Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A/Mira de Béjaïa

Faculté des Sciences Exactes

Département de Recherche Opérationnelle

MÉMOIRE DE FIN DE CYCLE

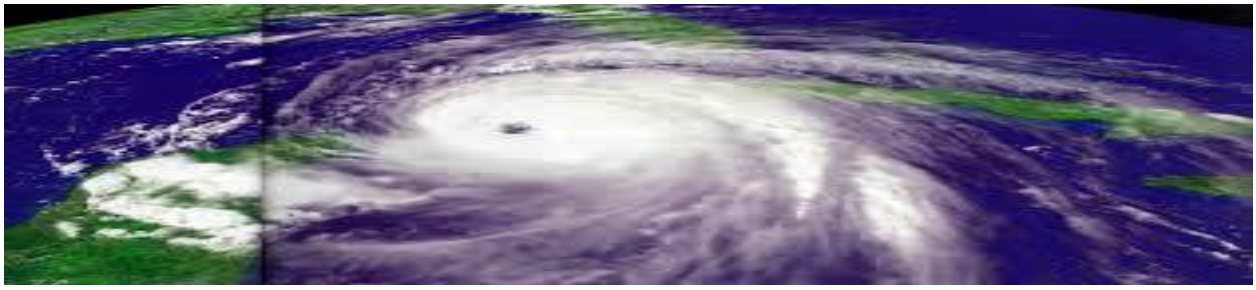
En vue de l'obtention du diplôme de Master

Option :

Modélisation Mathématique et Techniques de Décision

Thème

Analyse statistique du couple pluie-température du bassin versant de la SOUMMAM



Soutenu devant le jury composé de :

Présidente	M^{me}	*LAGHA* KARIMA
Examinateur	M^{r}	*MADANI* KHOUDIR
Examinatrice	M^{me}	*BARACHE* AICHA
Rapporteur	M^{r}	*AISSANI* DJAMIL
Rapporteur	M^{r}	*AKDIM* ABDELGHANI

Présenté par :

M^{r} *ADJLOUA* Elhocine

M^{r} *AZROU* Mebrouk

♣ *Promotion* 2014/2015 ♣

Remerciements

Nous tenons tout d'abord à remercier Dieu le tout puissant, qui nous a donné la force et la patience d'accomplir ce modeste travail.

Nous souhaitons adresser nos remerciements les plus sincères aux personnes qui nous ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable année universitaire.

Nous tenons à remercier très chaleureusement M^r A. AKDIM et M^r D. AISSANI qui nous ont permis de bénéficier de leur encadrement. Les conseils qu'ils nous ont prodigué, la patience, la confiance qu'ils nous ont témoignés ont été déterminants dans la réalisation de notre travail de recherche.

Nous tenons également à remercier M^r K. MADANI et M^r H. REMINI pour leurs conseils et leurs aides apréciables.

Nous n'oublions pas nos parents pour leur contribution, leur soutien et leur patience.

Nous adressons nos plus sincères remerciements à tous nos proches et amis, qui nous ont toujours soutenue et encouragé au cours de la réalisation de ce mémoire.

Je dédie ce modeste travail :

À ma trés chere garand mere Tounsia.

À mes chèrs parents que j'aime beaucoup pour leur précieux soutien.

À mes très chers frères : Noreddine, Arezki, Belaid, Tarik ,Mokrane et Axel.

À mes très chères sœurs : Djouhar, Zakia ,Nouria et Hafsa.

À toute ma famille, tantes, oncles, cousines et cousins.

À mon cher binôme : Elhocine.

À tout mes ami(e)s : Kimou, Ahmed, fawzi, Nounou, Djawida, Souad, Lydus, Djidji, Takfa, Amir, red, Khelil, Katia, Katia Hadid, Ghiles, Khaled, Massi, mourad, toufik,Touta, mouna, Chaben, Bilel biber, Samir, Jugo, Ossama, Missipsa et tout mes amis(es) du net.

À toute la promotion R.O 2015.

À tout ceux qui veulent réussir et combattent pour le faire.

Mebrouk

Je dédie ce modeste travail :

À mes parents qui m'ont été d'une aide très précieuse.

À ma trés chere garand mere Djohra, Marboura.

À tous mes frères (Hassan, Lyes) et mes sœures (Amira, Asma).

À toute ma famille, tantes, oncles, cousines et cousins.

À mon cher binôme : Mebrouk.

À mon ami et collègue de travail Faical.

À tous mes amis (Zaky, Amir, Amine, Moussa, Nabil, Billal, Fahim, Faouzi, Khaled, Halim, Youcef, Hakim, Raouf, Jigu, Yazid, Rabah, Bachir, Idir, Takfarinas).

À tous mes amies (Sonia, Sohila, Nadia, Linda, Farida, Wissam, Yasmine, Salima, Fouzia, Sabrina).

À toute la promo Recherche Opérationnelle 2014/2015.

ELHOCINE

Table des matières

Table des figures

Liste des tableaux

Introduction générale

A l'image des autres pays du monde, l'Algérie se voie concernée voire même affectée par le problème du changement climatique. Raison pour laquelle nombreux sont les travaux réalisés dans cette thématique, et qui dans leur ensemble visent à apporter une réponse aux multiples questions posées sur ce sujet, et étudier ces comportements climatiques par des facteurs qui permettent de déterminer la nature et le type du climat selon des indices climatiques connus durant les périodes observées.

Le changement climatique étant un scénario météorologique, affectant une région donnée, son impact est repéré en étudiant la composante climatique locale de la zone concernée, via ses paramètres mesurables dont la pluviométrie, la température, la vitesse de vent et tout autres facteurs ayant un rapport direct avec le climat et ces changements.

Cependant, les méthodes d'études différent en fonction du but recherché, allant des simples observations de terrain au niveau des stations météorologiques choisies dans la zone étudiée, ou d'une petite interprétation numérique et graphique aux traitements statistiques les plus complexes, utilisant des outils les plus performants et adéquats.

Dans le cadre de ce travail, nous allons adopter une nouvelle méthode de traitement des données numériques. En effet, le contexte général de notre étude vise à collecter des données de pluviométrie et de température concernant le bassin versant de la Soummam sur un laps de temps donné (de 1973 2014), traiter ces données avec des outils informatiques(R, Matlab, Excel...etc.) performants, puis comparer les résultats obtenus aux indices climatiques NAO affectant les zones méditerranéennes dont le bassin versant de la Soummam.

La première phase du travail consiste à collecter les données de pluviométrie et de température, relevant de station météorologique implantée à l'intérieur du bassin versant de la Soummam, située à l'aéroport de Bejaia, appartenant à un étage bioclimatique humide. Ces données qui ne sont pas toutes complètes, qui comporte des valeurs non attribuées ou non observées, requirent une reconstitution à l'aide des méthodes et des outils informatiques sophistiqués.

En deuxième phase, un traitement statistique pour l'ensemble des données, ce traitement choisis est cependant une analyse à haute fréquence avec des données journalières, et accompa-

gné d'une interprétation des résultats obtenus sous forme graphiques d'une excellente qualité selon l'orientation du travail, ce qui permettra de comprendre le comportement climatique de la région et l'influence de ces paramètres.

Dans la troisième et dernière phase de cette étude, nous comparons les résultats obtenus sous forme graphique ou analytique au modèl climatique global NAO. L'interprétation des résultats permettera de chercher l'impact de la NAO sur le climat du bassin versant de la Soummam, l'influence de ce modèl sur le comportement climatique et/ou sur les anomalies météorologiques.

Aussitôt les trois étapes de cette étude soient achevées, une interprétation finale est tirée, permettant de déduire l'efficacité de l'analyse dans le cadre de traitement des séries chronologiques relatives aux données climatiques, l'influence réel de la NAO sur le climat du bassin versant de la Soummam et le comportement de ce climat durant la période étudiée, au niveau de la zone d'étude choisie et finalement l'influence climatique observée sur le bassin versant de la Soummam durant la période en question.

Les présentations, les explications et les détails de ce travail sont données par les chapitres constituant ce manuscrit, selon une méthodologie chronologique adoptée dans le souci d'aboutir au but crucial de cette étude.

Chapitre 1

Généralités

Introduction

Ce chapitre est consacré à la définition de quelques concepts et généralités concernant les outils et les points essentiels de ce travail.

1.1 Définitions

1.1.1 La bioinformatique

L'informatique est un outil trés puissant et nécessaire vue les taches qui peuvent être accomplies avec, c'est un outil qui permis aux utilisateurs d'effectuer de nombreuses opération en un clin d'œil, l'archivage de données, traitement des données, analyse des séquences et leurs prédiction...etc. On trouve l'informatique comme un outil d'utilisation dans de nombreux domaines notamment la biologie d'où le terme bioinformatique.

La bioinformatique est apparue dans les années 1990, c'est une nouvelle discipline qui fusionne des disciplines de la biologie, l'informatique et traitement de l'information.

1.1.2 La climatologie

La climatologie est la science du climat. Mais son domaine d'application n'est pas restreint à ce dernier. Il s'agit d'une discipline beaucoup plus vaste. Elle emprunte à d'autres sciences des notions ou des résultats dont elle a besoin en faisant appel à des moyens techniques de plus en plus sophistiqués... . On peut en citer quelques unes : toutes les sciences concernant l'atmosphère comme la physique, la chimie, mais également la biologie, l'agronomie, l'hydrologie, l'économie, l'informatique... et surtout les statistiques pour le traitement et l'utilisation rationnelle des données.

C'est la branche de la géographie physique, et l'étude du climat, c'est-à-dire la succession des conditions météorologiques sur de longues périodes dans le temps. L'étude du temps à court terme est le domaine de la météorologie .En règle générale, le climat ne varie pas, ou assez peu, en un endroit donné du globe, sur une durée de l'échelle du siècle, mais sur des temps géologiques, le climat peut changer considérablement. La connaissance de nombreux paramètres, comme la température à différentes altitudes, l'influence des gaz à effet de serre, l'humidité relative, l'évaporation océanique, est nécessaire pour produire des modèles climatiques numériques et anticiper les changements du climat que l'on peut prévoir à plus ou moins long terme (30 ans).

Si la climatologie s'intéresse essentiellement à l'étude et à la classification des climats existants sur terre, une partie de la discipline traite aussi de l'interaction entre climat et société que ce soit l'influence du climat sur l'homme ou de l'homme sur le climat. La climatologie a un but d'analyser les éléments météorologiques qui constituent le climat et de rechercher des causes qui expliquent les différents climats et les fluctuations qui les accompagnent [1].

Dans la démarche climatologique, on distingue plusieurs phases associées à ces différents buts :

- La climatologie descriptive (ou analytique) : c'est l'étude géographique des conditions météorologiques caractérisant chaque région. Elle permet, à partir d'observation, la description des évolutions de l'atmosphère aux différents points du globe.

- La climatologie explicative (ou synthétique) : elle consiste a étudier les propriétés et l'origine des fluctuations ou des évènements climatiques avec une interprétation physique ou dynamique.

- La climatologie appliquée : c'est l'application de la climatologie a des domaines autres que l'atmosphère elle même puisque le climat agit constamment sur diverses sortes d'activités.

1.1.3 Le climat

Selon l'Organisation Météorologique Mondiale **(OMM)** Le climat est la distribution statistique des conditions de l'atmosphère terrestre dans une région donnée pendant une période donnée. Il se distingue de la météorologie qui désigne l'étude du temps à court terme et dans des zones ponctuelles. L'étude du climat est la climatologie.

La détermination du climat est effectuée à l'aide de moyennes établies à partir de mesures statistiques annuelles et mensuelles sur des données atmosphériques locales notamment la température, pression atmosphérique, précipitations, ensoleillement, humidité, vitesse du vent. Sont également pris en compte leur récurrence ainsi que les phénomènes exceptionnels [2]

1.1.4 Changement climatique

Plusieurs causes naturelles peuvent expliquer les bouleversements climatiques constatés au cours des temps [3].

Le climat terrestre varie selon les époques et les lieux. Les changements observés s'étalent généralement sur des longues périodes (plus de 30 ans) qui atténuent la perception que l'homme peut en avoir à un moment donné. Au cours des dernières décennies cependant, les changements climatiques semblent être accélérés. Dans ces conditions, il n'est pas surprenant que le public s'interroge sur la réalité de ces changements, leurs causes(astronomiques et géologiques)[3], leur devenir et, plus encore, leurs conséquences immédiates et lointaines sur les modes de vie tel que la santé, les écosystèmes et l'économie [4] ; même le milieu physique modifie (la température, le niveau d'eau des océans et les mers devient plus acides que avant). Le climat se modifie, plusieurs facteurs qui interviennent à ces changements climatiques notamment l'effet de serre, le soleil, et même le mode vie actuel, comme dans les pays riches et développés, dont l'Union européenne (UE) fait partie, qui explique le changement climatique. Les centrales qui transforment l'énergie en électricité et en chauffage, les déplacements en voiture et en avion, la fabrication des biens de consommation, l'agriculture, les déchets toxiques, toutes ces activités ont leur part de responsabilité dans le changement climatique.

Parmi les indicateurs de l'évolution climatique on cite :

1. L'augmentation de la température de surface sur la Terre.
2. La température des océans (les courants morins)
3. La réduction de la surface des glaces océaniques arctiques
4. Le niveau moyen des océans, ...

1.1.5 Les séries chronologiques

Une série chronologique est une série de valeurs réelles effectuées à des temps écartés régulièrement. les élément d'une serie chronologique sont séparés par des intervalles régulières, ça peut être des jours, des semaines, des mois, des années ...etc.

Les dates sont numérotées par des entiers positifs $n = 1; 2; :::$. Les données sont de nature très diverse ; il peut s'agir de relevés de phénomènes d'origine naturelle (température, hauteur des eaux, débit moyen journalier des riviéres, activité des taches solaires...etc), des séries économiques (indice de la consommation, taux de chômage, prix de matières premières...) ou financières (cours boursiers, taux de change...)[5].

1.1.6 Présentation du logiciel R i386 3.0.2

R est un environnement statistique créé par Ross Ihaka , Robert Gentleman[6] , il est à la fois un langage et un logiciel ; on cite quelques caractéristiques du R :

1. Un système performant de stockage et de manipulation des données ;
2. La possibilité d'effectuer des calculs matriciel et autres opérations complexes ;
3. Une large collection intégrée et cohérente d'outils d'analyse statistique ;
4. Un large éventail d'outils graphiques particulièrement flexibles ;
5. Un language de programmation simple et efficace qui inclue de nombreuses facilités[7].

