Régionalisation des normales annuelles des températures en Algérie par la méthode K-means .

Islam BOUSRI [1*] , Amina BOUCETTA [2], Salah SAHABI-ABED [12]

Abstract

L'objectif d'une régionalisation ou du zonage climatique est d'obtenir un découpage d'un territoire en zones homogènes, à l'intérieur desquelles le comportement climatique est similaire au paramètre étudié, ce qui représente une information cruciale pour les différentes applications nécessitant une minimisation de la variabilité spatiale du paramètre analysé.

La régionalisation dispose également d'une grande importance dans le contrôle des modèles de prévision numérique par rapport à l'observation (données climatologiques), puisqu' elle permet de visualiser le comportement du modèle par région homogène en termes de paramètre de découpage. Vu l'importance des données utilisées sur l'étude du paramètre concerné, notre choix dans cette étude s'est porté sur un ensemble de données comprenant les normales climatiques annuelles des températures (minimales, maximales et moyennes), les altitudes, les coordonnées géographiques (latitude et longitude), les distances par rapport à la mer et par rapport à la forêt de l'ensemble des stations du réseau professionnel.

Une régionalisation par rapport au paramètre température a été élaborée par la méthode du Kmeans, où nous avons utilisé la méthode d'Elbow pour la détermination du paramètre k qui va représenter le nombre de régions (clusters).

Ensuite, nous avons réalisé une analyse en composantes principales (ACP) afin de visualiser clairement les séparations des régions sur les principaux axes factoriels.

En final, grâce à cette méthode nous avons réussi à obtenir cinq régions, qui ont été représentées par des nuages de points et visualisées sur la carte de l'Algerie.

Keywords

K-means, La méthode de Elbow , clusters , ACP, Régionalisation, zonage

[1] *Direction de l'Exploitation Météorologique*

[2] *Direction de la Climatologie et de la Coordination Réseaux, Office National de la Météorologie, Dar El Beida, Alger*

*__Correspondant__: bousri.islam@gmail.com

Contents

1. Introduction

L'Algérie étant le plus grand pays de l'Afrique, est caractérisé par trois types de climat : le climat méditerranéen le long de la côte ainsi que la majeure partie nord, le climat de transition sur la bande collinaire et montagneuse du nord, qui est semi aride (modérément pluvieux) et enfin le climat désertique sur la grande surface occupée par le Sahara.

Par conséquent, il est intéressant d'étudier comment le climat et les zones climatiques en Algérie ont évolué dans le passé récent et vérifier leurs déplacements.

L'importance de la régionalisation climatique est diverse. Elle a pu mettre en évidence dans certains pays comme en France par exemple la réglementation thermique de 2012 qui a utilisé la notion de zones climatiques pour déterminer la consommation moyenne d'un bâtiment en énergie. Pour le contrôle des modèles de prévision numérique par rapport à l'observation, la régionalisation nous aidera à visualiser le comportement des modèles dans les régions qui dispose du même climat.

La régionalisation des paramètres météorologiques basée sur les outils statistiques renferme plusieurs approches et méthodes. DeGaetano et Shulman(1990)[1] ont appliqué

une technique de regroupement flexible aux trois premières composantes principales de plusieurs variables climatologiques afin d'identifier les régions de résistance cohérentes des plantes. Fovell et Fovell (1993) [2] ont utilisé K-means afin d'identifier les zones climatiques du territoire des Etats-Unis d'Amérique en fonction de la température et des précipitations.

En Algérie, A. Medjerab et L,Henia 2005 [3] ont réalisé une étude de régionalisation des pluies annuelles sur le Nord-Ouest de l'Algérie en utilisant l'analyse en composante principale (ACP), et S.TAIBI 2013 [4] a fait une évolution et régionalisation des précipitations au nord de l'Algérie sur la période 1936–2009 par l'ACP

En 2020, Wu Zeng a utilisé également l'algorithme de regroupement K-means pour identifier et exprimer l'agrégation régionale des données temporelles et spatiales dans les événements de température extrême à l'échelle nationale, ce qui fournit une nouvelle idée et une nouvelle méthode pour l'étude des événements climatiques extrêmes [5].

