
Bordj Bouarreridj	05 09 05	1974- 2010
Madjana	05 09 06	1974- 2010
Barrage K'sob	05 10 05	1974- 2010

Sur toute la période de fonctionnement des 5 stations, il ya 21 années d'observations commues qu'on va utiliser dans le pèsent travail.

Les 5 stations utilisées sont schématisées sur un extrait de la carte du réseau hydro-climatologique.

Figure V.2 : 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 Medjaz(station de référence).
- Station de Bordj Ghedir .
- Station de Bordj Bourreridj.
- Station de Medjana.
- Station de barrage k'sob.

1.3-Répartition des données

Toutes les méthodes sont utilisées pour estimer les données de la station de base ***m*** de Medjaz, 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% (16 années), la deuxième partie pour la validation avec 30% (5 années).

- La partie calage est constituée des données mensuelles des 15 années suivantes : 74-75 à 77-78 ; 80-83 ; 98-2003 ;.
- La partie validation constituée des données mensuelles des 6 années suivantes : 2005-2010.

Les données utilisées dans le calage et la validation sont présentés dans l'annexe (a ; b).

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.

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

- Racine de l'erreur moyenne quadratique (RMSE);
- Erreur moyenne absolue (MAE);
- Le coefficient de détermination R^2.

Les expressions des différents critères sont données par les équations V.1-4 :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{P}_i - P_i)^2} \quad V.1$$

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(\widehat{P}_i - P_i)}{P_i}\right| \quad V.2$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|(\widehat{P}_i - P_i)\right| \quad V.3$$

$$R^2 = \frac{[\sum_{i=1}^{n}((\widehat{P}_i - \overline{\widehat{p}})(p_i - \overline{p}))]^2}{\sum_{i=1}^{n}(\widehat{P}_i - \overline{\widehat{p}})^2 \sum_{i=1}^{n}(p_i - \overline{p})^2} \quad V.4$$

Où $\widehat{P}_i$: Valeur estimée ;

P_i : Valeur observée ;

n : Nombre total d'observation.

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 (generalized reduced gradient).

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

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.

1.6-Application des méthodes d'estimations

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

1.6.1-Méthode IDWM

a) Evaluation des facteurs de pondération d_{mi}

Pour IDWM les distances d_{mi} , sont mesurées sur la carte du réseau hydro-climatologique 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 Medjaz) et les autres stations *i* sont portées dans le tableau V.2.

Tableau V.2: Facteur de pondération d_{mi}

Station i	**50904**	**50905**	**50906**	**1005**
d_{mi}(m)	25034.09	20265.55	21388.65	8684.71

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.3 : Coefficients de corrélation entre la station de base et les autres stations i

Station pluviométrique i	**50904**	**50905**	**50906**	**1005**
Coef.de corrélation	0.80	0.77	0.59	0.90

1.6.3-Méthode d'ACP

L'application de la méthode de comblement des lacunes par l'Analyse en

Composantes Principales (AC P) a été faite à l'aide du logiciel HYDROLAB version 98.2.

1.6.3.1-Présentation du logiciel HYDROLAB

Logiciel l'HYDROLAB est un Programme développé par J.P. Laborde Professeur à l'Université de Nice - Sophia Antipolis-France en 1998. C'est un outil très simple parfaitement intégré, écrit en basic et présenté sous forme de macros sur EXCEL dont l'utilisation est universelle.

L'objectif de HYBROLAB est de répondre aux questions les plus fréquemment posées aux Hydrologues. Ces questions portent essentiellement sur :

- L'analyse uni variée (Ajustements) ;
- L'analyse multi variée (Régressions multiples) ;

- L'analyse en composantes principales (ACP) ;
- Le comblement de lacunes dans des séries de données ;
- La détection d'anomalies dans les séries de données ;
- L'analyse spatiale (variographie).
- Des fonctions statistiques classiques telles que F de Fisher-Snedecor, intégrale de Gauss… ;
- Des fonctions liées à l'estimation de l'évapotranspiration potentielle (Penmann, Durée du jour, radiation, Mc culloch, FAO,…) ;
- Une fonction pour passer de l'évapotranspiration potentielle à la réelle;
- Des fonctions de passage des coordonnées géographiques à différentes coordonnées Lambert.

1.6.4-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 Medjaz).

Tableau V.4 : Coefficients optimaux C_i

Ci / FFSGAM	C_1	C_2	C_3	C_4
FFSGAM[1]	0,0933	1,0208	-0,2221	3,9563
FFSGAM[2]	0,0789	1,0121	-0,5711	3,91515
FFSGAM[3]	0,0068	0,0832	-0,0001	16,4185
FFSGAM[4]	0,4313	0,1395	-1,7279	3,2835

Conclusion

Les différentes étapes de l'application des méthodes de cambrement de lacune sont présentées dans ce chapitre, Les Résultats obtenus dans cette partie du mémoire vont être traités et analysés dans le chapitre suivant.

CHAPITRE VI

Résultats et interprétations

Résultats et interprétations

Introduction

Dans ce chapitre nous allons présenter les résultats d'estimation et leurs interprétations, afin de juger quelle sont les méthodes qui donnent de bons résultats d'estimations et qui seront recommandées pour le comblement des données manquantes de précipitations à pas de temps mensuel.