Dans ce travail, nous avons utilisé un scripte prédefinit dans le logiciel R, en utilisant le package Amelia

1.1.7 Présentation du package Amelia

Amelia est un package qui complète R pour l'imputation multiple de données manquantes. Le package met en œuvre une nouvelle attente de maximisation avec des bottes d'algorithme de piégeage qui fonctionnent plus rapidement, avec un plus grand nombre de variables, et il est beaucoup plus facile à utiliser, à diverses chaînes de Markov, Monte Carlo donne essentiellement les mêmes réponses.

Le programme améliore également les modèles d'imputation en permettant d'avoir beaucoup d'informations potentiellement vastes et précieuses. Il inclut également des fonctionnalités d'imputer précisément les ensembles de données transversales, des séries de temps individuel, ou définir Softime série pour différentes sections. Un ensemble complet de diagnostic graphique sont également disponibles. Le programme est facile à utiliser, et cette algorithme est beaucoup plus robuste ; à la fois une ligne de commande simple et vaste interface utilisateur graphique sont inclus (Amelia).

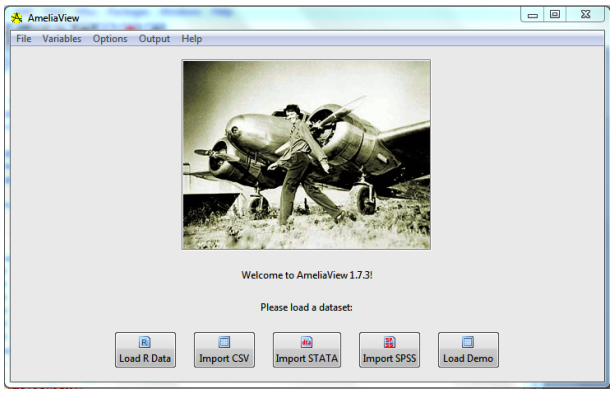


FIGURE 1.1 – L'interface du package Amelia

1.1.8 Présentation Matlab version 7.9.0.529 (R2009b)

Le logiciel Matlab (Matrix Laboratory) consiste en un langage interprété qui s'exécute dans une fenêtre dite d'exécution. L'intérêt de Matlab tient, d'une part, à sa simplicité d'utilisation : pas de compilation, déclaration implicite des variables utilisées [8].

C'est un outil très utilisé, dans les universités comme dans le monde industriel, qui intègre des centaines de fonctions mathématiques et d'analyse numérique de haut niveau dans de nombreux domaines (calcul matriciel, traitement de signal, traitement d'images, visualisations graphiques)...etc.

La programmation sous Matlab consiste à écrire des scripts de commandes Matlab, exécutables dans la fenêtre d'exécution. En outre, grâce aux diverses Toolboxes spécialisés (ensemble de scripts Matlab), Matlab s'enrichit au fur et à mesure.[9].

1.1.9 Présentation de la N.A.O

La NAO (Oscillation Nord Atlantique) est traditionnellement définie comme la différence de pression normalisée entre une station sur les Açores et l'autre sur l'Islande. Une version étendue de l'indice peut être dérivée de la moitié de l'hiver de l'année en utilisant une station dans la partie sud-ouest de la péninsule ibérique [10]. Ici, nous donnons des données pour SW Islande (Reykjavik), Gibraltar et Ponta Delgada (Açores). La NAO est calculée à partir de Gibraltar et SW Islande [10] .

Il existe d'autres modes de variabilité climatique : l'Oscillation Nord-Atlantique(NAO), l'Oscillation Arctique (AO), l'Oscillation Antarctique (AAO) et Pacific Decadal Oscillation (PDO) . . . etc.[11] qui sont montrés dans la Figure 1.2 .

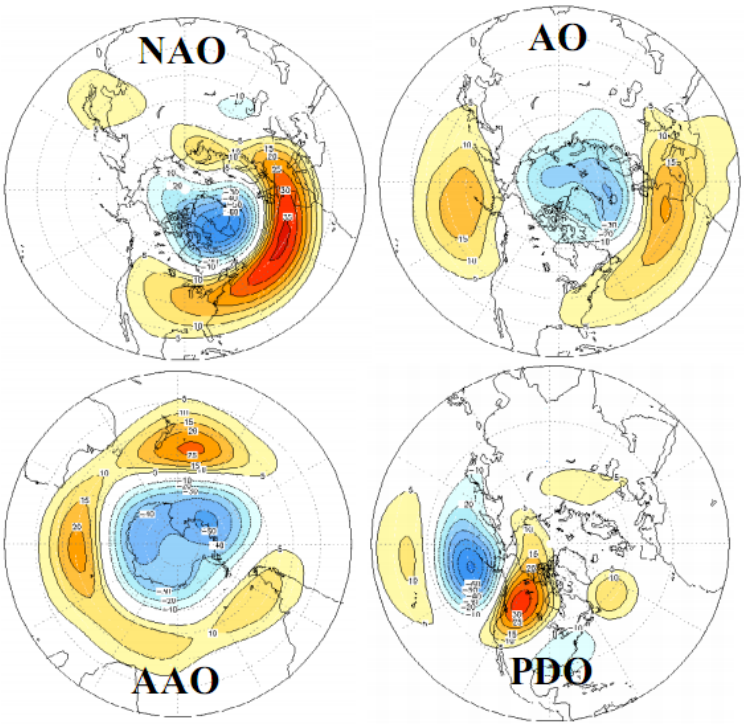


FIGURE 1.2 – Differentes Oscillations planétaires

L'indice NAO représente un "paquet climatique" qui permet de mettre en évidence les effets écologiques des fluctuations du climat [12]. Sur la côte nord-occidentale africaine. De plus, il est défini comme la différence de pression au niveau de la mer entre deux stations proches des "centres d'action" : au nord Stykkisholmur (Islande) et au sud soit Ponta Delgada (Açores), soit Lisbonne (Portugal), soit Gibraltar, voir la Figure 1.3 .

Une valeur très positive correspond à un système dépressionnaire en Islande qui provoque l'entrée de masses d'air humides mais tempérées sur l'Europe du nord. En revanche, pour des valeurs de NAO négatives, les flux maritimes sont déviés vers le sud en, association l'hiver, à un climat froid et sec [13].

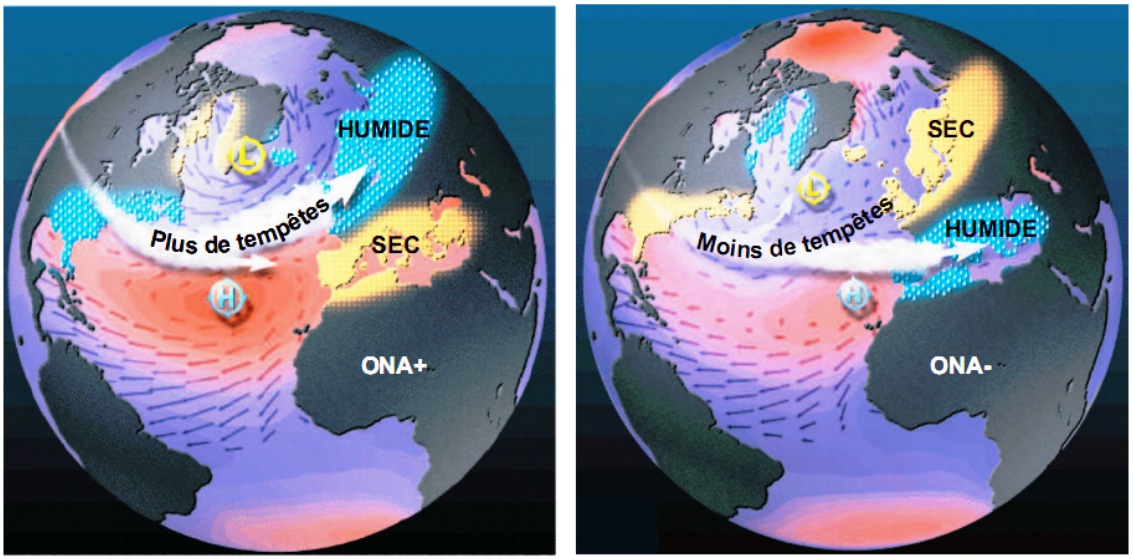


FIGURE 1.3 – Les indices NAO (positif et négatif)

1.2 Données (pluie-température)

La série chronologique des données de précipitations et de températures journalières au niveau du bassin versant du la Soummam (Bejaia), de la période 1973/2014 présente environ 15340 mesures pour la pluie, 15340 mesures pour la température (le nombre de jours de la période 1973/2014). Pour la pluie, chaque jour on compte la quantité d'eau journalière par unité (mm) au niveau de la station météorologique de l'aéroport de BEJAIA. En revanche, pour la température, on ne peut pas prendre une seule observation chaque jour car la variation de ces dernières change durant la journée même, donc on doit prendre la moyenne entre la température maximale et minimale pour chaque journée.

Chapitre 2

Techniques et Méthodes de Comblement de Données

Préambule

Afin d'analyser et de comprendre les changements climatiques et ses variations à l'échelle régionale, on passe inévitablement par la collecte et le traitement de données météorologiques traitant divers aspects. Dans cette étude, un soin particulier a été apporté sur la constitution d'un jeu de données climatiques qui soit le plus fiable et le plus complet possible sur la période d'étude retenue.

On s'intéresse aux données pluviométriques et températures qui proviennent da la station météorologique qui se situe au niveau de l'aéroport de BEJAIA. Ces données peuvent être fournies par les services de la Météorologie Nationale, l'Hydraulique (Agences de bassins), de l'Agriculture (Offices régionaux de mise en valeur Agricole,...), des Eaux et Forêts ou de l'Intérieur. La disparité de ces données pose souvent un problème de la qualité des interprétations.

2.1 Les causes de ce manque de données

Il y'a beaucoup de facteurs qui conduisent à des trous dans les données :

- Les erreurs accidentelles et aléatoires dues à :
  1. La collecte au cours de l'observation :
     - ◇ Pertes d'eau.
     - ◇ Absence de l'observateur non signalée.
     - ◇ Déguisement de la donnée ou décalage de jour.
     - ◇ Mauvaises conditions de mesure.

2. L'inscription sur les originaux et copie :
   ◇ Oublies de virgules, mauvaises interprétations des chiffres.
3. La transmission et saisie de données.
4. Le calcul des cumuls, moyennes, ...

– Les erreurs systématiques :

1. Déplacement du site d'observation au cours du temps.
2. Modification de l'environnement immédiat du poste de mesure :
   ◇ Déboisement, boisement, urbanisation, construction d'un barrage, ...
3. La non conformité du matériel de mesure (défaut d'appareillage non remarqué).

L'objectif

Notre objectif dans cette partie, c'est de compléter une base de données de pluviométriques et de températures pendant une période de 42 ans, avant de commencer le traitement à l'échèlle régionale du bassin versant de la Soummam. Notre étude permet d'évaluer un manque de données et de déterminer les meilleurs modèles mathématiques et outils informatiques permettant de les combler.

2.2 Comblement de données

2.2.1 Les données manquantes

Une donnée manquante peut être définie comme une donnée qui était visée par le processus de collecte, mais qui n'a pas pu être obtenue.

Une série incomplète est une série pour laquelle les valeurs de certains attributs sont inconnues, ces valeurs sont dites manquantes.

Soit l'observation est un vecteur des valeurs de certains indicateurs ou attributs, les valeurs manquantes peuvent être de deux natures :

1. Valeur manquante totale : c'est-à-dire que toutes les valeurs de l'observation manquent.
2. Valeur manquante partielle, c'est-à-dire que l'observation est présente avec quelques valeurs manquantes.

Exemple

	Attributs			
Observations	a_1	a_2	a_3	a_4
w_1	56	98	10	5
w_2	40	87	21	?

FIGURE 2.1 – Exemple d'un manque de données partielles

2.3 Les mécanismes des données

Il existe plusieurs mécanismes pour le traitement des données manquantes. Donc avant de passer à ce traitement on doit faire une évaluation du mécanisme des valeurs manquantes et choisir après une méthode adéquate pour le traitement.

On distingue trois mécanismes de données manquantes :
Les données manquantes selon un mécanisme complètement aléatoire, les données manquantes selon un mécanisme aléatoire et les données manquantes selon un mécanisme non aléatoire.

2.3.1 Mécanisme complètement aléatoire (MCAR, Missing Completely At Random)

Ce cas de figure se présente lorsque la probabilité de valeur manquante d'une variable ne dépend ni de la variable elle-même ni d'une autre variable de la base de données, et on écrit :

$$P(r|X_{obs}, X_{miss}) = P(r)$$

X_{obs} : représente les données observées.
X_{miss} : représente les données manquantes.
Par exemple : Si les données manquantes pour la variable " diagnostic " d'un registre s'expliquent par le fait que les laboratoires d'analyse n'arrivent plus à fournir les résultats en raison d'un dysfonctionnement, on dit que les diagnostics manquants suivent un mécanisme complètement aléatoire, car ils ne dépendent ni de la variable " diagnostic ", ni d'une autre variable du registre ; ils dépendent de l'état de fonctionnement du laboratoire.

2.3.2 Mécanisme aléatoire (MAR, Missing At Random)

On parle de données manquantes selon un mécanisme aléatoire lorsque la probabilité de valeur manquante d'une variable ne dépend pas de la variable elle-même, mais d'une autre variable de la base de données, et on ecrit :

$$P(r|X_{obs}, X_{miss}) = P(r|X_{obs})$$

X_{obs} : représente les données observées.

X_{miss} : représente les données manquantes.

Par exemple : Lorsque les données manquantes pour le variable " diagnostic " d'un registre sont observées seulement chez les femmes, la probabilité que le diagnostic manque ne dépend alors que du sexe. Dans ce cas, les données manquantes, pour la variable " diagnostic ", suivent un mécanisme aléatoire, car c'est le fait d'être homme ou femme qui explique les données manquantes et non le type de diagnostic.

2.3.3 Mécanisme non aléatoire (NMAR, Not Missing At Random)

On parle de données manquantes selon un mécanisme non aléatoire lorsque la probabilité de valeur manquante d'une variable dépend de la valeur de cette variable elle-même, et on écrit :

$$P(r|X_{obs}, X_{miss}) = P(r|X_{miss})$$

X_{obs} : représente les données observées.

X_{miss} : représente les données manquantes.

Par exemple : Un adolescent qui sort de lui même d'un essai longitudinal sur l'obésité parce qu'il constate qu'il a grossit.

2.4 Les méthodes de comblement

En statistique, on parle de valeur manquante lorsqu'on a pas d'observation pour une variable donnée pour un individu donné.