Dans la présente étude, nous avons réalisé une régionalisation des températures normales annuelles par la méthode k-means.

Un ensemble d'apprentissage x(1), ... , x(m), l'objectif est de prédire les centroïdes k et une étiquette c(i) pour chaque point de données :

Step 1. Initialiser les centroïdes des clusters $u_1, u_2, \ldots\ldots, u_k$ de façon **aléatoire** ,
Step 2. Répétez jusqu'à la convergence : {

For every i,set

$$c^{i} := \operatorname{argmin} \| x^{i} - u_{j} \|^{2}$$

For each j,set

$$u_j := \frac{\sum_{i=1}^{n} 1\{c^{i}=j\} x^{i}}{\sum_{i=1}^{n} 1\{c^{i}=j\}}$$

Tableau 1: Algorithme de K-means.

2. Aspect théorique

2.1 k-means

Le « k-means » a été utilisé pour la première fois par James MacQueen en 1967[6], bien que l'idée originale a été proposée par Hugo Steinhaus en 1957 [7] c'est l'un des algorithmes de "clustering" les plus populaires. Cet algorithme est basé sur le concept que chaque groupe peut être représenté par son centre.

Dans l'objectif de trouver ces groupes, k-means commence par attribuer aléatoirement des K-centres puis tente de rattacher les points les plus proches à ces centres.

Algorithme K-means :

Les principales fonctionnalités de cet Algorithme sont :

- 1.Il Initialise les centroïdes des clusters.
- 2. Il Affecte les points de données aux clusters dont le centroïde est le plus proche, sur la base de la distance euclidienne.
- 3. Il permet la mise à jour des centroïdes des clusters : Pour chaque cluster, la moyenne des valeurs de tous les points qui y appartiennent, devient la nouvelle valeur du centroïde.

Enfin, la procédure consiste à répéter les étapes 2-3 jusqu'à ce que les centroïdes cessent d'être mis à jour.

2.2 La méthode du coude

L'une des tâches les plus difficiles que nous pouvons rencontrer lors de l'utilisation des K-means est le choix du nombre de clusters. Une des méthodes utilisées pour déterminer la meilleure valeur K est la méthode du coude Elbow qui est basée sur l'idée que pour avoir une bonne qualité d'un cluster, il faut que l'observation attachée à ce cluster soit aussi proche que possible du centre du cluster.

Nous utiliserons la SSE (Sum Squared Error) dans cette méthode pour déterminer la somme de la différence au carré entre le centre du cluster et chaque observation liée à ce cluster.

Il est donc logique de supposer que nous allons essayer de minimiser la SSE.

L'idée de la méthode Elbow est d'exécuter le clustering k-means sur l'ensemble de données pour une gamme de valeurs de k et pour chaque valeur de k de calculer la SSE.

Le nombre optimal de clusters correspond à la valeur minimale optimale de SSE qui représenterait une articulation du coude.

2.3 Analyse en composantes principales (ACP)

L'analyse en composantes principales (ACP) introduite par K. Pearson en 1901 [8] et développée par H. Heotelling en 1933 [9] qui est une méthode très puissante pour explorer la structure d'un grand nombre p de données quantitatives par élément observé. L'analyse de ces données doit prendre en compte leur caractère multidimensionnel et révéler les connexions existantes entre leurs composantes.

Dyer (1975)[10] et Richman (1986) [11] ont démontré l'apport de la procédure de rotation, qui redistribue l'information contenue dans les K premières composantes entre K nouvelles composantes : la composante principale (CP) dérivée d'une ACP avec rotation individualise et stabilise davantage les structures spatiales.