1-Estimation des données manquantes

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

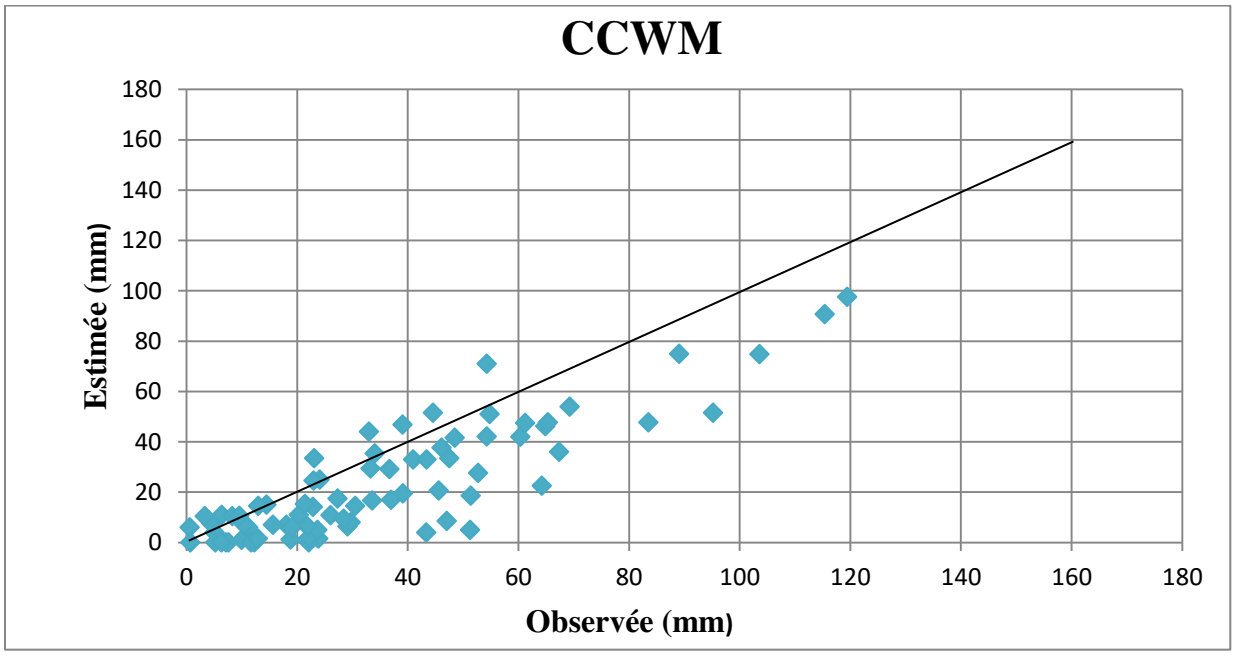
Tableau VI.1 : Critère de comparaison

Méthode / Critère	CCWM	IDWM	ACP	FFSGAM[1]	FFSGAM[2]	FFSGAM[3]	FFSGAM[4]
RMSE	**17.91**	**21.21**	**12,47**	**11.34**	**11.34**	**11.34**	**11.34**
MAE	**14.18**	**16.67**	**10,13**	**8.06**	**8.06**	**8.06**	**8.06**
R^2	**0.76**	**0.70**	**0.88**	**0.82**	**0.82**	**0.82**	**0.82**

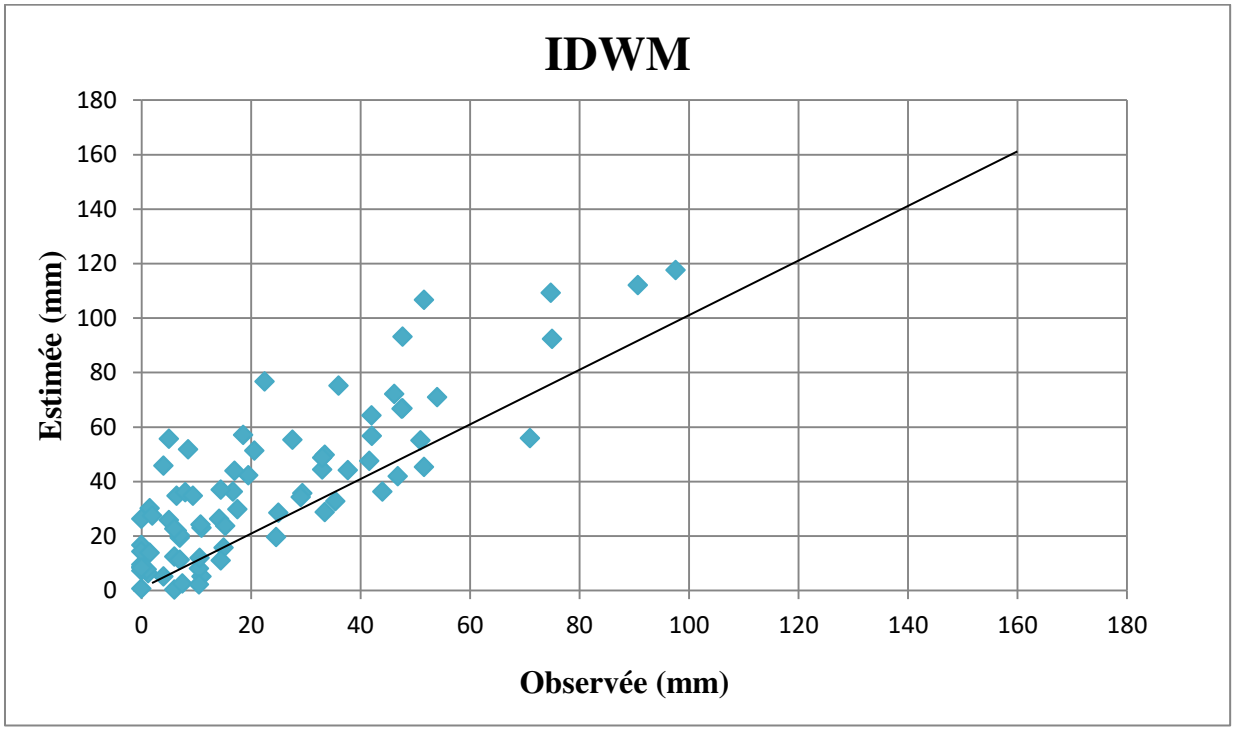
b-L'analyse des critères présentés par tableau VI.1 nous permet de tirer que :

- Toutes les méthodes utilisées ont donné de très bons résultats d'estimations
  La méthode d'ACP a donné des résultats plus performants que toutes les autres méthodes.
- Les quatre modèles de FFSGAM convergent vers les mêmes valeurs.

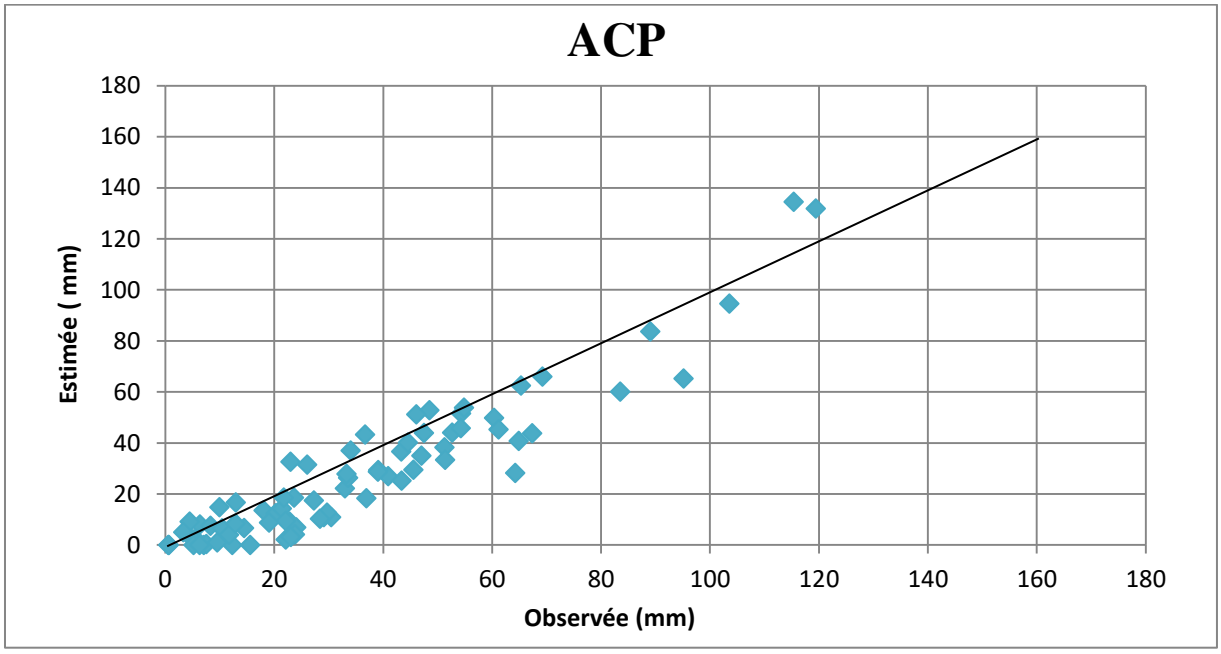
Figure VI.1 : montre les graphes des valeurs estimées en fonctions de celles observées