Les solutions sont différentes car de nombreuses techniques de traitement de données manquantes ont été développées dans les années 90, sans prétendre être exhaustives, on identifiait déjà plus d'une vingtaine, pour la plupart issues des recherches en statistique. Depuis, les chercheurs en intelligence artificielle et fouille de données (Data mining), se sont mis à étudier la question et à développer de nouvelles techniques, recenser l'ensemble de ces techniques serait fastidieux. Aussi avons-nous opté pour une mise en évidence des principales caractéristiques des différentes méthodes. Nous pouvons alors présenter les techniques les plus utilisées par rapport au mécanisme de donnée. voir le tableau suivant :

Les modèles mathématiques de comblement les plus connus sont

2.4.1 Procédures de suppression

Étude des cas complets (Listwise deletion)

Cette méthode permet de se ramener à une base de données complète par réduction de la dimension du problème. Pour cela, tous les exemples de la base contenant des valeurs

Hypothèses	**Définitions**	**Méthodes suggerées**
MCAR	P(M|Y,φ)=P(M|φ) $\forall$ Y,φ	Listwise et Pairwise
MAR	P(M|Y,φ)=P(M|Y_{obs},φ) $\forall$ Y_{miss},φ	Regression et imputation multiple
NMAR	P(M|Y,φ)=P(M|Y_{miss},φ)	models mixtes

TABLE 2.1 – Définition formelle des mécanismes de données manquantes

manquantes sont supprimés (Il s'agit de supprimer toutes les rangées de la matrice de données présentant des données perdues). Par conséquent, cette méthode sacrifie un grand nombre de données [14], la suppression de 10% de données de chaque variable dans une matrice de cinq variables peut facilement provoquer l'élimination de 59% des observations de l'analyse [15], on en a observé une baisse de dimension de l'échantillon de 624 à 201 avec l'utilisation de la méthode de suppression listwise.

Les techniques statistiques d'analyse de données ayant besoin d'un nombre suffisant d'observations pour que leurs résultats soient valides. Dans des cas qui ne sont pas rares, où la quasi-totalité des exemples possède des valeurs manquantes, elle devient même inutilisable. D'autre part, les statistiques, telles que la moyenne ou la variance, seront fortement biaisées, à moins que le mécanisme de génération des données ne soit complètement aléatoire [16].

Malgré le fait que la grande perte de données réduit la puissance et l'exactitude statistiques [17], cette technique, du fait de sa simplicité est fréquemment l'option implicite pour l'analyse dans la plupart des progiciels statistiques. Cette méthode est aussi appelée par certains auteurs la méthode d'analyse des données disponibles (available-case analysis).

Étude des cas disponibles (Pairwise deletion)

Dans cette méthode de type (Data deletion), en couplant toutes les variables actuelles dans la matrice de données, et puis en supprimant les données couplées contenant des informations absentes.

On ne considère que les cas où ces variables sont complètement observées. Par exemple : si la valeur de l'attribut A est absente pour une observation, les autres valeurs pour le reste des attributs de la même observation pourraient encore être employées pour calculer des corrélations, telles que celle entre les attributs B et C. Comparée à la première méthode (Étude des cas complets), [18] cette méthode conserve beaucoup plus des données qui auraient été perdues si on employait la méthode d'étude des cas complets. Il s'agit d'une autre méthode proposée par les logiciels statistiques, mais générant la problématique :
le nombre d'observations (n) varie pour le calcul de chaque valeur de la nouvelle base de

données, le risque d'obtenir une base de données réduite est grand et, encore une fois, la représentativité sera biaisée si les données manquantes ne sont pas distribuées de façon complètement aléatoire.

Les études de Monte Carlo ont montré que la suppression par la méthode "Listwise deletion" donne des évaluations moins précises pour les paramètres d'estimation. La méthode "Pairwise deletion" est uniformément plus précise, bien que les différences puissent, parfois être minimes [19]. Cette méthode est appliquée losque le cas où le nombre des valeurs manquantes est relativement faible [20].

2.4.2 Procédures de remplacement

Ces procédures visent à se ramener à une base complète en trouvant un moyen adéquat pour remplacer les valeurs manquantes. On nomme ce procédé "imputation" (complétion ou substitution).

De façon générale, des procédures de remplacement peuvent être employées dans certains cas, tant qu'on a une bonne raison pour remplacer. Il est facile d'exécuter ses procédures et certaines sont incluses comme options avec les logiciels statistiques. Les avantages les plus importants de ces procédures sont la conservation de la dimension de la base de données, par conséquent, de la puissance statistique d'analyse. Dans une plus ou moins large mesure, toutes les procédures de remplacement sont décentrées s'il y a une distribution non-aléatoire des valeurs manquantes. Cependant, le remplacement des données manquantes est approprié quand les corrélations entre les variables sont faibles [17],[21].

De différentes procédures de remplacement de données manquantes ont été élaborées au cours des années. On constate que les différences entre les diverses méthodes diminuent avec : une plus grande dimension de la base de données, un plus petit pourcentage des valeurs manquantes et une diminution au niveau des corrélations entre les attributs [19]. On a deux catégories de remplacement pouvant être distinguées :

- Une imputation simple (imputation par la moyenne, imputation par la régression. . ..).
- Une imputation multiple.

2.4.3 L'imputation simple

Imputation par la moyenne

Les valeurs manquantes de chaque attribut sont remplacées par la moyenne des valeurs non-manquantes (mais on ne tient pas compte des autres variables).

Il y a deux variantes de l'imputation par la moyenne : Imputation par la moyenne totale, imputation par la moyenne des sous-groupes. Pour l'imputation par la moyenne totale,

la valeur absente d'un attribut est remplacée par la moyenne des valeurs de cet attribut de toutes les observations. Pour l'imputation par la moyenne de sous-groupe (classe), la valeur manquante est remplacée par la moyenne du sous-groupe (classe) de l'attribut en question.

L'inconvénient de cette méthode est la sous-estimation de la variance et de biaiser la corrélation entre les attributs, cela veut dire que la distribution des données est loin d'être préservée.

Cette manière de procéder serait encore moins recommandable que l'utilisation de la méthode d'étude des cas complets [22]. Il est attendu que, même si les données sont manquantes selon un mécanisme complètement aléatoire, l'estimé de la moyenne de la distribution sera valide, mais, par contre, l'estimé de l'écart type s'avère automatiquement biaisé [17]. En remplaçant les valeurs manquantes par une valeur constante, la variance de l'attribut s'avère inévitablement réduite. Il s'en dégage que la valeur de l'erreur type, diminuée par la réduction de la variance et par l'augmentation de la taille de l'échantillon, sera artificiellement plus petite que ce qui aurait dû être observé [22].

Les études ont été quelque fois peu concluantes concernant l'efficacité de cette substitution. La substitution par la moyenne est moins précise que la méthode Listwise deletion, alors que d'autres, ont prouvé que la substitution par la moyenne est plus précise que le "Listwise deletion" et le "Pairwise deletion".

C'est pour cela que cette méthode de comblement par défaut est utilisée sur la plupart des langages de traitement statistique, notamment le langage R.

Exemple illustratif :

A=	1	2	0	3	5	0	7	1	3	moy $\simeq$ 2.44

TABLE 2.2 – Exemple d'une série de données complètes A

	S1	S2	S3	moyenne
$A^{'}$=	{ 1 , 2 , 0}	{ NA , 5 , 0 }	{ 7 , 1 , 3 }	$moy_{S1} \simeq 2.375$ $moy_{S2} \simeq 2.5$ $moy_{S3} \simeq 3.66$

TABLE 2.3 – Exemple d'une série de données incomplètes $A^{'}$

Mais cette méthode n'est pas toujours valide, car dans notre exemple, si la dimension augmente, donc la moyenne diminue (existence d'une tendance des résidus).

L'imputation par régression

C'est une approche en deux étapes : d'abord, on estime les rapports entre les attributs,

et puis on utilise les coefficients de régression pour estimer la valeur manquante [23]. La condition fondamentale de l'utilisation de l'imputation par régression est l'existence d'une corrélation linéaire entre les attributs. La technique suppose également que les valeurs manquantes sont au hasard.

Dans le contexte des valeurs manquantes, deux modèles de régression sont en général employés : la régression linéaire et la régression logistique. Cette dernière est plutôt utilisée pour traiter les variables discrètes, alors que la régression linéaire est appliquée sur des variables continues [17].

Pour chacune de ces méthodes, il est possible de tenir compte de l'information de classe en n'utilisant que les observations d'une même classe pour estimer les paramètres de régression.

L'inconvénient de cette méthode, c'est les hypothèses qui sont faites sur la distribution des données. Supposer une relation linéaire entre les variables, revient à faire des hypothèses qui sont rarement vérifiées, dans cette situation, le remplacement des valeurs manquantes par des valeurs prédites basées sur un modèle biaisé ne constitue pas un traitement approprié.

Ces méthodes, seraient beaucoup plus efficaces, exclusivement dans le cas où le modèle de régression est adéquat [24]

L'imputation par hot-deck

L'imputation hot-deck est une procédure qui consiste à remplacer les valeurs manquantes d'une observation par des valeurs empruntées à d'autres observations similaires, définies comme étant celles pour lesquelles les valeurs sont les plus identiques à celles de l'observation présentant une donnée manquante, l'hypothèse sur laquelle elle s'appuie est que les probabilités de présence des valeurs sont égales dans les cas d'imputation.

Même si ce type de méthodes préserve les distributions des variables, elles risquent d'altérer les relations entre les variables [24].

2.4.4 L'imputation multiple

Ces dernières méthodes souffrent d'un défaut majeur : elles sont déterministes en ce sens que deux individus qui ont les mêmes valeurs des autres variables auront la même valeur imputée.

On considère Y comme une variable aléatoire, dont la loi conditionnelle est de $(Y/X_1, X_2, X_3, X_4, \ldots\ldots\ldots\ldots X_p)$. La solution la plus élaborée rendue possible par les moyennes de calculs actuels est l'imputation multiple : on effectue plusieurs tirages, ce qui

conduit à plusieurs tableaux de données que l'on analyse séparément. Les résultats obtenus sont ensuite regroupés pour étudier la variabilité attribuable aux données manquantes.

Le problème est assez complexe en réalité sur les calculs : si l'on utilise un modèle de la régression pour estimer la valeur manquante

$$Y = \beta_0 + \beta_1 X + ... + \beta_p X + \epsilon$$

il ne suffit pas de tirer des valeurs dans la distribution du résidu ϵ, mais il faut tenir compte du fait que les coefficients β_j du modèle sont estimés, donc aléatoire. On doit donc tirer aussi des valeurs des β_j dans leur distribution a posteriori qui elle-même dépend des valeurs manquantes.

Elle est disponible sur certains logiciels (ex : langage R avec un traitement par l'ACP et la régression, logicielle SPSS) elle semble constituer une méthode préférable à celles présentées précédemment, Dans tous les cas [25], soulignons qu'elle est préférable à la méthode "Listwise deletion".

2.4.5 Procédures basées sur un modèle

Maximum de vraisemblance

Sous sa forme plus simple, l'approche de maximum de vraisemblance pour analyser des données manquantes, suppose que les données observées sont tirées d'une distribution normale multi variée [26]. Au lieu d'imputer des valeurs aux données manquantes, ces méthodes définissent un modèle à partir des données disponibles et basent les inférences de ce modèle sur la vraisemblance de la distribution des données sous ce modèle.

Maximisation d'espérance

Une dernière approche assez fréquente consiste à utiliser l'algorithme de maximisation d'espérance "EM" (Expectation-Maximization) pour estimer les valeurs manquantes [27], qui est un processus itératif [28] . Il est généralement utilisé pour estimer les paramètres d'une densité de probabilité. Il peut être appliqué sur des bases de données incomplètes, et présente l'avantage de procéder à l'estimation des valeurs manquantes en parallèle de l'estimation des paramètres. On suppose l'existence d'un modèle de génération des données, par exemple un mélange de gaussiennes pour les variables continues. Les paramètres du modèle sont calculés suivant la méthode du maximum de vraisemblance, de manière itérative, ont présenté une nouvelle méthode de substitution basée sur une version simplifiée d'EM.

A partir d'une estimation par défaut des valeurs manquantes, les paramètres du modèle sont ré-estimés, à chaque itération, à partir de la matrice complète, de manière à accroître la vraisemblance des données. Le modèle avec ses nouveaux paramètres est alors utilisé pour ré-estimer les valeurs manquantes. Puis on recommence jusqu'à ce que la convergence soit atteinte (ou considérée comme telle). À la fin de l'exécution de l'algorithme, on dispose non seulement des paramètres de notre modèle, mais également d'une matrice de données complétée.

Cette technique est très coûteuse en temps de calcul comme beaucoup d'approches itératives [29]. De plus, elle demande la spécification d'un modèle de génération des données. Cette tâche implique de faire un certain nombre d'hypothèses, ce qui est toujours délicat. Pour ces raisons, l'application d'EM pour remplacer les données manquantes n'est pas toujours envisageable. Un grand nombre de méthodes de traitement des valeurs manquantes sont explorées dans les diverses études relevées dans la littérature, il est impensable de tenter d'inclure l'ensemble de ces diverses méthodes dans une seule recherche ;

Les méthodes de traitement les plus souvent comparées ont donc été considérées, Et les figures suivantes présentent l'utilisation des différentes méthodes discutées dans ce manuscrit [30].

Technique	Description	Champ d'application	Avantage	Inconvénient	Référence
Procédures de suppression					
Étude des cas complets (Listwise deletion)	Supprime toutes les observations dont certaines données sont manquantes.	Il convient d'éviter	Facile à utiliser (par défaut dans la plupart des logiciels statistiques).	Sacrifie une grande quantité de données et a un impact négatif sur les paramètres d'estimation (corrélation – régression) et sur la puissance statistique.	Kim and Curry (1977) Raymond (1986) [30], Malhotra (1987) , Little and Rubin (2002)
Étude des cas disponibles (Pairwise deletion)	Crée une matrice de corrélation avec les valeurs disponible (chaque couple de variables est pris deux à deux)	Lorsque les données sont relativement faible (moins de 10 %).	Préserve davantage les données et est plus précise que la suppression listwise	Corrélations ou covariances biaisées	Gleason and Staelin (1975) , Kim and Curry (1977) , Raymond (1986) , Roth (1994)

FIGURE 2.2 – Première partie

Procédures de remplacement					
Imputation par la moyenne totale (Total mean substitution)	Remplacer par la moyenne de toute la série de données observées Les D.M	Lorsque les corrélations entre les variables sont faibles (r < |.20|) et le taux des leurs manquantes moins que 10%.	Préserve la taille de la base de données et la rend facile à utiliser.	La sous-estimation de la variance et de biaiser la corrélation entre les variables (la distribution des données est loin d'être préservée).	Ford (1976) [88], Raymond (1986), Little and Rubin (1987), Kaufman (1988), Quinten and Raaiimakers (1999).
Imputation par la moyenne de chaque classe (Subgroup mean substitution)	Remplacer par la moyenne des valeurs disponibles les D.M de même classe.	Quand il est facile de définir les classes (classification supervisée).	Donne de meilleurs résultats, par rapport à Imputation par la moyenne totale.	La sous-estimation de la variance et de biaiser la corrélation entre les variables (la distribution des données est loin d'être préservée).	Ford (1976).