Par conséquent, l'ACP développe de nouvelles variables artificielles et des représentations graphiques qui permettent de visualiser les relations entre les différentes composantes et les relations entre les groupes variables et les

individus.

3. Données utilisées

3.1 Analyse descriptive des variables

Les données utilisées sont des normales climatiques annuelles des températures (min, max et moy), le paramètre altitude et les coordonnées géographiques latitude et longitude, les distances par rapport à la mer et par rapport à la forêt de 60 stations sélectionnées du réseau professionnel de l'Office National de la Météorologie (ONM).

4. Résultats et discussion

4.1 Interprétation de la matrice de corrélation

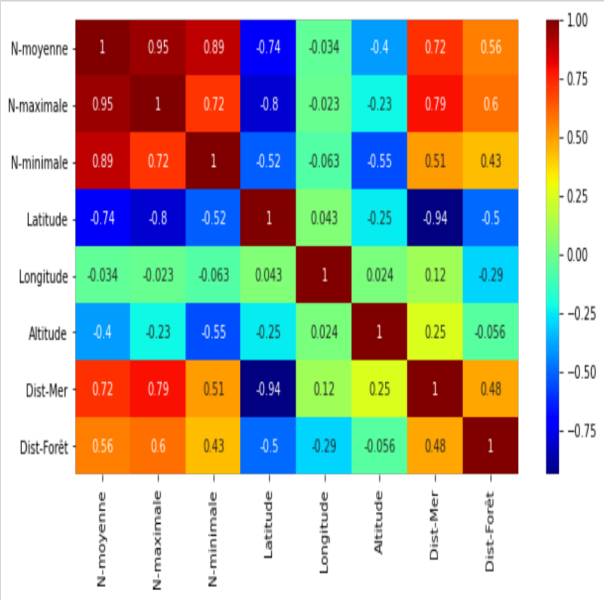

Figure 1. La Représentation de la matrice de corrélation

Nous constatons d'après **la figure 01** que : les normales climatiques de la température **moyenne** sont **fortement et positivement** corrélées avec les températures normales maximales, minimales et la variable distance mer. En revanche, elles sont fortement et négativement corrélées avec la latitude. Elles sont aussi corrélées moyennement négativement avec l'altitude et **moyennement positivement** avec la distance mer.

Concernant les normales maximales, nous remarquons que celles-ci sont **fortement et positivement** corrélées avec les normales minimales, la distance par rapport à la mer et la distance par rapport à la forêt et plutôt négativement fortement corrélées avec la latitude et **moyennement négativement** avec l'altitude.

Sur la figure 01, on constate que les normales minimales présentent une corrélation moyenne positive avec la distance mer, la distance forêt et **moyenne négative** avec la variable altitude. Les latitudes sont **négativement corrélées** avec l'altitude, la distance mer, la distance forêt et présente une corrélation moyenne avec la distance forêt et l'altitude et significative avec la distance mer.

Les longitudes sont **moyennement corrélées** avec les deux distances, **négativement** par rapport à la mer et positivement pour la forêt.

Les altitudes présentent une **corrélation positive** et moyenne avec la distance mer. Les distances mer et forêt sont moyennement positivement corrélées entre elles.

4.2 Synthèse de l'Interprétation de la matrice de corrélation

On remarque dans la matrice de corrélation que toutes les variables sont **fortement** à **Moyennement** corrélées entre elles, soit positivement ou négativement, contrairement à la variable longitude qui ne présente pas de corrélation avec les autres variables sauf une corrélation moyenne positive avec la variable distance mer et négative avec la variable distance forêt. La variable altitude aussi présente des corrélations moyennes avec toutes les autres variables à l'exception des variables distance forêt et longitude qui montrent des corrélations très faibles ou quasi nulles.