a : Méthode CCWM



b : Méthode IDWM



c : Méthode ACP

2-Interprétations sur les graphes

Il apparait dans la figure VI.1 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 FFSGAM s'alignent sur la première bissectrice mieux que les méthodes CCWM et IDWM.
- Le nuage des points estimés par la méthode ACP s'alignent sur la première bissectrice mieux que toutes les autres méthodes.

3-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) C_i, on se propose de tester les quatre modèles de la FFSGAM en prenant ces coefficients égaux à l'unité.

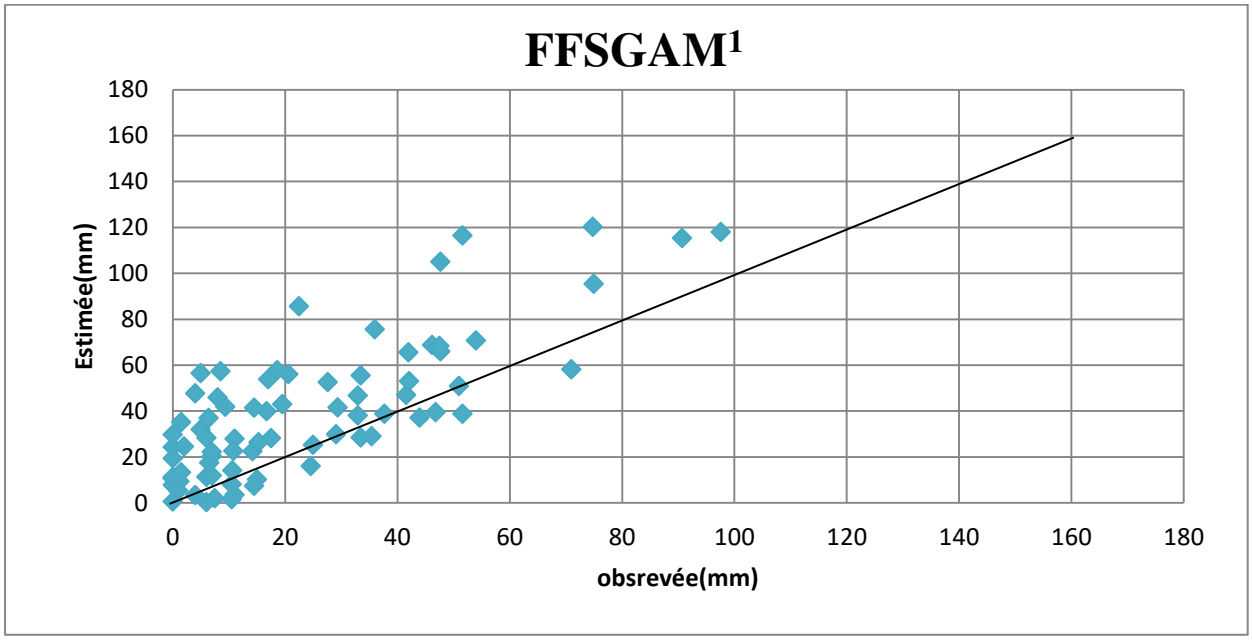
3.1 Estimation avec C_i=1 :

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

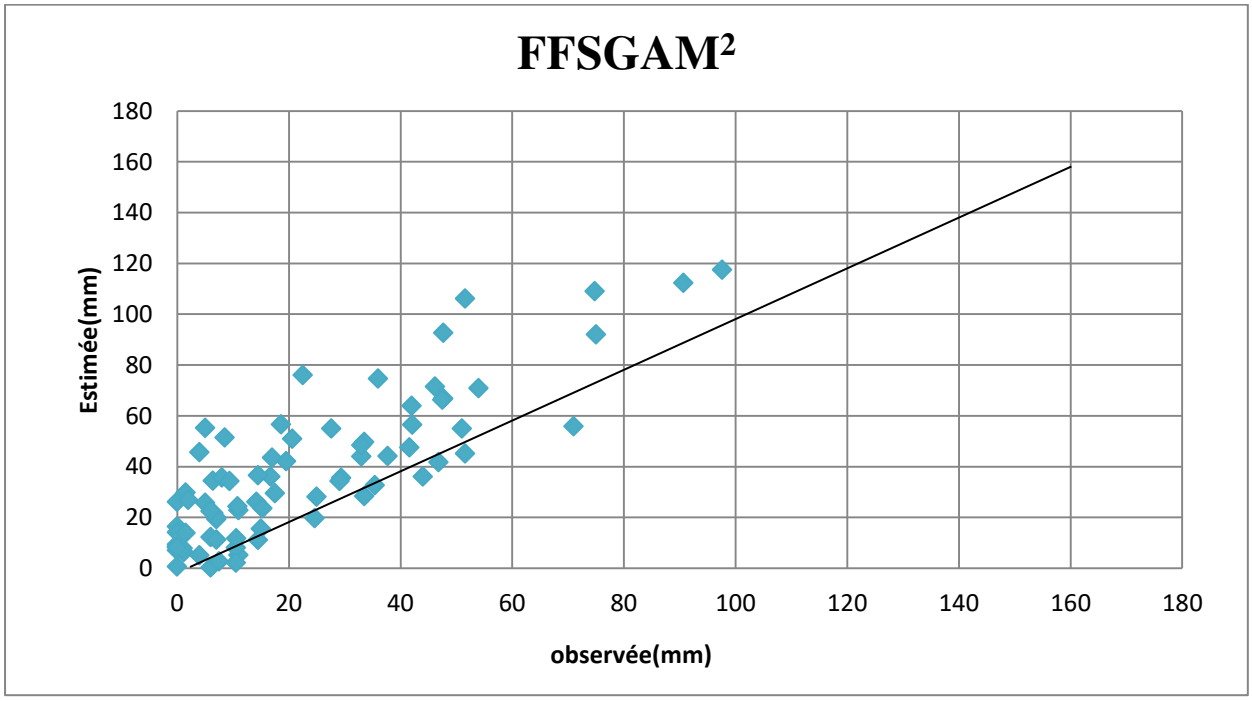
Tableau VI.2 :Critère de comparaison (Estimation C_i=1)