FIGURE 2.3 – Deuxième partie

Imputation par régression (Régression imputation)	On utilise les valeurs disponibles pour estimer les paramètres d'un modèle de régression, puis utiliser ces paramètres pour estimer la valeur manquante.	Lorsque plus de 20 % des données sont manquantes et les variables sont fortement corrélées.	Préserver l'écart des données estimées par rapport à la moyenne et la forme de la distribution	Distorsions des degrés de liberté et augmentation artificielle des relations entre les variables .	Frane (1976), Raymond and Roberts (1987), Little and Rubin (2002).
L'imputation hot deck (Hot-deck imputation)	Remplacer une valeur manquante par la valeur de la même variable à partir d'un cas similaire dans l'ensemble de données .	Lorsque la similarité entre les cas est facile à déduire.	Préserve les distributions des variables	Risquent d'altérer les relations entre les variables	Ford (1983) [88], Sinharay, Stern et Russel (2001)
Imputation multiple	Estimez m > 1 ensembles de valeurs plausibles pour les données manquantes sont créés. Chacun de ces ensembles est utilisé pour remplir les données manquantes et ainsi créer m ensembles complets de données, et elles	Sous l'hypothèse que les valeurs manquantes sont aléatoires.	L'induction statistique (écart-type, p-values, etc.) qui découle de l'IM est généralement valide car elle incorpore l'incertitude engendrée par les données manquantes.	La complexité de calcul des m matrices (contraintes d'espace mémoire et de temps de traitement). Ne permet pas de seulement compléter une base de données... mais oblige à réaliser une analyse statistique.	Rubin (1978), (Schafer et Graham, (2002) [, Little et Rubin, 2002), Fichman et Cummings (2003). Paul D.Allison(2001)

FIGURE 2.4 – Troisième partie

2.4.6 Les méthodes de traitement (cas d'une série chronologique)

Le tableau 2.4 nous montre la méthode adéquate pour le comblement des données dans notre cadre d'etude [31]

Taux des données manquantes	**Méthodes proposées pour traiter les données manquantes**
$< 5\%$	Toutes les méthodes fonctionne bien
$[5\% - 10\%]$	En utilisant une valeur constante e.g modèle simple peut-être bon si la correction entre les variables est négligeable les variables est négligeable les methodes d'imputation simple e.x regression, travaille bien plusieurs méthodes d'imputation sont les bonnes choix
$[10\% - 15\%]$	Méthodes d'imputation simples pourraient être biaisées imputation multiple fortement recommandé
$> 15\ \%$	Plusieurs méthodes d'imputation sont presque toujours fiables

TABLE 2.4 – Les méthodes adéquates selon le pourcetage des données manquantes

Chapitre 3

Comblement et validation

3.1 Comblement des lacunes

3.1.1 Température

Pour les températures journalières, enregistrées durant la période $1973 - 2014$, nous utilisons à la fois, des températures minimales et maximales afin de déterminer les températures moyennes. Cette procédure est conventionnellement adoptée dans toutes les études météorologiques faisant référence aux températures moyennes.

Dans l'idéal, il conviendrait de disposer des températures moyennes déterminées à partir de l'intégration au cours du temps des températures mesurées en surface. Cependant, peu de stations sont actuellement capables de fournir de telles données. Par conséquent, nous sommes amenés à calculer les températures moyennes à partir des minimales et des maximales.

Le nombre total de lacunes (données manquantes) dans l'ensemble de la série des températures moyennes journalières est de 4.7%, essentiellement observées durant la période 1973/2014.

La reconstitution des lacunes (données manquantes) à partir de la technique de l'imputation multiple, et cela se fait à l'aide de l'outil qu'on a cité auparavant (Amelia).

La figure 3.1 montre les différentes valeurs essentielles pour la procédure de comblement, notamment le maximum et le minimum des températures moyenne observés, la moyenne et l'écartype de la série, ainsi que le nombre des données manquantes. On peut constater que dans la majorité des cas, les composantes principales retenues expliquent en général plus de 95% de la variabilité des températures, ce qui permet d'attribuer une confiance aux données reconstituées à partir de la technique et l'outil utilisé.

Variable	Transformatior	Lag	Lead	Bounds	Min	Max	Mean	SD	Missing
Janvier					1	20.75	11.99	2.496	80/1302
Février					0	24.25	12.2	2.842	57/1302
Mars					5.75	27.5	13.62	2.784	60/1302
Avril					0	34.1	15.32	2.699	74/1302
Mai					10	29.7	18.2	2.627	52/1302
Juin					0	49.1	21.73	2.798	80/1302
Juillet					19.5	34.5	24.68	2.398	76/1302
Aout					19	37.95	25.44	2.29	50/1302
Septembre					0	47.2	23.52	2.605	42/1302
Octobre					11.6	31.5	20.49	2.979	37/1302
Novembre					0	32.35	16.09	2.962	21/1302
Décembre					-2.4	22.1	10.51	3.453	94/1302

FIGURE 3.1 – Données statistiques de la série 1973-2014

La carte des données manquantes enregistrées dans la série des températures moyennes est illustrée dans la figure 3.2, la couleur rouge indique les données observées, et la couleur crème représente les données non attribuées (manquantes).

Les mois sont stockés selon un ordre décroissant, allant des mois présentant le plus de données manquantes aux mois là ou il y a le moins de données manquantes.

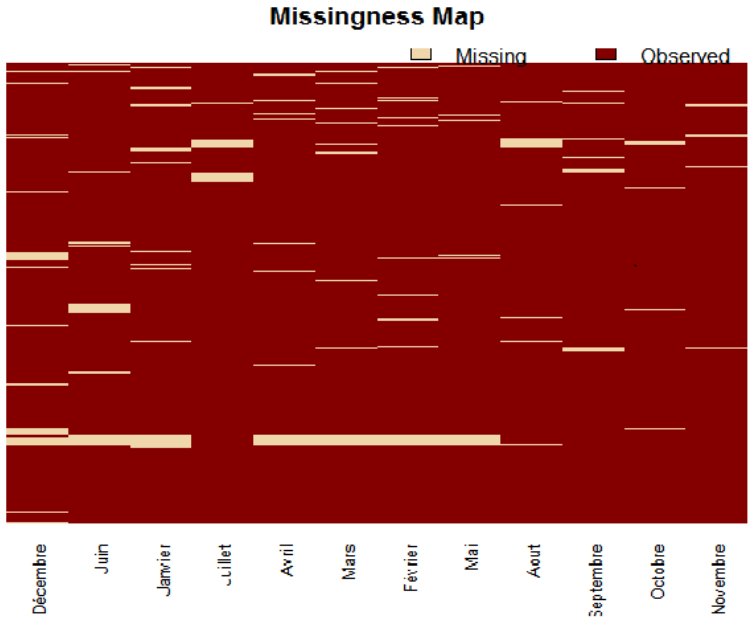


FIGURE 3.2 – Carte des données (observées et manquantes) 1973-2014

Variable	Transformation	Lag	Lead	Bounds	Min	Max	Mean	SD	Missing
janvier					0	133.3	3.186	8.782	177/1292
fevrier					0	64.6	2.56	6.661	175/1292
mars					0	104.1	2.047	6.543	158/1292
avril					0	96.9	1.825	6.743	154/1292
mai					0	48.5	1.159	4.336	109/1292
juin					0	92.8	0.4746	3.635	79/1292
juillet					0	71.8	0.2204	2.679	60/1292
aout					0	40.3	0.3354	2.303	45/1292
septembre					0	100	1.491	7.249	88/1292
novembre					0	83.8	2.094	7.334	95/1292
octobre					0	115.1	2.724	8.452	110/1292
decembre					0	138.2	3.26	9.191	197/1292

FIGURE 3.3 – Données statistiques de la série 1973-2014

3.1.2 Précipitations

Pour les précipitations journalières, qui sont mesurées durant la période 1973 – 2014 en mm/m^2 (millimètre/mètre carrée), le pourcentage des données manquantes des précipitations est un peu élevé par rapport à la série de la température, qui est de 9.72%. La figure 3.3 illustre les différentes valeurs essentielles pour la procédure de comblement.

Avec ce pourcentage des données non attribués, la méthode de l'imputation multiple est la plus adéquate comme le montre le tableau 2.4 .

La carte des données manquantes enregistrées dans la série des précipitations est illustrée dans la figure 3.4 , la couleur rouge indique les données observées, et la couleur crème représente les données non attribuées (manquantes).

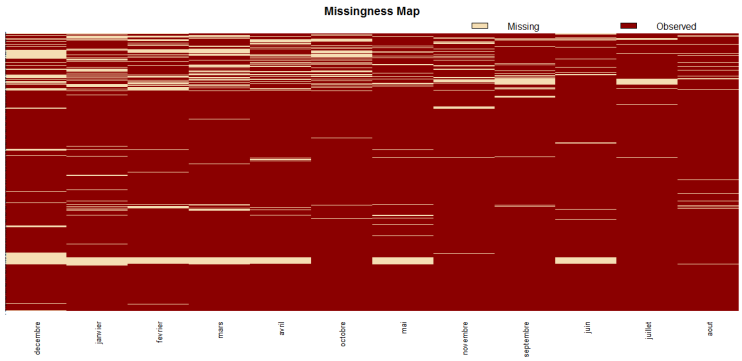


FIGURE 3.4 – Carte des données de la pluviomètrie (observées et manquantes) 1973-2014

3.2 Coefficient de Corrélation

3.2.1 Analyse de liaison

Le coefficient de corrélation permet de mesurer l'intensité de la liaison entre deux caractères quantitatifs. C'est donc un paramètre important dans l'analyse des régressions linéaires (simples ou multiples). En revanche, ce coefficient est nul ($r = 0$) lorsqu'il n'y a pas de relation linéaire entre les variables (ce qui n'exclu pas l'existence d'une relation autre que linéaire). Par ailleurs, le coefficient est de signe positif si la relation est positive (directe, croissante) et de signe négatif si la relation est négative (inverse, décroissante).

Ce coefficient varie entre -1 & $+1$; l'intensité de la relation linéaire sera donc d'autant plus forte que la valeur du coefficient est proche de $+1$ ou de -1, et d'autant plus faible quand il est proche de 0.

1. Une valeur proche de $+1$ montre une forte liaison entre les deux caractères. La relation linéaire est ici croissante (c'est-à-dire que les variables varient dans le même sens).
2. Une valeur proche de -1 montre également une forte liaison mais la relation linéaire entre les deux caractères est décroissante (les variables varient dans le sens contraire).
3. Une valeur proche de 0 montre une absence de relation linéaire entre les deux caractères.

C'est une technique qui permet d'étudier la relation qui pourrait exister entre deux séries d'observation (X, Y) de facteur distinct : Corrélation positive, c'est-à-dire à toute augmentation au niveau de l'observation X correspond une augmentation au niveau de l'observation Y. Les deux variables varient dans le même sens et avec une intensité similaire.

Exemple : La taille et le poids sont en corrélation négative, c'est-à-dire à toute augmentation au niveau de X correspond une diminution au niveau de Y. Les deux variables varient dans deux sens opposés et avec une intensité similaire.

3.2.2 Analyse graphique

L'analyse graphique est une bonne manière de comprendre les différentes caractéristiques énumérées ci-dessus. Le graphique "nuage de points" est l'outil privilégié numéro 1 pour bien interpréter les résultats.

Nous plaçons en abscisse la variable, en ordonnant la variable, chaque observation est positionnée dans le repère ainsi constitué. L'intérêt est multiple : nous pouvons situer les proximités entre les individus ; étudier la forme globale des points, voir notamment s'il existe une forme de liaison ou de régularité ; détecter visuellement les points qui s'écartent des autres, les observations atypiques ; vérifier s'il n'y a pas de regroupement suspects, laissant entendre qu'il y a en réalité une troisième variable qui influence le positionnement des individus ... etc.

Dans la figure 3.5, nous illustrons quelques types de liaisons qui peuvent exister entre deux séries :

3.2.3 Coefficient de corrélation de Bravais-Pearson

Le coefficient de corrélation de Bravais-Pearson est un indice statistique qui exprime l'intensité et le sens (positif ou négatif) de la relation linéaire entre deux variables. C'est une mesure de la liaison linéaire, c'est-à-dire de la capacité de prédire une variable X par une autre Y à l'aide d'un modèle linéaire.

Le coefficient r de Bravais-Pearson entre deux variables X et Y se calcule à partir de la covariance et des écart-types $\sigma_x\sigma_y$ en appliquant la formule suivante :

$$[r_{xy} = \frac{COV_{xy}}{\sigma_x\sigma_y}]$$

Avec :

$$COV_{X,Y} = \frac{\sum(X_i - \overline{X})(Y_j - \overline{Y})}{n}$$

3.3 Analyse des résidus

On estime l'erreur de l'ajustement par les résidus, si l'écart entre la série de référence (origine) et l'autre série imputée à base de la première, et à partir de la représentation graphique des résidus qui nous donnent la possibilité de déterminer la relation entre les séries.

On va marquer la tendance des résidus qui nous permet d'identifier la relation entre elles, si la tendance est sur l'abscisse des X i.e. $Y = 0$, donc il existe une liaison forte et l'erreur est trop petite, mais s'il y a une tendance positive ou négative, donc pas d'existence de liaison.