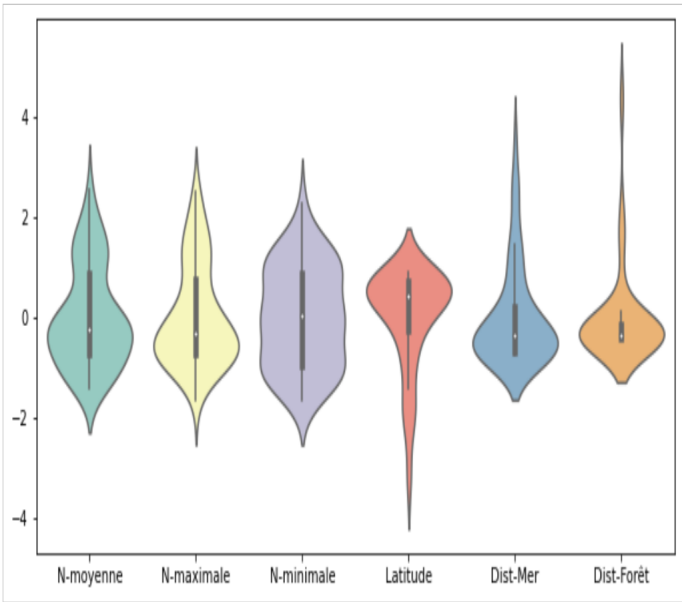

Figure 2. La Représentation de la matrice de corrélation

Les distributions des températures maximales montrent que la queue de haut est plus longue, ce qui signifie la présence d'une **asymétrie positive** ; on voit aussi que la courbe de densité de probabilités des températures moyennes a presque la même allure que celle des normales maximales. Contrairement au minimales, on voit plutôt une symétrie avec une courbe plus plate que la normale.

La distance forêt présente une courbe très pointue, ce qui signifie **une dispersion très faible** avec une queue de haut très longue, cette variable comme le reste des autres par rapport aux stations de notre réseau, la distance forêt

ne varie presque pas sauf pour quelques stations, qui sont généralement celles du Sahara comme elles se trouvent vraiment très loin des forêts.

Pour les courbes des latitudes et des distances mer, on observe pratiquement la même courbe, mais inversée. Cela peut être confirmé dans la matrice de corrélation où on a enregistré des corrélations avec l'ensemble des variables plus au moins les mêmes avec des signes différents. Ce qui est parfaitement logique parce qu'à chaque fois que la latitude diminue en allant vers le sud, la distance par rapport à la Méditerranée augmente à l'exception de quelques stations qui se trouvent au sud-ouest algérien et qui sont plus proches de l'océan Atlantique qu'à la mer méditerranée.

4.3 Application K-means

En mettant dans un graphique les différents nombres de clusters K en fonction de SSE, on retrouve le nombre de classes à sélectionner. En effet, on remarque sur ce graphique,