FFSGAM / Critère	FFSGAM^1	FFSGAM^2	FFSGAM^3	FFSGAM^4
RMSE	23.92	21	31.10	19.66
MAE	18.46	16.51	23.59	14.71
R^2	0.62	0.70	0.46	0.67

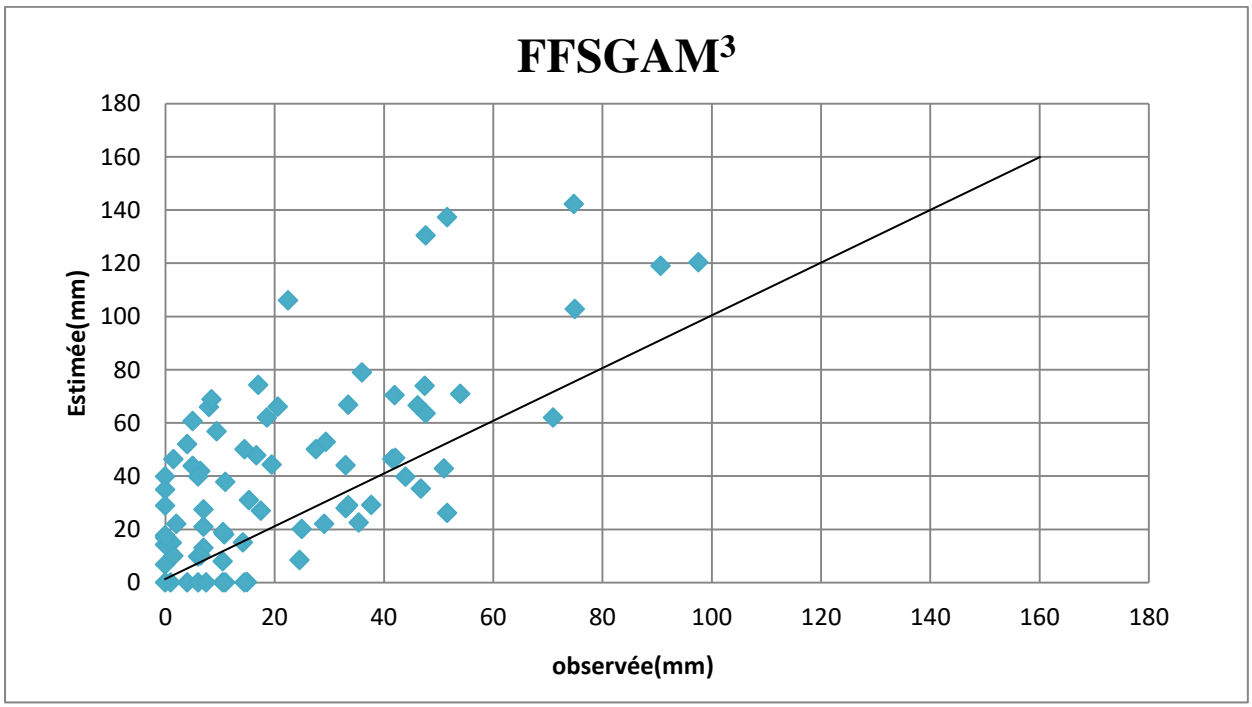
Le tableau VI.2 : montre que même avec des coefficients locaux C_i égaux à l'unité tous les modèles de la FFSGAM ont donné des résultats acceptables



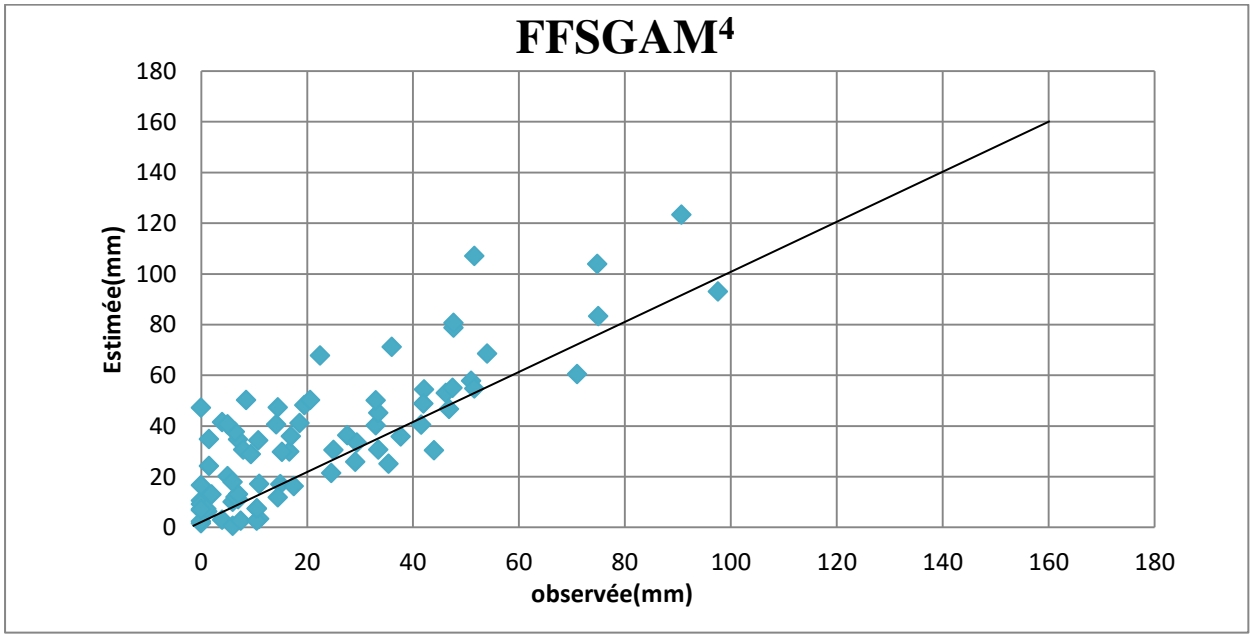
a : méthode FFGSAM^1



b : méthode FFGSAM²



c : méthode FFGSAM³



d: méthode FFGSAM^4

La Figure VI.2 : valeurs estimées en fonction de celles observées (Estimation avec $C_{i=}1$).

Conclusion

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

La méthode ACP a donné des résultats plus performants que toutes les autres méthodes, cela en peut être justifié par le fait que cette méthode calculée par A.C.P. sur la matrice des coefficients de corrélation des stations de base constitue les vecteurs régionaux. Ces derniers représentent des stations virtuelles sans lacunes et sans erreurs, utilisées par la suite pour la critique des données et le comblement des lacunes annuelles des postes pluviométriques.