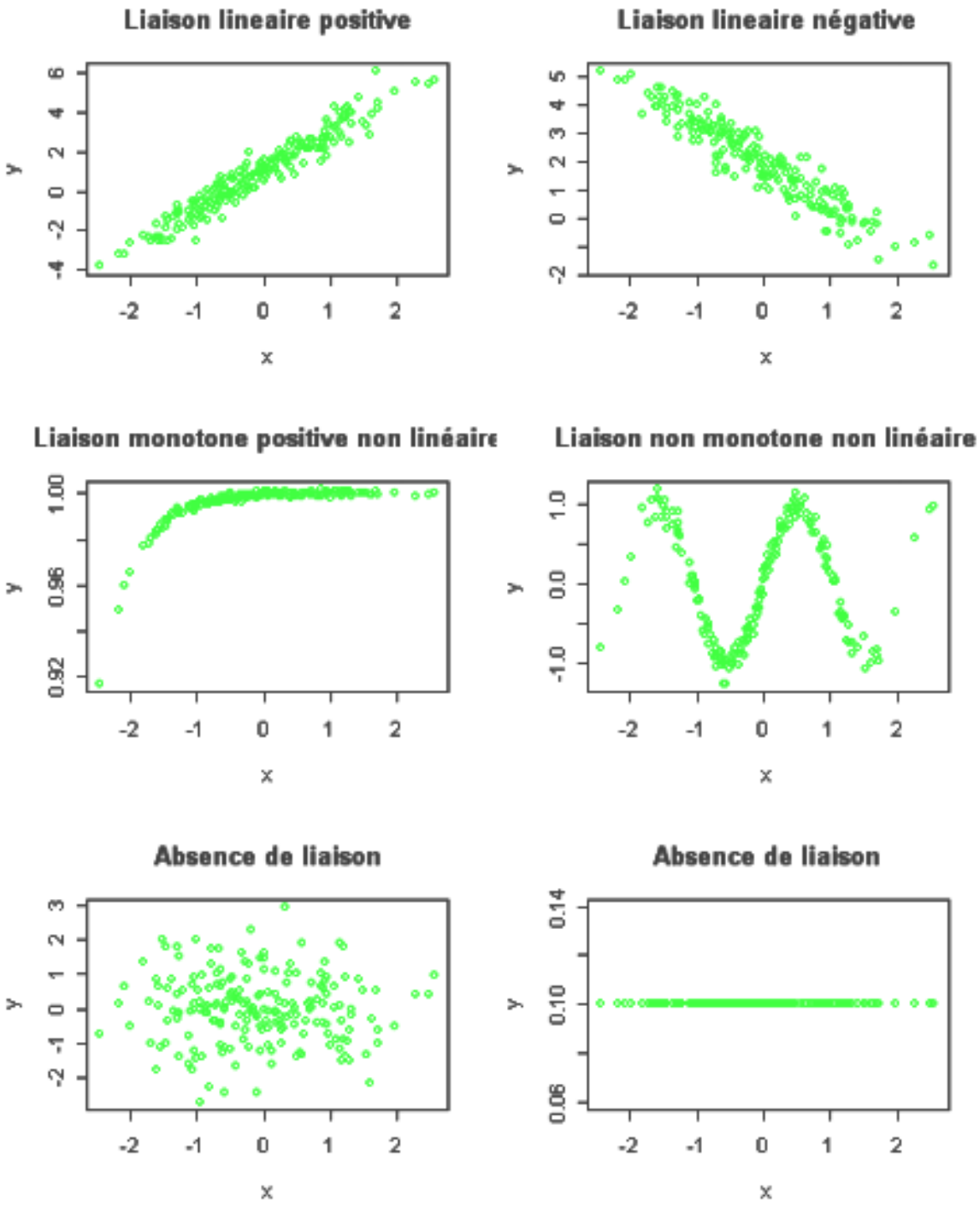


FIGURE 3.5 – Quelques liaisons entre deux séries d'obsèrvations

3.4 Intèrprétation

Le coefficient de corrélation sert avant tout à caractériser une relation linéaire positive ou négative. Il s'agit d'une mesure symétrique. Plus il est proche de 1 (en valeur absolue), plus la relation est forte. $r = 0$ indique l'absence de corrélation. Et avec les résidus, notre étude est analysée avec la variation de la tendance.

A partir des résidus et du coefficient de corrélation, on a confirmé la validité de l'imputation de notre jeu de données (Températures-Précipitations). Pour la température, le pourcentage des données manquantes est de 4.7% qui n'excède pas cinq pourcent (5%), par rapport à la norme. Avec ce pourcentage, les données reconstitués sont valables pour n'importe quelle méthode utilisée.

Par contre, pour les précipitations, le pourcentage de données manquantes est environ 9.7%, pour valider cette base, on a pris une série de 12 ans complète, on a enlevé le même niveau des lacunes (données manquantes) observées, puis on a passé a l'imputation de cette base.

Donc, on aura deux séries de données, la série de référence et la série imputée, et pour valider cette méthode pour ces données, on a passé au coefficient de corrélation et l'analyse des résidus. La figure 3.6 montre la représentation graphique des 2 séries de données.

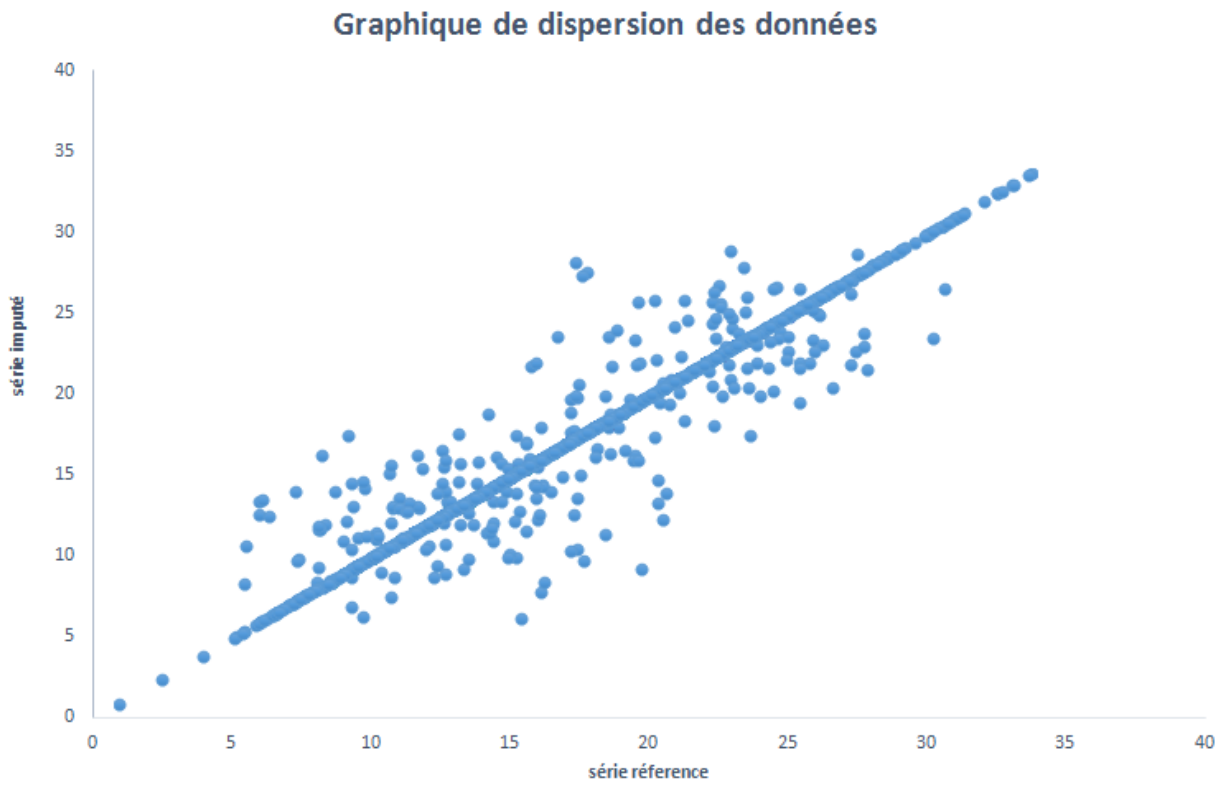


FIGURE 3.6 – Représentation graphique de la corrélation entre la série de référence et la série imputée

3.4.1 Coefficient de corrélation

Les résultats obtenus à partir des données ne sont pas compatibles avec une absence de corrélation, autrement dit, la corrélation est significative.

La figure [3.5] représente le nuage de points qui montre une liaison linéaire positive et très affine. Avec un coefficient de corrélation $r = 0.987$.

D'après la figure 3.6, il est démontré une étroite ressemblance entre les données de référence et celles comblées, d'ou l'efficacité de la méthode adoptée dans la démarche de comblement des manques.

La tendance des résidus est parallèle à l'axe $y = 0$.

3.4.2 Les résidus

La figure 3.7 montre la représentation des deux séries (référence et imputé) et la tendance linéaire des résidus.

Les résidus avec leurs tendance, d'après la graphe, leur tendance est sur l'abscisse des X (linéaire), ça signifie que les deux séries sont en corrélation forte. De plus, le calcul de la moyenne des erreurs estimées $M = 0.1605$, à partir de la moyenne obtenue. Ce résultat montre

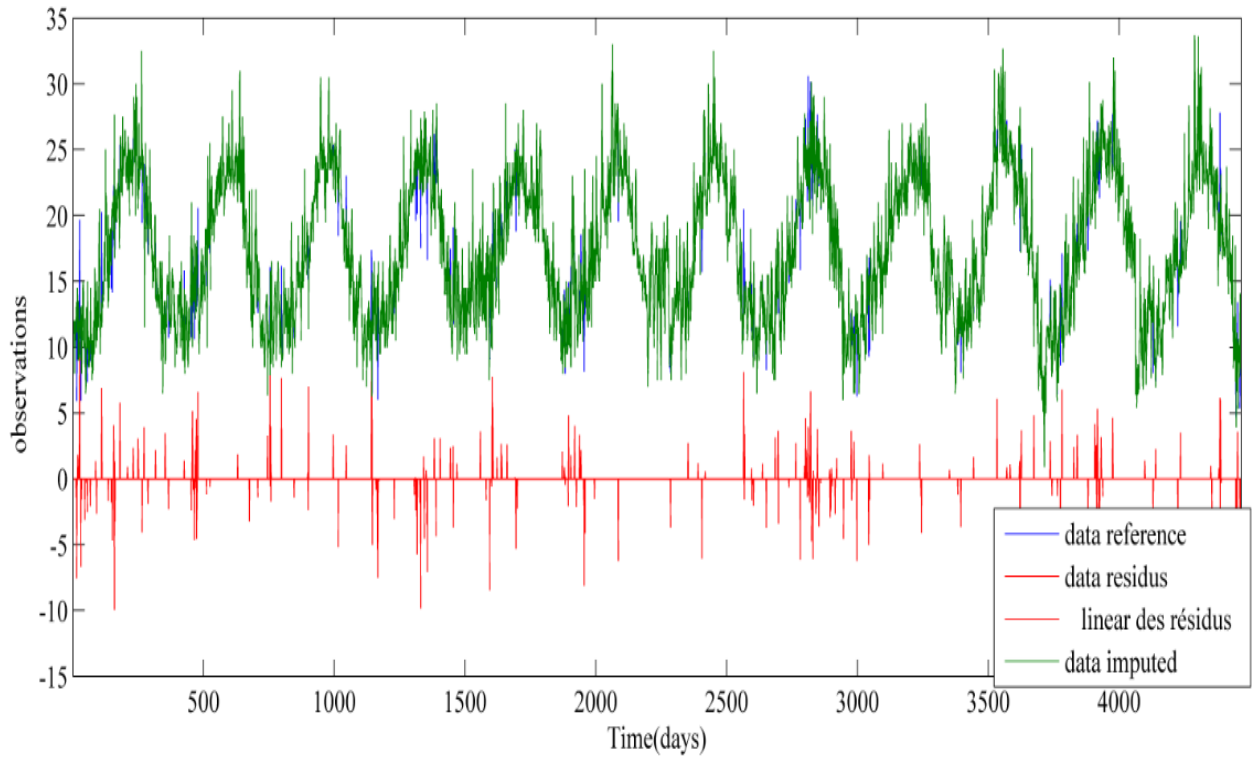


FIGURE 3.7 – Représentation graphique de la série de référence et la série imputée

la variation des erreurs par rapport à l'axe des X, qui sont proche de zéro et dans un intervalle restreint, qui confirme les résultats obtenue avant.

D'après les deux tests effectués, la base de données imputée obtenue est valide et fiable. **Remarque** :La confirmation de ces résultats est faite aussi avec d'autre tests de corrélation comme Spearman au seuil de $\alpha = 0.05$, avec un coefficient de corrélation $r = 0.988$

Chapitre 4

Analyse statistique des données climatiques

4.1 Généralités sur l'analyse statistique des données climatiques

Depuis ces dernières décennies, les différents Etats accordent un intérêt croissant au changement climatique qui demeure une notion très vaste (réduction de la pluviométrie, augmentation de la température, effet de serre,...

Plusieurs manifestations climatiques récentes de grande ampleur ont poussé la communauté mondiale à s'intéresser aux changements climatiques et à leurs impacts sur les ressources en eau. Parmi elles, on peut citer la sécheresse qui a affecté les pays du Maghreb, l'Algérie en particulier, depuis les années 1970. L'Algérie a connu durant son histoire de nombreuses périodes de sécheresses d'ampleur variable. Certaines ont eu des répercussions parfois dramatiques sur les conditions de vie de la population, notamment rurale ($1943 - 1948$) d'ou la répartition du climat de la région en uniquement deux saisons, une sèche et une autre pluvieuse comme cité par les travaux d'Auguste Sabatier ; les sécheresses les plus sévères et les plus persistantes sont celles relevées durant les années1980 à 1990 où le déficit pluviométrique a été estimé à 50% pour les régions du centre et de l'Ouest de l'Algérie. et à 30% à l'Est. L'année 1988/1989 a été classée comme année sèche pour l'Algérie.

Les domaines bioclimatiques de notre région nord Algérie sont définis selon les indices climatiques qui combine les précipitations et les températures et même les autres facteurs de climat, afin de caractériser le rythme climatique à dominance méditerranéenne régissant le climat du nord-est Algérien. On définit cinq grands types de bioclimats méditerranéens [35] :

- Le domaine humide : ce domaine est caractérisé par une pluviométrie supérieure à 900 mm et une forte humidité de l'air, il est caractéristique de la région littorale .

- Le domaine subhumide : il est caractérisé par une pluviométrie supérieure à 600 mm, il est aussi caractéristique de la région littorale où il partage sa dominance avec le domaine humide bien qu'il soit plus développé que ce dernier.