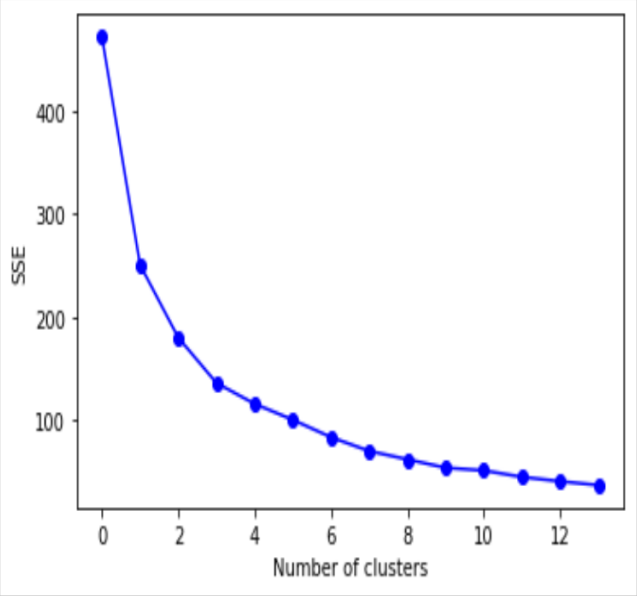

Figure 3. Représentation de la méthode du Elbow

la forme d'un bras où le point le plus haut représente l'épaule et le point où K vaut 13 représente l'autre extrémité de la main. Le nombre optimal de clusters est le point représentant le coude. Ici le coude peut être représenté par un K équivalent à 4 ou 5 ou 6 : Il s'agit du nombre optimal de clusters. Généralement, le point du coude est celui du nombre de clusters à partir duquel la SSE ne se réduit plus significativement. En effet, la chute de la courbe de la SSE entre 1 et 6 clusters est significativement plus grande que celle entre 8 clusters et 13 clusters.

Comme il n'y a pas de solution unique à un problème de classification, nous avons choisi la valeur médiane entre les trois valeurs associées aux coudes 4, 5 et 6.

Nous allons visualiser les régions sur l'ensemble des variables à l'exception de :

Table 1. Répartition des stations par région.

Régions	Nombre de stations
Region1	22
Region2	22
Region3	6
Region4	5
Region5	5

- La normale des moyennes est similaire à la normale des maximales
- La latitude est similaire à la Dist-Mer (distance par rapport à la mer)
- La distance par rapport à la forêt ou Dist-Forêt présente une très faible dispersion ce qui implique que les clusters seront moins visibles

4.4 Application ACP

Pour les valeurs propres, nous avons montré dans la figure 5 les valeurs propres associées à chaque vecteur propre qui représente un axe factoriel comme suite :

[ax1 ,ax2 ,ax3 ,ax4 ,ax5 ,ax6 , ax7 ,ax8]

[0.559 , 0.199 , 0.143 , 0.06 ,0.024 ,0.008 ,0.005 , 0.0]

Concernant le taux d'inertie observée sur le premier plan factoriel, il est à 75.81 % et le taux d'inertie sur un plan factoriel 3D est à 90.16 %.

Une représentation graphique sur un plan factoriel 3D qui combine les trois principaux axes factoriels et qui représente 90.16 % de l'information permet de visualiser la séparation des clusters plus clairement.

Généralement, lorsque nous disposons d'un grand jeu de données contenant des variables continues, une ACP est utilisée pour réduire uniquement la dimension des données avant la classification hiérarchique des données avec l'approche HCPC (Hierarchical Clustering on Principal Components ou Classification Hiérarchique sur Composantes Principales). Dans notre cas, nous avons travaillé avec l'approche K-means sans passer par une réduction des axes.

Comme nous avons utilisé l'ACP seulement pour visualiser plus clairement la séparation des groupes.

D'après la carte de répartition des régions climatiques en Algérie, réalisée avec cette méthode, nous avons constaté l'apparition de cinq différentes régions :

Région 01 : Le Littoral (zone côtière) avec un climat méditerranéen, en plus d'une zone contenue M'Sila, Boussaâda et Barika vu qu'il y a plusieurs paramètres en commun.

Région 02 : Hauts plateaux, caractérisé par des étés plus chauds et moins humides, les températures minimales sont très importantes la nuit.

Région 03 : pré Saharienne, caractérisée par des étés secs et chauds.

Région 04 et 05 : Sahara Orientale, Occidentale, caractérisée par des étés plus secs et plus chauds que la région