Conclusion générale

Conclusion générale

Dans différents domaines utilisant les données hydrométéorologiques, nous sommes souvent en présence du problème des données manquantes. La question qui se pose est : qu'elle est la meilleure méthode qui permet d'estimer le mieux les données hydrométéorologiques manquantes.

En Algérie, généralement les services hydrologiques reconstituent les observations de précipitations manquantes par les méthodes simples (moyenne, moyennes pondérées par la tendance annuelle), l'analyse en composantes principales ou par l'analyse de régression.

Cependant dans d'autres pays tel qu'aux Etats unis, plusieurs méthodes sont utilisées pour reconstituer les observations pluviométriques manquantes, telles que les méthodes de pondérations classiques (méthodes de pondérations par la distance inverse, par le coefficient de corrélation) , et plus récemment, les méthodes basées sur l'intelligence artificielle (telles que les méthodes basées sur les algorithmes génétiques, sur les réseau de neurones artificiels) ont été introduites.

A cet égard nous avons mené une étude comparative entre les différentes méthodes d'estimation des données manquantes dans les enregistrements de précipitation. L'étude a été faite au pas de temps mensuel en utilisant les méthodes suivantes :

1-Méthodes classiques : CCWM, IDWM et l'ACP.

2-Méthode basée sur l'intelligence artificielle (basée sur les algorithmes génétiques): FFSGAM.

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

La méthode ACP a donné des résultats plus performants 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 Hydrologic Engineering, 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 partial weather station network. Agricultural Forest Meteorology 84, 255–271.

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

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 knowledgebased approach to the statistical mapping of climate. Climate Research 22, 99–113.

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

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.

Giustolisi, O., Savic, D.A., 2004. A novel genetic programming strategy: evolutionary polynomial regression. In: Liong, S.-Y., Phoon, K.-K., Babovic (Eds.), Proc. of the 6th International 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 by evolutionary 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 Academic Publishers, 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. - "Eléments d'hydrologie de surface" - Cours polycopié de l'Université de Nice Sophia Antipolis - 1998 - Nice - 195 pages.

LABORDE J. P. - "Notice explicative de la carte des évapotranspirations potentielles de l'Algérie du Nord" - Projet de Coopération algéro-allemande n°94 21 83 5 - 1997 – GTZ ANRH Alger - 41 pages.

LABORDE.J.P.1998. NOTICE D'UTILISATION DU LOGICIEL HYDROLAB. C.N.R.S.France .

Salas, J.D.-J., 1993. Analysis and modeling of hydrological time series. In: Maidment, D.R. (Ed.), Handbook of Hydrology. Mc-Graw-Hill, New York.

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

Seo, D.-J., Krajewski, W.F., Bowles, D.S., 1990b. Stochastic interpolation of rainfall data from rain gages and radar using cokriging—2. Results. Water Resources Research 26 (5) 915–924.

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

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

Sullivan, D.O., Unwin, David J., 2003. Geographical Information Analysis. Wiley, New York.

Touaibia Iman.(2011) Département hydraulique faculté de technologie .université de Telmcen BP.230-1300 Algérie. ESTIMATION BIAIS DU MODELE REGRISSIF PUISSANCE CAS DU BASSIN VERSANT DU K'SOB.

RYANJ. LONGMAN , ANDREWJ. NEWMAN, THOMASW. GIAMBELLUCA, MATHEWLUCAS . (JULY2020) Characterizing the Uncertainty and Assessing the Value of Gap-Filled Daily Rainfall Data in Hawaii.

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. Journal of Chemical Information and Computer Sciences 38, 189–199.

Simanton, J.R., Osborn, H.B., 1980. Reciprocal-distance estimate of point rainfall. Journal of Hydraulic 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.

Teegavarapu, R.S.V., 2007. Use of universal function approximation in variancedependent interpolation 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. Journal of Hydroinformatics 11 (2), 133–146.

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

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

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

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

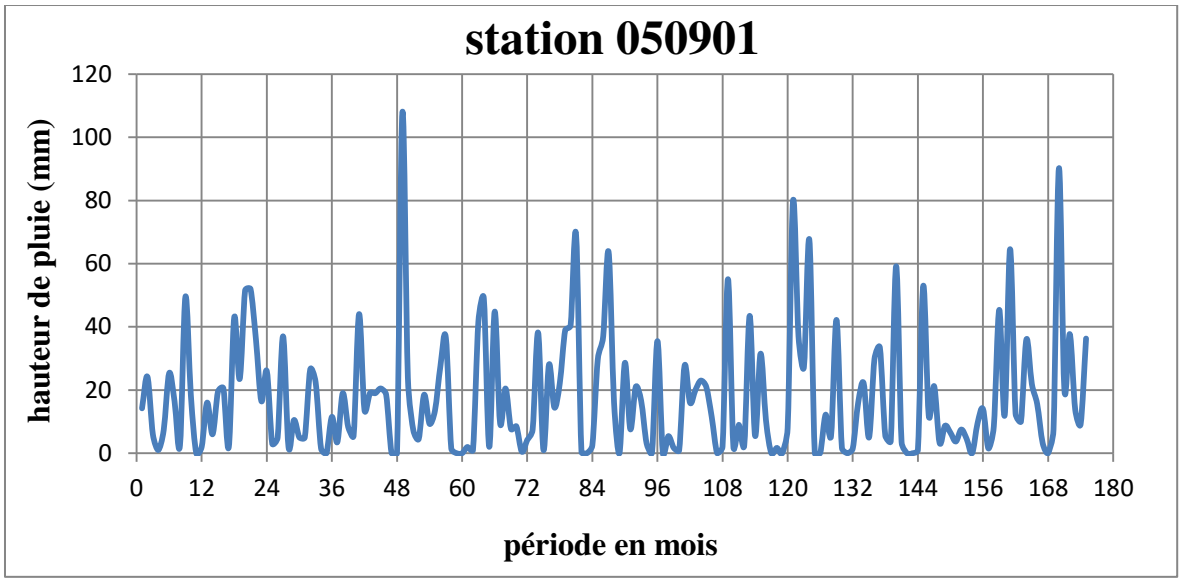
Wang, F., 2006. Quantitative Methods and Applications in GIS. CRC Press. Wei, T.C., McGuinness, J.L., 1973. Reciprocal Distance Squared Method: A Computer Technique for Estimating Area Precipitation. Technical Report ARS-Nc-8. US Agricultural Research Service, North Central Region, OH, USA.