- Le domaine semi-aride : il est caractérisé par une pluviométrie qui fluctue entre 300 mm et 600 mm, il est localisé au niveau des bassins intérieurs du Tell (Mila-Ferdjioua) et est représenté par une poche au niveau de la région de Guelma, il est nettement développé dans les hautes plaines qui le prolongent vers le Sud.

- Le domaine subaride : il est caractérisé par une pluviométrie inférieure à 350 mm, il est représenté par une bande au Sud du piémont de l'Aurès et Nememcha et qui s'élargit au niveau de la région de M'sila ; cette bande ne comprends pas le Hodna.

- Le domaine aride : il est caractérisé par une pluviométrie inférieure à 150 mm, il fait place au pied des massifs de l'Atlas.

4.2 Repésentation de la zone d'étude

4.2.1 Situation géographique

La région d'étude qui est le bassin versant de la Soummam se situe à environ 230 km à l'est d'Alger et appartient administrativement à la wilaya de Bejaia [33]. Il fait partie du Sahel littoral de l'Algérie, il est limité au Nord par les montagnes de la grande Kabylie (massif du Djurdjura), par la mer méditerranéenne et les chaînes côtières de la petite Kabylie. Au Sud, il est limité par les monts de Hodna.

La superficie du bassin-versant de l'oued Soummam est d'environ 9125 Km^2, avec le numéro 15 dans le répertoire de l'Agence Nationale des Ressources Hydrauliques (ANRH)[34].

4.2.2 Localisation de la station

il existe plusieurs stations météorologiques au niveau de la wilaya de Bejaia, dont Bejaia, Elkseur, Sidi-Aich, Akbou, Tichy, Allaghan, Tazmalt,..., qui appartiennent au bassin, nous avons choisi de travailler sur une seule station météorologique, celle de Bejaia, implantée à l'aéroport de Bejaia.

Elle est située au Nord-est entre les latitudes 36° 43 et 36° 1, 20 Nord et les longitudes 5° 4 et 5° 1, 20 Est et avec une altitude de 6 m. Le nombre de paramètre météorologiques qu'elle peut fournir est important (températures, précipitations, humidité, la vitesse vent, débit, durée d'ensoleillement ...). Dans notre travail on s'intéresse seulement aux séries de températures et de pluie (données journalières) pour cette étude.

D'après l'organisation mondiale de la météorologie (O.M.M), les données d'une station reste valable dans une zone d'étude qui s'étale sur une superficie d'un rayon de 50 km à vol d'oiseau. Dans la figure [4.2], on montre la carte de localisation du secteur étudié.

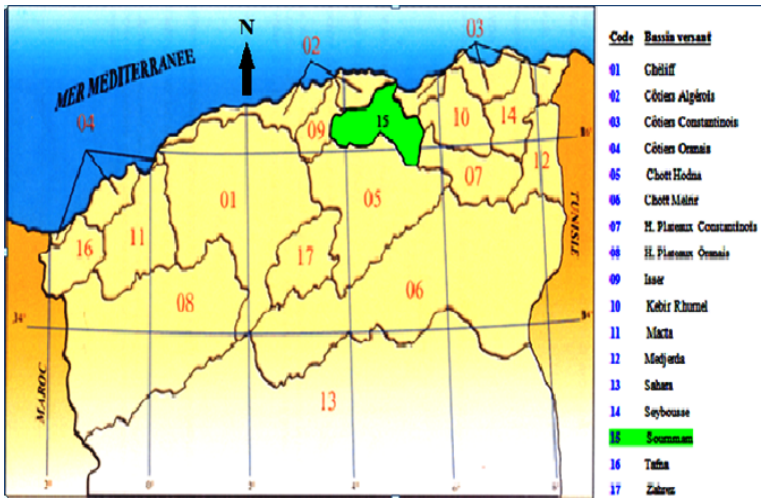


FIGURE 4.1 – Carte du Bassin versant de la Soummam

FIGURE 4.2 – Carte de localisation du Bassin versant de la Soummam

4.3 Les indices climatiques

4.3.1 Indice d'aridité de E.Martonne

L'indice d'aridité de De Martonne (1927) a été tiré à partir de la modification du facteur de pluie de Lang en 1923. Cet indice permet de caractériser le pouvoir évaporant de l'air à partir de la température, selon l'équation suivante :

$$I_{DM} = \frac{P}{T + 10}$$

Où :

- P : les hauteurs annuelles des précipitations en mm ;
- T : les températures moyennes annuelles en $^{\circ}C$;

– 10 : Constante, utilisée pour éviter les valeurs négatives lorsque la température moyenne de l'air est inférieure à $0^{\circ}C$.

Cet indice simple a été largement utilisé par les géographes, il prend des valeurs d'autant plus élevées que le climat est plus humide et d'autant plus faibles que le climat est plus sec.

Pour illustrer les climats existant selon cet indice de De Martonne avec des exemples de région pour chaque type de climat, voir le tableau [4.1] qui donne la classification des climats selon l'indice de DE Martonne.

Indice	Type du climat	Exemple de régions
$0 < I < 5$	Hyper aride	Désests absolus Ex : Reg de Tanezrouft(Sahara), Atacama (Chili)...
$5 < I < 10$	Aride	Région désertiques Ex : le désert de sahara, le désert du Thar(inde)...
$10 < I < 20$	Semi-aride	Le Sahel(Afrique),Chaco(Argentine),Nordeste (Brésil)....
$20 < I < 30$	Semi-humide	La région miditérranéenne...
$30 < I < 50+$	Humide	

TABLE 4.1 – Classification des climats selon l'indice de De Martonne

4.3.2 Indice de quotient pluviométrique d'Emberger

Pour classer et caractériser les climats de différentes régions méditerranéennes, Emberger a défini en 1955 le quotient pluviométrique noté (Q2)[36] combinant trois facteurs climatiques primordiaux (moyenne des températures extrêmes, précipitation et valeur de l'évaporation grâce à l'amplitude extrême $M-m$). Cet indice est précieux pour comprendre les variations floristiques et phytosociologiques de la région étudiée.

Les quotients pluviométriques, s'expriment par la formule suivante :

$$Q2 = \frac{2000P}{M^2 - m^2}$$

M **:** est la moyenne des températures, en Kelvin, du mois le plus chaud ;

m **:** est la moyenne des températures, en Kelvin aussi du mois le plus froid ;

P **:** est la moyenne des précipitations en millimètres.

Le degré de température mésurée en Kelvin, doit être exprimée en degré Celsius(C) telle que : $0^{\circ}C = 273^{\circ}K$ ou $K = {}^{\circ}C + 273$

Le climagramme d'Emberger

Le climagramme du quotient pluviothermique d'Emberger (Q2), illustré dans la figure [4.3], deux grands ensembles peuvent être tirés de cette figure. Le premier sur la partie supérieure du climagramme qui comprend les sites des stations du littoral, dans l'étage bioclimatique humide et sub-humide avec les stations du nord Algérien qui sont situées près de la mér Méditerranée, dans l'étage humide (Jijel) et l'étage sub-humide supérieure (Skikda, Bejaia et d'El-Kala,...), et inférieur (Annaba, Alger,...).

Le deuxième ensemble comprend les stations de l'intérieur qui s'étalent sur le sub-humide moyen au semi-aride inférieur, avec des variantes allant d'un hiver tempéré à un hiver frais. Seule la station de l'Est située dans l'étage saharien supérieur à hiver chaud(Biskra,Tamanraset,...).

Le climagramme d'Emberger indique la migration de la totalité des stations, ces migrations peuvent être soit dans le même étage bioclimatique, soit d'un étage à un autre ou d'une variante hivernale à une autre.

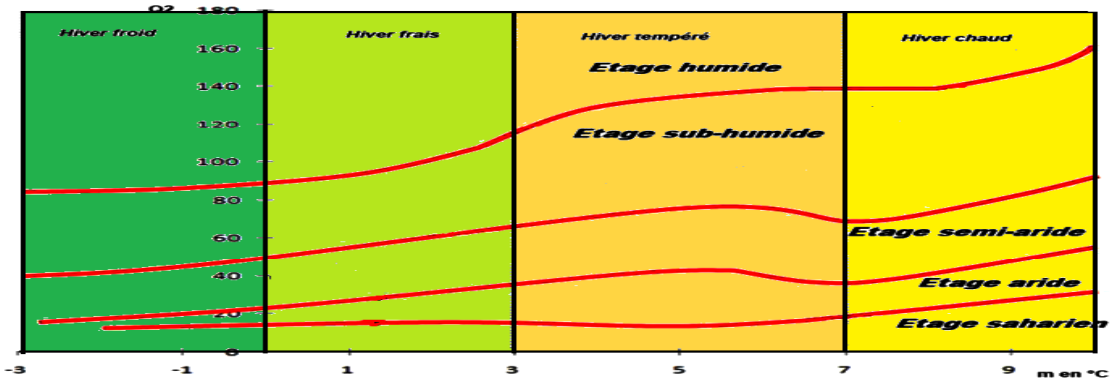


FIGURE 4.3 – Climagramme du quotient pluviothermique d'Emberger

4.3.3 Indice ombrothermique de Gaussen

D'aprés Gaussen (1953), un mois est sec si le quotient des précipitations mensuelles P exprimées en mm, par la température moyenne T exprimée en $^{\circ}C$, est inférieur à 2.

La représentation sur un graphique se fait de cette manière : en abscisse des x nous indiquons le temp (la période d'étude), et à l'axe des ordonnées les précipétations (à gauche) et la températeurs (à droite). Par la suite, on ordonne les températures et les précipitations, ce qui nous permet d'avoir le diagramme ombrothermique qui met immédiatement en évidence les périodes sèches et les périodes pluvieuses.

Les échelles prises sont telles que $1^{\circ}C$ corresponde à $2mm$ de précipitations.

Nous considérons que nous avons une période humide chaque fois que la courbe des précipitations passe au-dessus de la courbe des températures et une période sèche dans le cas inverse.[35]

4.3.4 Indice pluviométrique d'Angot

Cet indice a été proposé par Angot, au début du siècle. Il étudie l'évolution des précipitations au cours d'une année, ainsi que leur répartition saisonnière. Pour cela, il fait intervenir les sommes des précipitations mensuelles qu'il calcule selon la formule suivante :

$$I_a = \frac{\sum P(6moislespluschauds)}{\sum P(6moislesplusfroids)}$$

Où :

– P : précipitations mensuelles en mm.

D'aprés Angot, on aura deux cas, lorsque I_a est inférieur à $1(I_a \leq 1)$, donc la période froide est plus arrosée que la période chaude, et dans l'autre s I_a est supérieur à $1(I_a > 1)$, c'est-à-dire la période chaude est plus arrosée que la période froide.

Cet indice étudie le rapport entre les précipitations de la saison chaude et de la saison froide au cours d'une année (période)[35] .

4.3.5 Indice pluviométrique annuel de Moral

Proposé par Moral en 1964, cet indice est bien adapté pour la classification des climats situés dans la zone intertropicale. Il se calcule, selon la formule suivante :

$$I_M = \frac{P}{T^2 - 10T + 200}$$

Où :

- P : les hauteurs annuelles des précipitations en mm ;
- T : les températures moyennes annuelles en $^{\circ}C$.

A partir de cet indice, on peut classifier le climat selon la variation de la valeur I_M, on distingue deux types de climat :

1. Climat sec, si $I_M < 1$.
2. Climat humide, si $I_M > 1$ [35].

4.4 L'analyse des données Pluviométriques et de Températures

La pluviométrie est la mesure du volume des précipitations en un temps et un lieu donné, l'étude de leurs caractéristiques, de leur répartition, ... Les précipitations sont un peu plus fortes dans les régions de forêts et elles sont rares voir absente dans les zones arides ou sahariennes.

La température est un état énergétique de l'air se manifestant par un échauffement plus ou moins important. Abaissement, changement, différence, élévation,... . Dans cette étude, un choix des températures moyennes journalières a été effectué. Ces températures sont calculées en prenant la moyenne arithmétique des diverses températures extrêmes journalières observées au cours d'une période de $24h$ par les thermomètres à maximums et à minimums.

Notre région d'étude est la région méditerranéenne en général. Elle est sont caractérisée par des précipitations irrégulières et une répartition inégale dans l'espace et dans le temps. Cela peut se vérifier notamment par la nette décroissance des pluies du Nord vers le Sud.

4.4.1 Variabilité de Température

La température est un paramètre important pour l'analyse du climat, de par leur variabilité, et leur répartition spatiotemporelle.

La figure 4.4 montre la variation des températures durant la période 1973/2014 avec les différentes tendances linéaires obtenues par la regression linéaire.

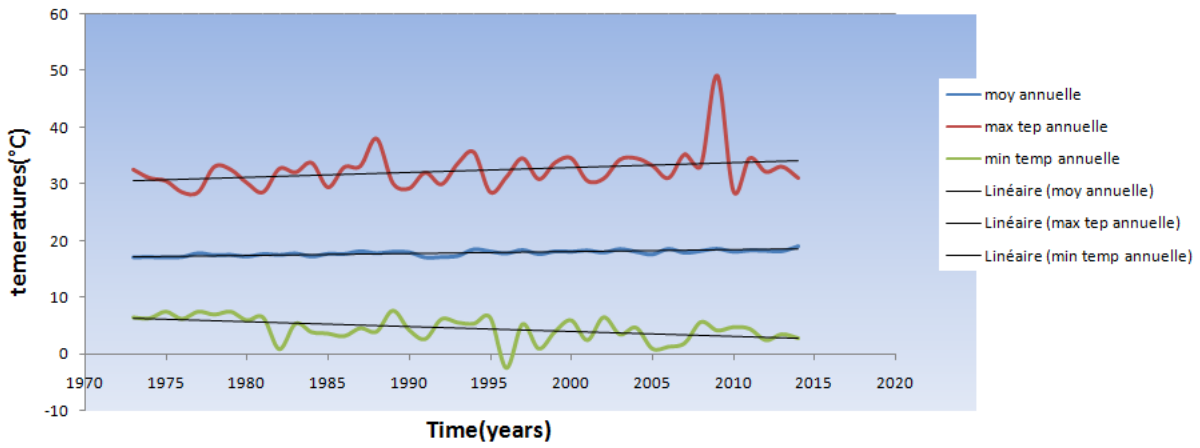


FIGURE 4.4 – Variation des températures annuelles

1. La température Moyenne : La figure 4.4 montre que les températures moyennes n'ont pas vraiment changé, mais elle présentent une legère tendance à l'augmentation tout au long des 42 *ans* derniers.
2. La température Maximale : Les maximas (températures maximales annuelles) ont une tendance remarquable à l'augmentation, ce qui veut dire que le nombre des journées plus chaudes a augmenté.
3. La température Minimale : contrairement aux températures maximales, les minimas (les températures minimales) ont une tendance remarquable à la baisse, ce qui veut dire que le nombre des journées les plus froides a diminué durant les 42 dernières années .