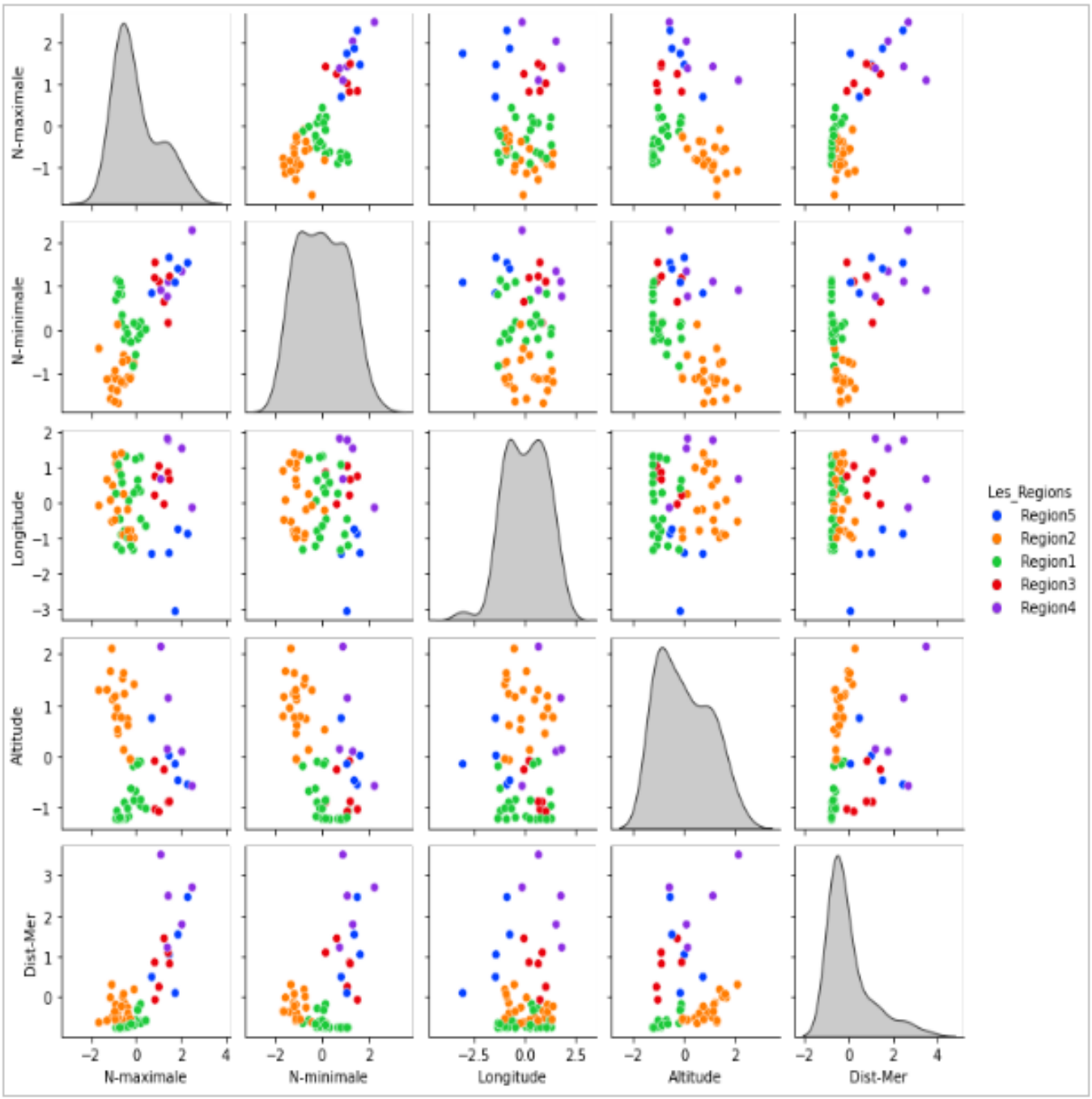

Figure 4. Les différentes régions dans des nuages de points.