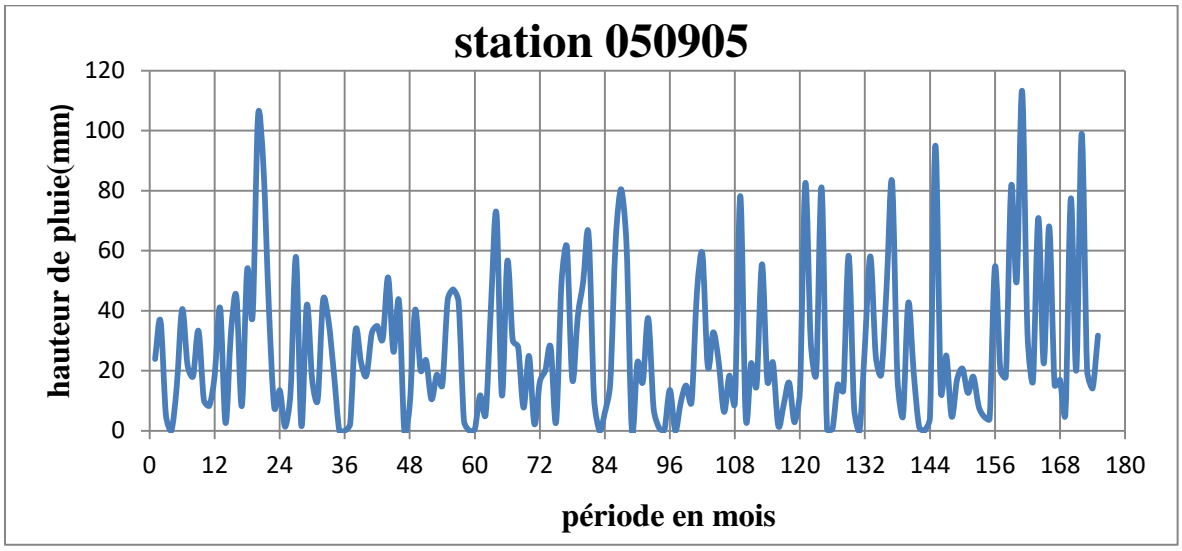
Zurada, J.M., 1992. Introduction to Artificial Neural Systems. Boston, MA, USA.

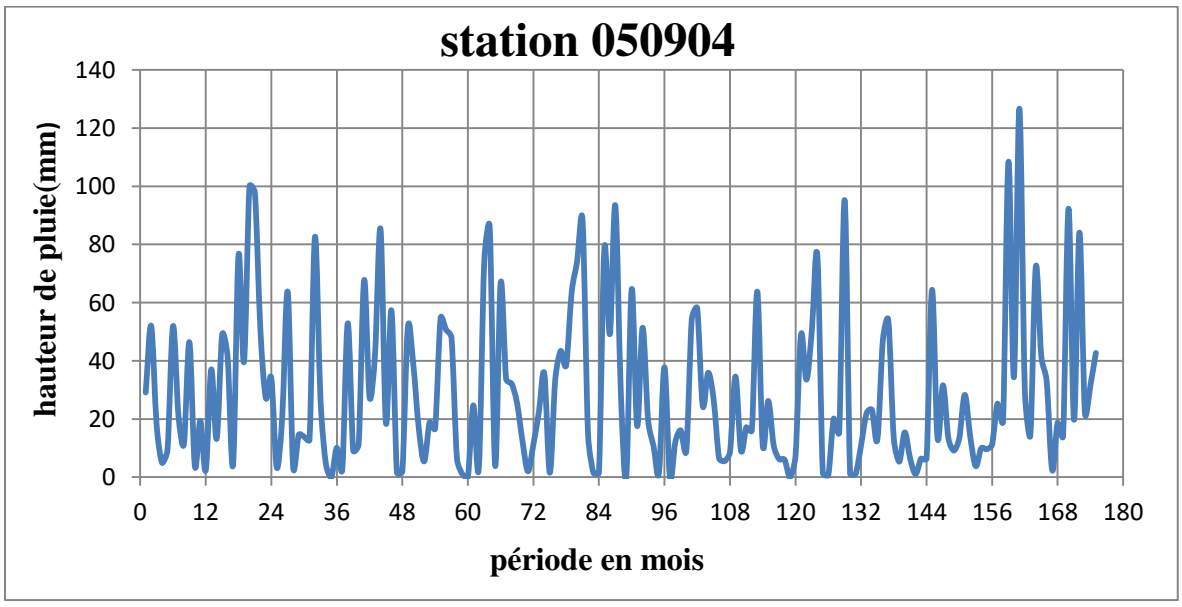
Mémoire de fin d'étude master en hydraulique Arabi. M (2017/2018) Sidi Bel Abbe « etude comparative des méthodes de comblement de lacune dans les enregistrements de précipitations ».