4.4.2 Variabilité des précipitations

La figure 4.5 nous montre aussi les variations des précipitations durant les derniers 42 ans.

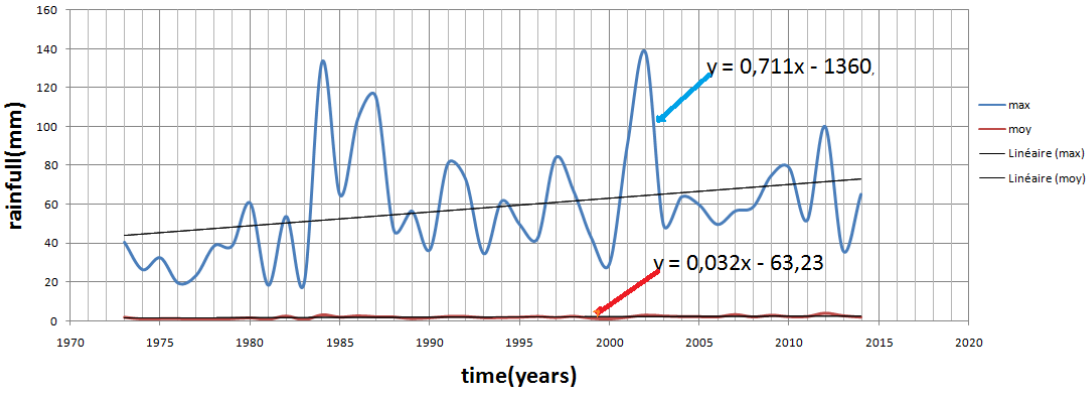


FIGURE 4.5 – Variation des précipitations maximales et moyennes

De même que les température, les précipitations sont un paramètre important pour l'analyse du climat. Mais on remarque que les précipitations minimales mensuelles sont nulles, puisque il existe pas des pluies qui se présentent durant tout un mois. La figure 4.5 montre une augmentation des maximas (précipitations maximales mensuelles) durant les 42 ans, sachant que cette augmentation est de 0.711 mm/an. Mais les précipitations moyennes annuelles présentent une legère augmentation de 0.032 mm/an.

4.4.3 Régime saisonnier

L'importance de la variation saisonnière des précipitations et des températures, concorde avec son rôle primordial de régisseur des secteurs sensibles telles que : les activités agricoles et le mode de vie.

Pour mieux comprendre le régime pluviométrique et température saisonnier, il faut d'abord répartir les années et les diviser en quatre trimestres, de sorte que les mois initiaux de chaque trimestre contienne soit un solstice, soit un équinoxe [32].

Cette méthode définit quatre saisons de la manière suivante :

1. La saison hivernal comporte le mois de décembre, janvier et février (DJF) ;
2. Le printemps est déterminé par le mois de mars, avril et mai (MAM) ;
3. L'été comporte le mois de juin, juillet et août (JJA) ;
4. L'automne est défini par le mois de de septembre, octobre et novembre (SON).

Précipitation

Pour les précipitations, mesurer toutes les quantités recueillies durant chaque trimestre de l'année, qui représente le cumul trimestriel de la saison, et l'ensemble de ces cumuls donnent le cumul des précipitations annuel. La figure 4.6 montre la variation des quantités de la pluie durant tout la période enregistrée ($1973 - 2014$).

Variation saisonnière des précipitations

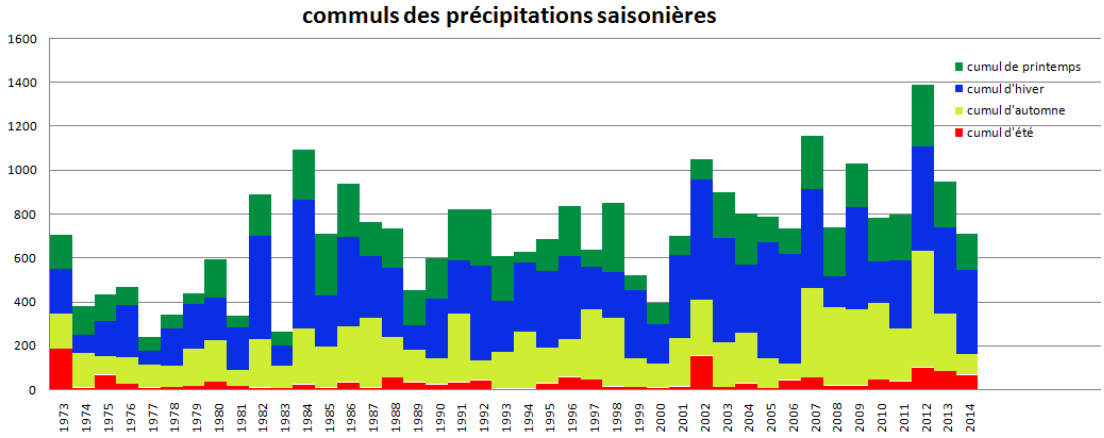


FIGURE 4.6 – Variation saisonnière des précipitations

Interprétation

La figure 4.6, le cumul des précipitations saisonnières confirme l'appartenance de notre région d'étude à l'étage bioclimatique humide. Cela est confirmé par le manque de précipitation et un cumul très réduit de la saison d'été. Par ailleurs, les cumuls d'hiver sont importants affichant des volumes des précipitations importants.

D'après la figure 4.6, on constate une augmentation considérable dans les cumuls saisonniers, l'évolution est significative à partir des années 1990.

Tendances des variations des pluies

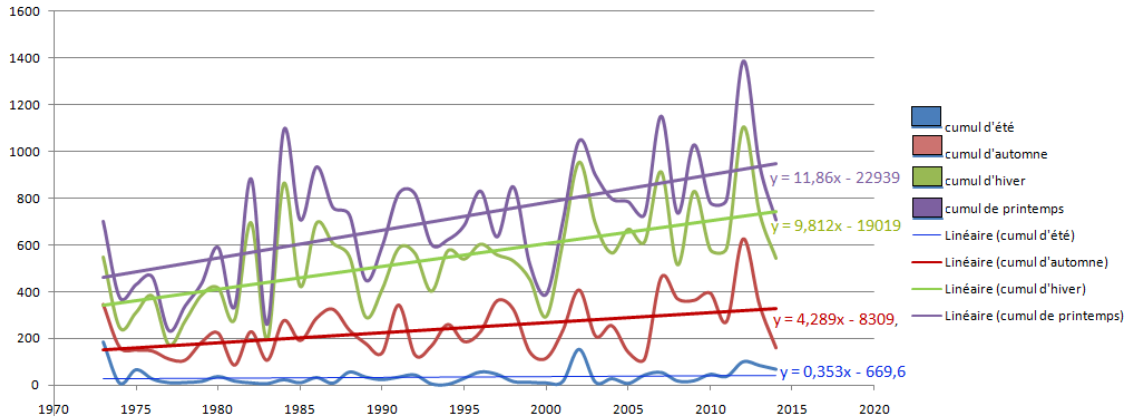


FIGURE 4.7 – Variation des précipitations avec leurs tendances

Interprétation

La figure 4.7, les courbes des cumuls de la pluviométrie et leurs tendances confirment ce qui a été observé par la variation saisonnière des précipitations (figure 4.6). L'augmentation se fait d'une manière très apparente, particulièrement pour les saisons du printemps et d'hiver.

Température

L'obsèrvation des températures durant les saisons de l'année est importante pour étudier le compertement climatique d'une région, et elle permet de juger l'aridité et la sécheressité d'un secteur ou une région. Dans cette analyse, on s'intéresse aux températures moyennes, maximales et minimales. La figure 4.8 montre la variation saisonnière des températures.

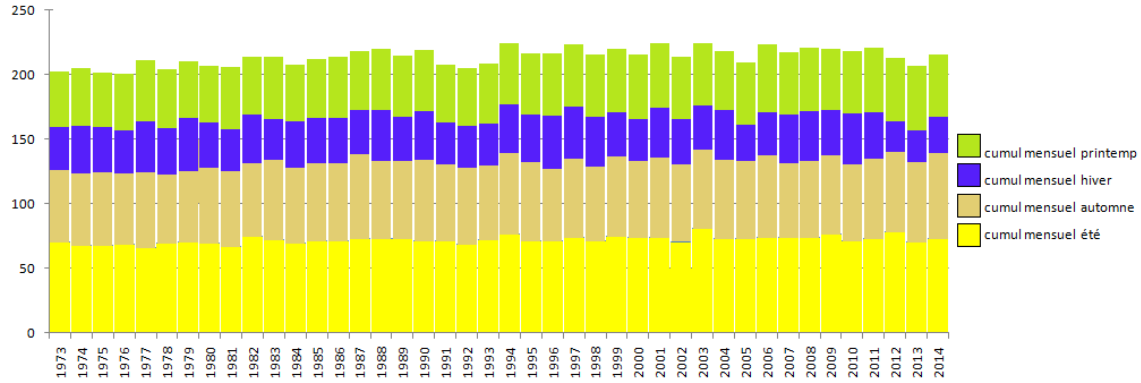


FIGURE 4.8 – Variation saisonnière des températures

Interprétation

La figure 4.8 montre la répartition des températures selon les quatre saisons existantes pour la région du bassin versant de la Soummam. On voit clairement que le cumul de la saison d'été est important. Il vient en second le cumul d'automne, celui du printemps est moins considérable que les deux premiers. Après il vient le cumul d'hiver.

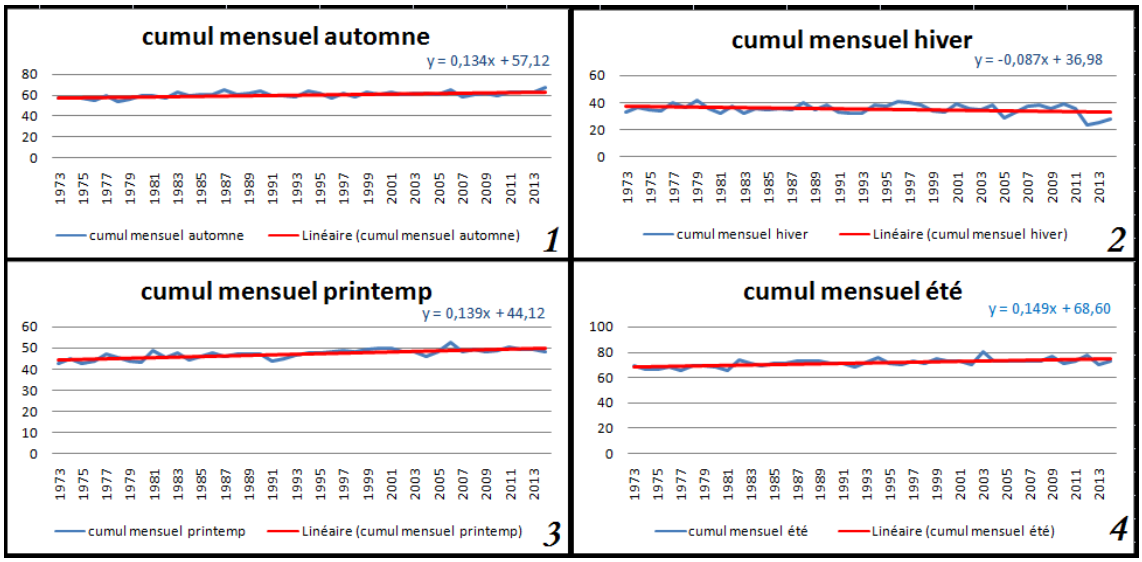


FIGURE 4.9 – Variation des températures avec leurs tendances

Interprétation

La figure 4.9 montre les tendances des variations saisonnières de la température. On voit bien des tendances en augmentations pour les trois saisons l'été, l'automne et le printemps (la partie 1, 3 et 4 de la figure 4.9). Mais une tendance à la baisse du cumul d'hiver qui s'explique dans la partie 2 de la figure [4.9] et cela est due à l'augmentation des cumuls annuels des précipitations (voir la figure 4.6 et la figure 4.7)

4.4.4 Indice de DE Martonne

Avec l'application de cet indice pour la région d'étude "bassin versant de la Soummam", il permet de classifier le type de climat pendant une période d'une année, et cela sera généralisé pour toute la période des 42 *ans*. La figure 4.10 montre la variation du climat et les étages bioclimatiques existants en fonction des deux facteurs climatiques observés (température & pluie).

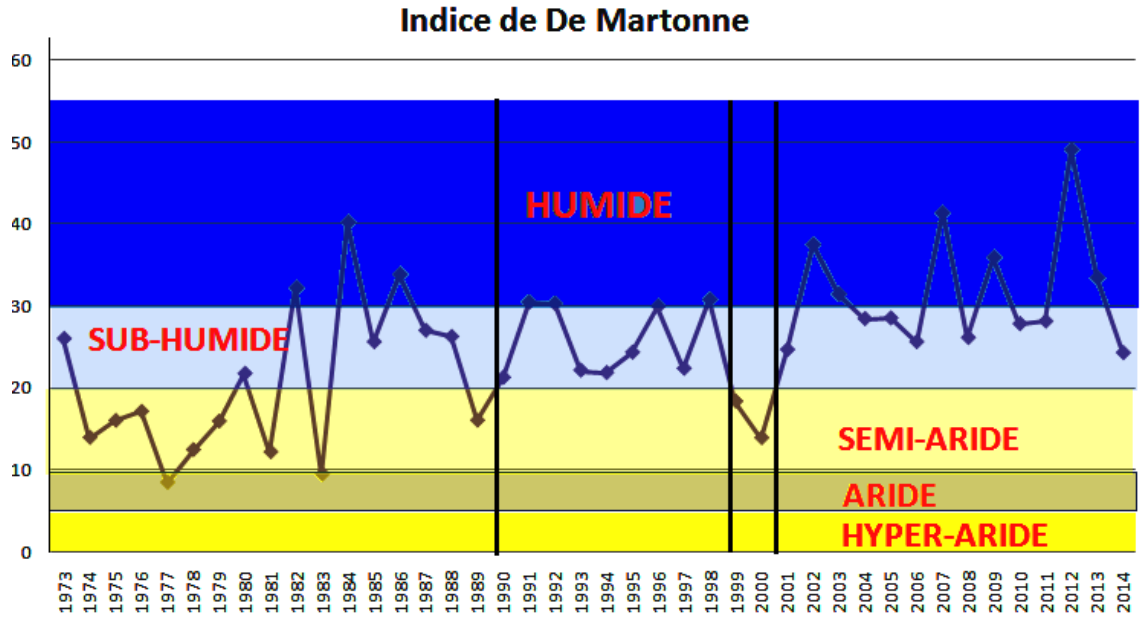


FIGURE 4.10 – Variation de l'indice de DE Martonne pour la période 1973/2014

Interprétation

L'indice de De Martonne a permis de séparer la période d'étude 1973 – 2014 en 4 zones distinctes à comportement pluviothermique distinct.