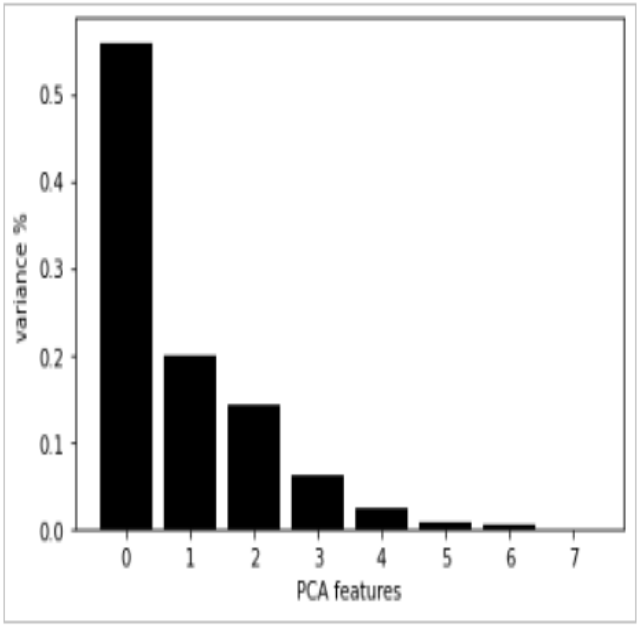

Figure 5. Le pourcentage de l'inertie par axe factorielle

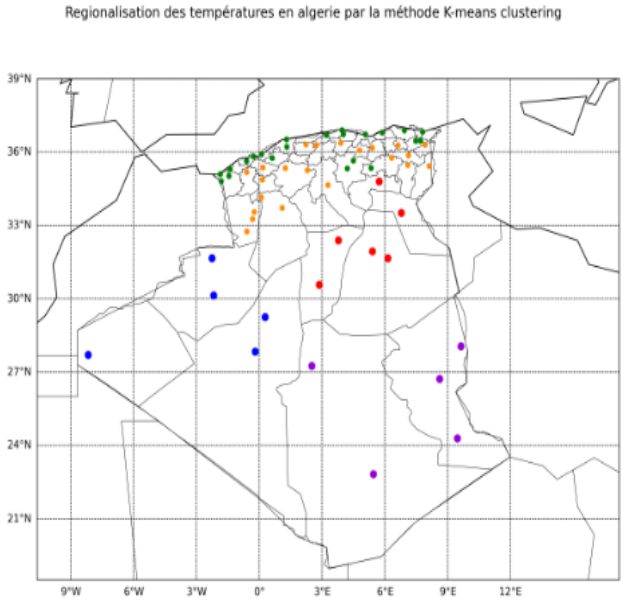

Figure 7. Répartition des régions sur la carte de l'Algérie

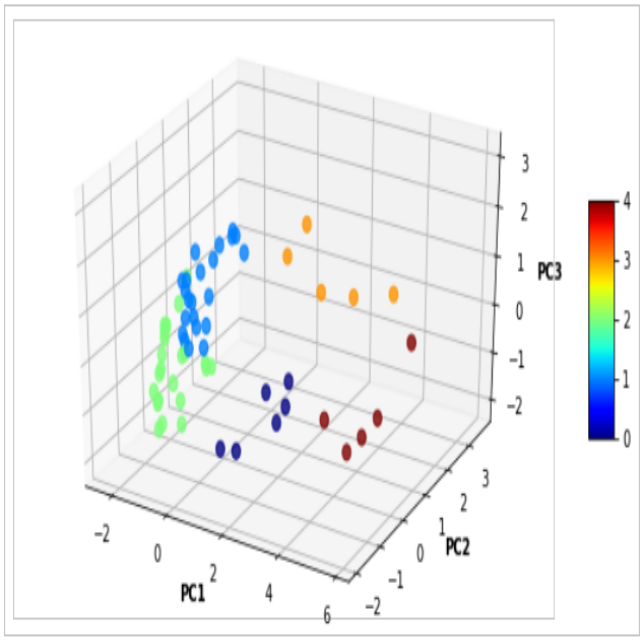

Figure 6. Projection des clusters sur les trois premiers axes factoriels

précédente, avec un climat désertique et sec, une légère différence de température entre ces deux régions, due au paramètre distance à la mer, que nous avons pris en considération dans notre zonage, comme la région 05 est plus proche à l'océan Atlantique.