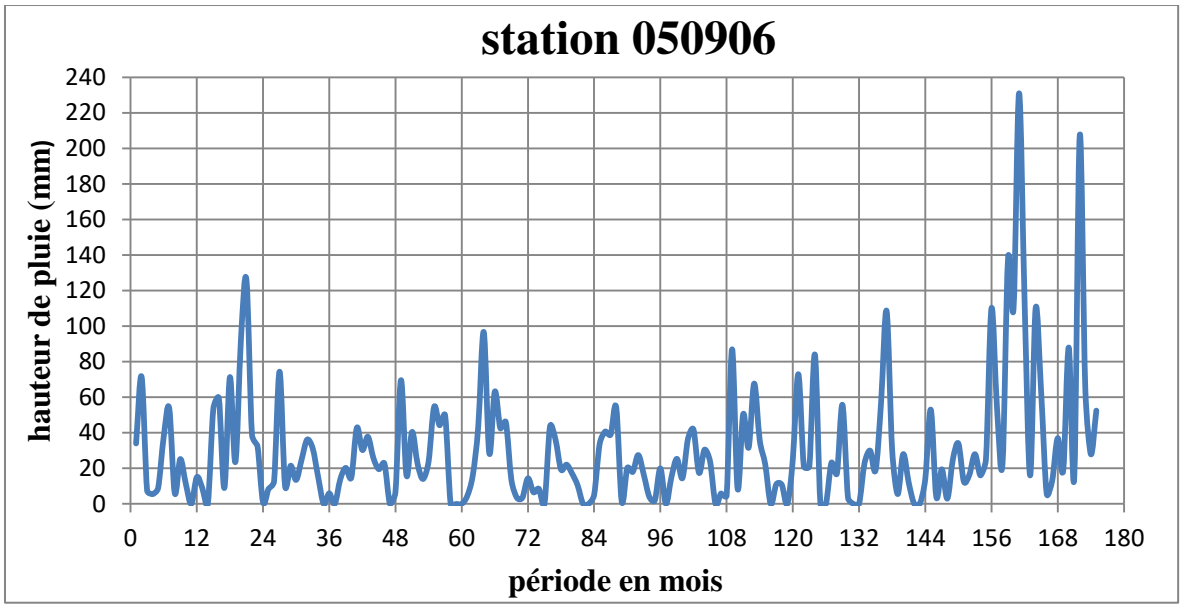
ANNEXES

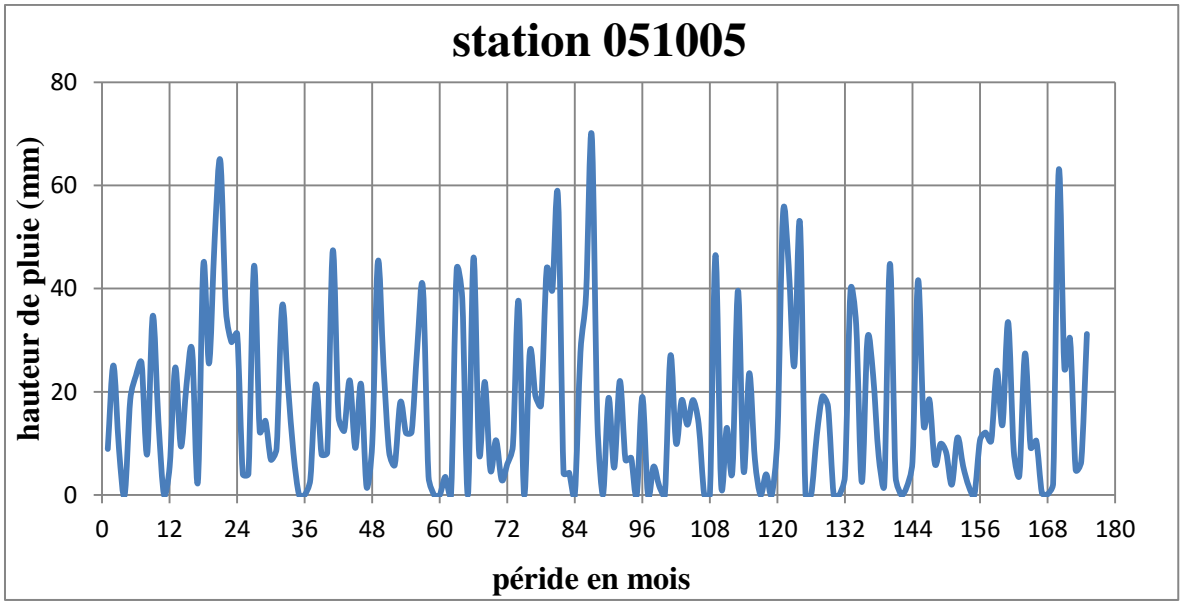
a) Données pluviométriques mensuelles utilisées pour le calage