La période 1973 – 1990

On constate une oscillation faisant varier l'étage bioclimatique de notre zone d'étude l'aride, le semi-aride, subhumide et l'humide. Le comportement climatique pendant cette intervalle enregistre des pics bas et des pics hauts la positionnant dans l'étage humide, les valeurs extrêmes de cette intervalle sont les suivantes :

1. La valeur minimale :8.47 mm correspondante à l'année 1977.
2. La valeur maximale : 40.25 mm correspondante à l'année 1984.

Cette variation est due aux fluctuations de la température et de la pluviométrie.

La période 1990-1999

La zone d'étude durant cette intervalle s'est démarquée par une position bioclimatique subhumide accompagnée d'une timide intrusion dans l'étage bioclimatique humide. Cela s'explique par l'élévation des cumuls pluviométriques illustrés précédemment.

1. La valeur minimale : 18.39 mm correspondante à l'année 1999.
2. La valeur maximale : 30.81 mm correspondante à l'année 1998.

La période 1999-2001

Durant cette période, on observe une chute vers l'étage semi-aride qui s'explique par le déclin du cumul pluviométrique démontré dans la figure [4.7] avec une valeur minimale de 13.93 mm correspondante à l'année 2000.

La période 2002-2014

Le comportement climatique de la région indique une reclassification de la zone dans les étages subhumide et l'humide avec une dominance humide. Cela est due à la baisse de la température durant cette période et aux élévations considérables des cumuls pluviométriques

1. La valeur minimale : 24.32 mm correspondante à l'année 2014.
2. La valeur maximale : 49.12 mm correspondante à l'année 2012.

4.5 L'influence de l'indice Nao au bassin versant de la Soummam

4.5.1 Description

La NAO (Oscillation du Nord Atlantique) est une variation du climat naturelle qui a des impacts importants sur le climat de l'Europe de l'ouest, des environs du Nord de l'Afrique et de l'Est de l'Amérique du Nord. La NAO a des effets bien plus importants en hiver qu'en été. C'est vers 1920 que les deux météorologues, l'autrichien Friedrich et l'anglais Gilbert Walker ont découvert l'Oscillation du Nord Atlantique.

Cette Oscillation a aussi une certaine influence, car elle détermine le positionnement et la trajectoire des dépressions de l'hémisphère Nord. La variation de ce phénomène dépend de la pression atmosphérique.

Comme il est cité dans le premier chapitre, l'indice NAO représente un "paquet climatique" qui permet de mettre en évidence les effets écologiques des fluctuations du climat[12]. Sur la côte nord-occidentale africaine dont la région de la Soummam.

Question

Est ce que le phénomène de la NAO influence réellement le climat du bassin versant de la Soummam ?

Selon les calculs effectués sur les données climatiques (température et pluie), une comparaison effèctuée aux indices NAO, les résultats sont donnés par le tableau 4.1.

	ECR T	I NAO	ECR T min	ECR T max	ECR PL	ECR PL max
ECR T	1					
I NAO	-0.193	1				
ECR T min	0.967	-0.145	1			
ECR T max	0.920	-0.227	0.867	1		
ECR PL	-0.514	0.103	-0.5325	-0.446	1	
ECR PL max	-0.338	0.060	-0.358	-0.269	0.844	1

TABLE 4.2 – La corrélation entre les indices climatiques du bassin de la Soummam et l'indice NAO

- ECR T : Ecart Centré et Réduit de la température moyenne annuelle.
- I NAO : l'indice NAO 1973/2014.
- ECR t min : Ecart Centré et Réduit de la température moyenne annuelle.
- ECR t max : Ecart Centré et Réduit de la température maximale annuelle.
- ECR pl : Ecart Centré et Réduit de la pluviométrie annuelle.
- ECR p max : Ecart Centré et Réduit de la pluviométrie maximale annuelle.

4.5.2 Interprétation des résultats

Le tableau 4.2 montre les différentes corrélations entre l'indice NAO et les indices climatiques calculés pour le bassin versant de la Soummam. Il est clair que les faibles corrélations démontrent l'absence de l'influence de la NAO sur le climat du bassin de la Soummam.

Conclusion générale

Les études sur le climat sont nombreuses et d'une importance capitale pour comprendre les anomalies qui se présentent et de prendre les précautions possibles pour éviter les catastrophes naturelles qui peuvent survenir.

Dans ce mémoire, une étude climatique sur le bassin versant de la Soummam est faite soigneusement en commençant par le traitement de jeux de données climatiques concernant les deux paramètres, pluie et température, par un comblement des lacunes en utilisant un outil informatique et statistique. Il s'agit de l'environnement statistique R avec un package spécifique Amelia.

L'obtention des données complètes et fiables nous a permi de comprendre pas mal de choses concernant notre climat local (Soummam), notamment les variations temporelles des paramètres climatiques pris en compte.

Après avoir comblé et analysé statistiquement nos données, il nous a été permis de conclure ce qui suit :
1- Le comblement de données permet de tirer des informations plus précises, sans aucune perte d'information, véhiculée par le manque de données.
2- Le climat local présente une augmentation de la température et une augmentation des précipitations.
3- L'indice de De Martonne (Indice d'aridité), permet d'observer un passage significatif dans l'ordre suivant : aride, semi-aride, subhumide et humide.
4- Les corrélations obtenues entre nos données et ceux des indices positif et négatif de la NAO s'avèrent non significatives, par conséquent, le model global NAO ne présente aucune influence sur le bassin versant de la Soummam.
5- En perspective de ce que nous avons étudié, il sera d'un intérêt capital d'aborder, dans les travaux à venir, les points suivants :

a) Prendre d'autres stations pouvant recouvrir toutes la zone d'étude, et calculer d'autres indices climatiques.

b) Elargir l'étude vers d'autres paramètres, vent, humidité, débit . . . .

c) Aborder des études comparatives aux régions voisines, voir distantes.

Bibliographie

[1] Reynolds, R. W., & Smith, T. M. (1995). A high-resolution global sea surface temperature climatology. *Journal of Climate, 8(6), 1571-1583.*

[2] Lespinas, *F. (2008). Impacts du changement climatique sur l'hydrologie des fleuves côtiers en région Languedoc-Roussillon (Doctoral dissertation, Thèse de Doctorat, Université de Perpignan Via Domitia).*

[3] Magnan, A. (2009). Proposition d'une trame de recherche pour appréhender la capacité d'adaptation au changement climatique. *VertigO-la revue électronique en sciences de l'environnement, 9(3).*

[4] Bitar, G. (2010). Impact des changements climatiques et des espèces exotiques sur la biodiversité et les habitats marins au Liban. *Rapports et procés-verbaux des réunions Commission internationale pour l'exploration scientifique de la Mer Méditerranée, 39, 452.*

[5] Braun, J. M. (1974). *Etude des séries chronologiques multiples par l'analyse des données (No. CEA-R-4561). SIS-74-3082.*

[6] Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology, 163(12), 1149-1156.*

[7] Paradis, E. (2002). R for Beginners..

[8] Gilbert, W., Henrion, D., Bernussou, J., et Boyer, D. (2007). ATOL : outil de contrôle moteur sous Matlab. *In Congres des Doctorants EDSYS, Albi.*

[9] Quarteroni, A. M., Saleri, F., & Gervasio, P. (2011). *Calcul scientifique : cours, exercices corrigés et illustrations en MATLAB et Octave. Springer Science and Business Media.*

[10] Hurrell, J. W. (1995). Decadal trends in the North Atlantic Oscillation : regional temperatures and precipitation. *Science, 269(5224), 676-679.*

[11] Marini, C. (2011). *Etude des causes et effets de la circulation méridienne de retournement Atlantique (Doctoral dissertation, Université Pierre et Marie Curie-Paris VI).*

[12] Stenseth, N. C., Ottersen, G., Hurrell, J. W., Mysterud, A., Lima, M., Chan, K. S., ... & Ådlandsvik, B. (2003). Studying climate effects on ecology through the use of climate indices : the North Atlantic Oscillation, El Nino Southern Oscillation & beyond. *Proceedings of the Royal Society of London B : Biological Sciences, 270(1529), 2087-2096.*

[13] Antoniadou, T., Besse, P., Fougères, A. L., Le Gall, C., & Stephenson, D. B. (2001). L'oscillation atlantique nord (NAO) et son influence sur le climat européen. *Revue de statistique appliquée, 49(3), 39-60.*

[14] Crettaz de Roten, F., & Helbling, J. M. (1991). Une estimation de données manquantes basée sur le coefficient RV. *Revue de statistique appliquée, 39(2), 47-57.*

[15] Roth, P. L. (1994). Missing data : A conceptual review for applied psychologists. *Personnel psychology, 47(3), 537-560.*

[16] Magnani, M. (2004). Techniques for dealing with missing data in knowledge discovery tasks. *Obtido http ://magnanim. web. cs. unibo. it/index. html, 15(01), 2007.*,

[17] Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data. John Wiley & Sons.*

[18] Bennane, Abderrazak (2010) *Traitement des valeurs manquantes pour l'application de l'analyse logique des données à la maintenance conditionnelle. Mémoire de maîtrise, École Polytechnique de Montréal.*

[19] Raymond, M. R. (1986). Missing data in evaluation research. *Evaluation & the health professions, 9(4), 395-420.*

[20] Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8(3), 430-457.*

[21] Raaijmakers, Q. A. (1999). Effectiveness of different missing data treatments in surveys with Likert-type data : Introducing the relative mean substitution approach. *Educational and Psychological Measurement, 59(5), 725-748.*

[22] Rousseau, M. (2006). *L'impact des méthodes de traitement des valeurs manquantes sur les qualités psychométriques d'échelles de mesure de type Likert (Doctoral dissertation, Université Laval).*

[23] Bocquet, J. P., Neto, M. M., da Rocha, M. T., & de Morais, M. X. (1978). Estimation de l'âge au décès des squelettes d'adultes par régressions multiples. *Contribuições para o Estudo da Antropologia Portuguesa, 10, 107-167.*

[24] Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological methods, 6(4), 317.*

[25] Fichman, M., & Cummings, J. N. (2003). Multiple imputation for missing data : Making the most of what you know. *Organizational Research Methods, 6(3), 282-308.*

[26] Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Operations Management, 24(1), 53-62.*

[27] Vellido, A. (2006). Missing data imputation through GTM as a mixture of t-distributions. *Neural Networks, 19(10), 1624-1635.*

[28] (Ruud, P. A. (1991). Extensions of estimation methods using the EM algorithm. *Journal of Econometrics, 49(3), 305-341.*

[29] (Magnani, M. (2003). Technical report on rough set theory for knowledge discovery in data bases. *University of Bologna.*

[30] R. Lo Presti , E. Barca & G. Passarella. Water Research Institute of the National Research Council. *Departement of Bari, Viale F. De Blasio, 5 70123 Bari, Italy.*

[31] Rosselalo Lo Presti and al. *A methodologie for missing data applied to daily rainfall data in the cardelaro river bassin(Italy)*

[32] Meddi, H., & Meddi, M. (2007). Variabilité spatiale et temporelle des précipitations du Nord-Ouest de l'Algérie. *Geographia Technica, 2, 49-55.*

[33] Mouni, L., Merabet, D., Arkoub, H., & Moussaceb, K. (2009). Etude et caractérisation physico-chimique des eaux de l'oued Soummam (Algérie). *Science et changements planétaires/Sécheresse, 20(4), 360-366.*

[34] Ladjal, R. (2013). *Problématique de la mobilisation et de la préservation des ressources hydriques dans le Sersou (Bassin Cheliff amont Boughzoul) (Doctoral dissertation).*

[35] Karim, F. A. Changement climatique ou variabilité climatique dans l'Est algérien.

[36] Boudjedjou L. (2010) ;Etude de la flore adventice des cultures de la région de Jijel.

Résumé

Les études portées sur le climat et ses changements doivent faire appel à des techniques d'analyse très performantes. Dans le cadre de notre étude, nous avons choisi deux facteurs climatiques, la pluviométrie et la température pour évaluer les changements climatiques au niveau du bassin versant de la Soummam. Une série de la pluviométrie et une autre de la température non toutes complètes qui sont traitées en premier lieu par des outils informatiques, statistiques et mathématique (R, Matlab, Excel, XLStat...) afin de les rendre complètes (comblement des lacunes) et prêtes pour l'étude, elles feront objet d'une analyse statistique à haute fréquence.

Après avoir comblé les lacunes, une analyse est apportée pour la validation des méthodes permettant de combler ces lacunes. Les séries de la pluviométrie et de la température sont traitées graphiquement et analytiquement à l'aide des outils informatiques choisis afin de déterminer la nature et le comportement du climat du bassin versant de la Soummam.

La réalisation de cette étude nous permettra de tester l'efficacité de cette analyse statistiques dans le cadre de traitement de données relatives aux études climatiques, comparer les résultats au model climatique NAO influençant le bassin méditerranéen et évaluer l'impact climatique et humain sur le bassin versant de la Soummam durant la période étudiée par rapport à la station météorologique choisie.

Mots clés **:** Pluviométrie, température, données manquantes, comblement, indice NAO, climat, analyse.

Abstract

The studies carried on the climate and its changes need to use high-performance analysis techniques. As part of our study, we chose two climatic factors, rainfall and temperature to assess climate changes in basin of the Soummam. A series of rainfall and another temperature not complete all that is processed primarily by computer tools, statistics, and mathematics to make them complete (filling gaps) and ready, they will subject to a hight frequency statistical analysis.

After having filled the gaps, an analysis is made for the validation of methods for filling these gaps. The series of rainfall and temperature are processed graphically and analytically using the it tools selected to determine the nature and climate of the behavior of the basin of Soummam.

the completion of this study will allow us to test the effectiveness of the statistical analysis in the data processing framework for climate studies. compare the results to the overall NAO model influencing the Mediterranean and assess the climate and human impact on the basin slope of the Soummam during the study period compared to the selected weather station.

Key words **:** Rainfall, temperature, missing data filling, index NAO, analysis.