Ci-dessous un tableau représentatif des stations météorologiques catégorisées selon la région climatique d'après les résultats issus de la classification :

5. Conclusion

Une étude de régionalisation des normales annuelles de température par la méthode du Kmeans a été réalisée. L'utilisation de la méthode de Elbow a révélé l'existence de cinq régions.

Une analyse en composantes principales (ACP) a été réalisée afin de visualiser clairement les séparations des régions sur les principaux axes factoriels.

Nous avons illustré les cinq régions déduites par l'analyse de la méthode appliquée dans cette étude par des nuages de points et sur une carte. Cette régionalisation a permis de mettre en évidence des zones climatiques homogènes du point de vue températures annuelles ; Cette régionalisation présentera un intérêt important pour une utilisation au bénéfice de beaucoup de secteurs socio- économiques (énergie, agriculture…). Il sera également intéressant d'utiliser cette technique pour régionaliser d'autres paramètres météorologiques pertinents (comme les précipitations).

Table 2. Répartition des stations météorologiques et leurs appartenances dans leurs régions climatiques

Région 1	Région 2	Région 3	Région 4	Région 5
ALGER DAR-EL-BEIDA	AIN-SEFRA	BISKRA	DJANET	ADRAR
ANNABA	BATNA	EL-GOLEA	ILLIZI	BECHAR
ARZEW	B-B-ARRERIDJ	EL-OUED	IN-AMENAS	BENI ABBES
BARIKA	BOUIRA	GHARDAIA	IN-SALAH	TIMIMOUN
BEJAIA-AEROPORT	CONSTANTINE	HASSI-	TAMANRASSET	TINDOUF
BENI SAF	DJELFA	MESSAOUD		
BOU CHEGOUF	EL KHEITER	OUARGLA		
BOU SAADA	EL-BAYADH			
CHLEF	KHENCHELA			
DELLYS AFIR	KSARCHELLALA			
GHAZAOUET	MASCARA			
GUELMA	MATEMORE			
JIJEL AEROPORT	MECHERIA			
MAGHNIA	MEDEA			
MOSTAGANEM	MILIANA			
MSILA	NAAMA			
ORAN SENNIA	OUM EL			
RELIZANE	BOUAGHI			
SKIKDA	SAIDA			
TENES	SETIF AIN-SFIHA			
TIZI OUZOU	SIDI BEL ABBES			
TLEMCEN ZENATA	SOUK AHRAS			
	TEBESSA			
	TIARET			

References

[1] Arthur T DeGaetano and Mark D Shulman. A climatic classification of plant hardiness in the united states and canada. *Agricultural and Forest Meteorology*, 51(3-4):333–351, 1990.

[2] Robert G Fovell and Mei-Ying C Fovell. Climate zones of the conterminous united states defined using cluster analysis. *Journal of climate*, 6(11):2103–2135, 1993.

[3] Abderrahmane Medjerab and Latifa Henia. Régionalisation des pluies annuelles dans l'algérie nord-occidentale. *Revue Géographique de l'Est*, 45(2), 2005.

[4] SABRINA Taibi, MOHAMED Meddi, DOUDJA Souag, and Gil Mahé. Évolution et régionalisation des précipitations au nord de l'algérie (1936–2009). *Climate and land surface changes in hydrology, IAHS Publ*, 359:191–197, 2013.

[5] Wu Zeng, YingXiang Jiang, ZhanXiong Huo, and Kun Hu. Clustering analysis of extreme temperature based on k-means algorithm. In *International Conference on Artificial Intelligence and Security*, pages 523–533. Springer, 2020.

[6] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[7] Hugo Steinhaus et al. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci*, 1(804):801, 1956.

[8] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[9] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[10] Thomas GJ Dyer. The assignment of rainfall stations into homogeneous groups: an application of principal component analysis. *Quarterly Journal of the Royal Meteorological Society*, 101(430):1005–1013, 1975.

[11] Michael B Richman. Rotation of principal components. *Journal of climatology*, 6(3):293–335, 1986.