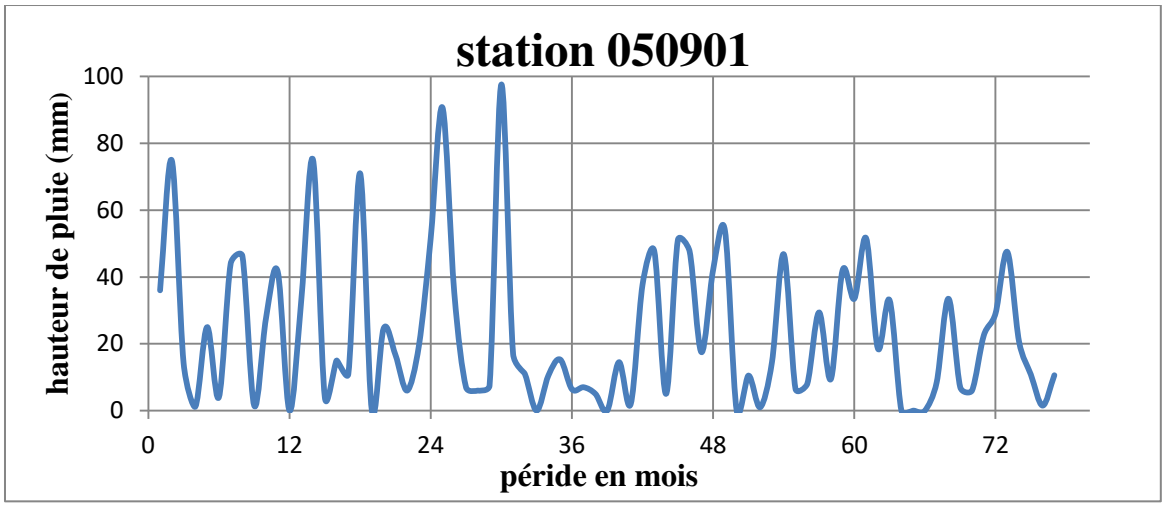


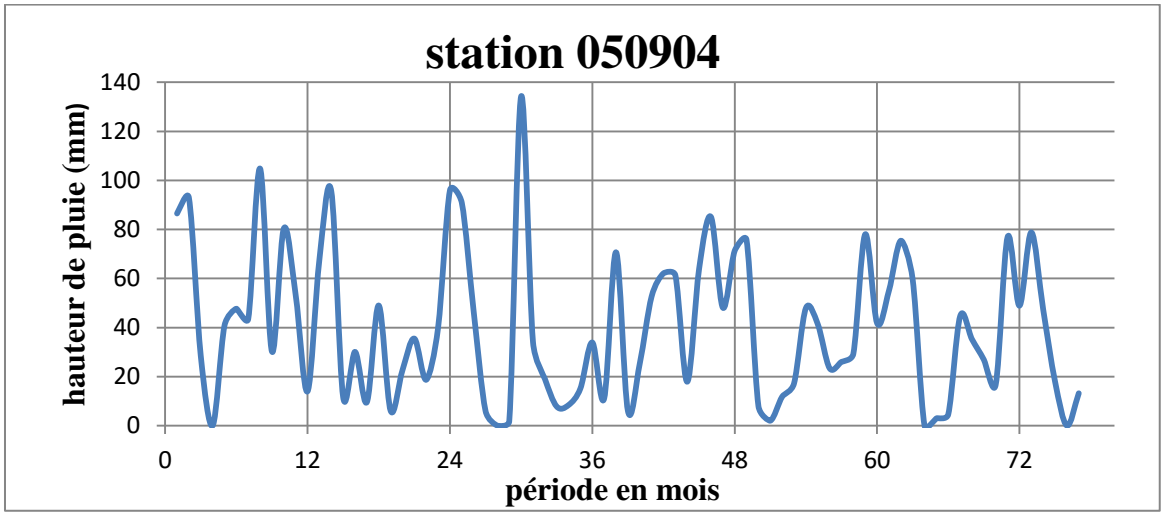


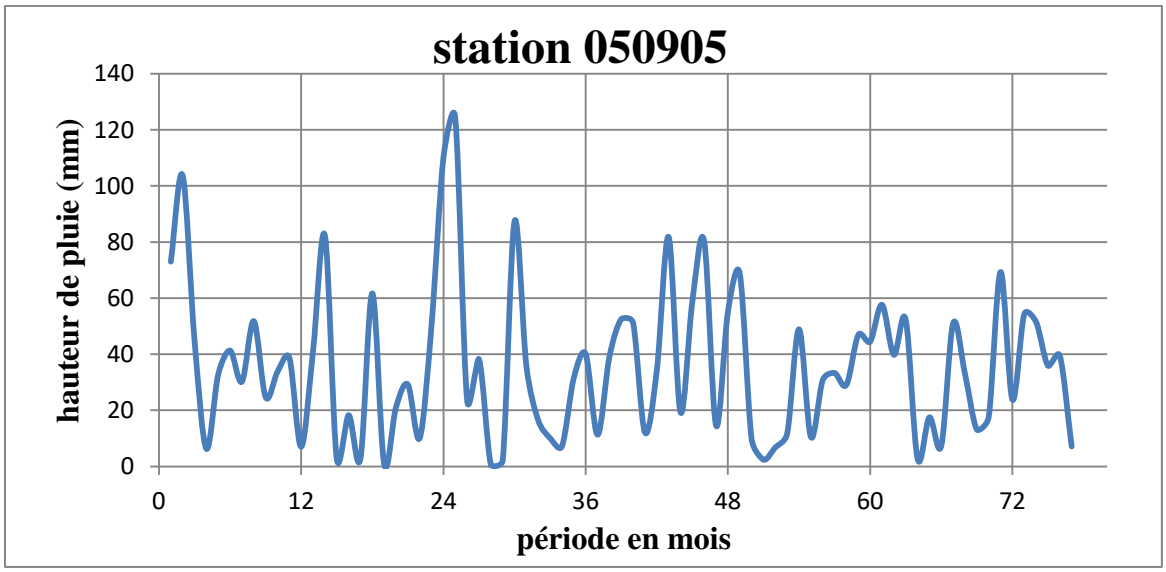

station 050906
hauteur de pluie (mm)
240
220
200
180
160
140
120
100
80
60
40
20
0
0 12 24 36 48 60 72 84 96 108 120 132 144 156 168 180
période en mois

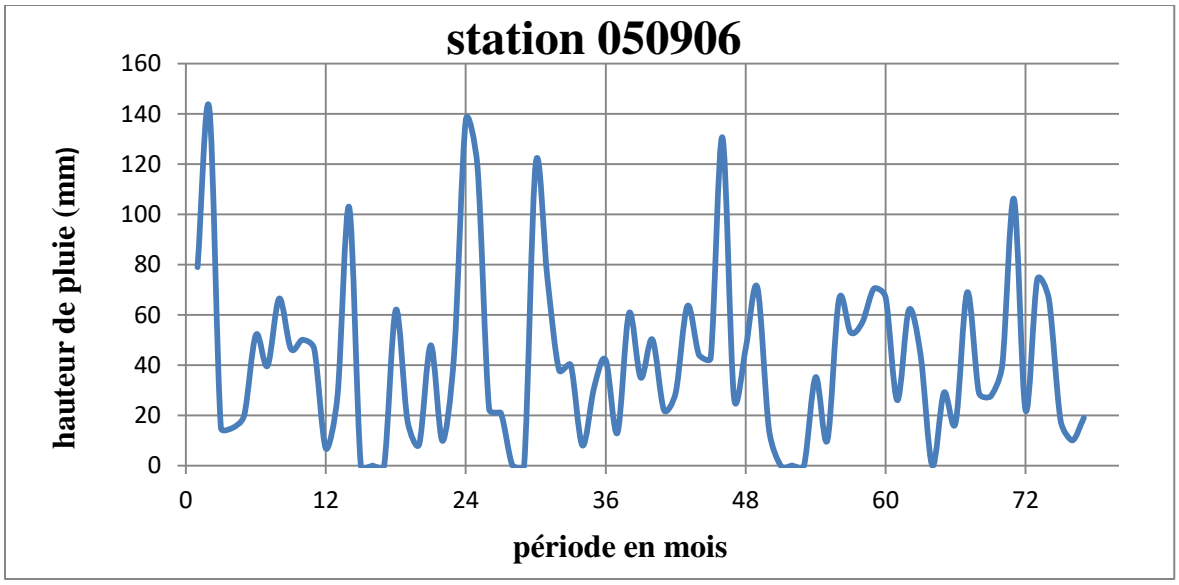

station 051005
hauteur de pluie (mm)
80
60
40
20
0
0 12 24 36 48 60 72 84 96 108 120 132 144 156 168 180
péride en mois

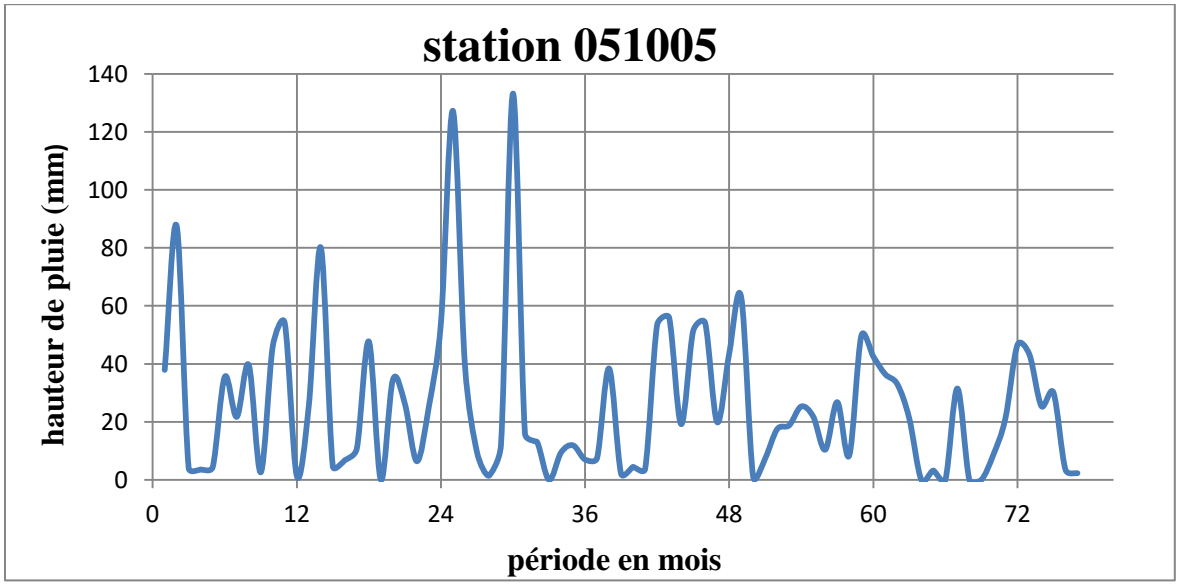
b) Données pluviométriques mensuelles pour la validation








station 050906
hauteur de pluie (mm)
160
140
120
100
80
60
40
20
0
0
12
24
36
48
60
72
période en mois


station 051005
hauteur de pluie (mm)
140
120
100
80
60
40
20
0
0
12
24
36
48
60
72
période en mois