

Université des Sciences et de la Technologie
Houari Boumediene

Master 2 : Systèmes Informatiques Intelligents

Mini-Projet Data-Mining

PARTIE 1 : ANALYSE DE DONNÉES ET
PRÉTRAITEMENT



Réalisé par
DIB Fella
IGHILAZA Lina

Introduction

Comprendre l'interaction entre les propriétés du sol et les conditions climatiques est essentiel pour une agriculture efficace, la préservation de l'environnement et la mise en place de politiques adaptées. L'Algérie, avec sa diversité climatique et son vaste territoire, offre une opportunité unique d'exploiter et d'apprendre de ses données sur les sols et le climat. Ce projet s'appuie sur des techniques de science des données pour explorer, nettoyer et préparer ces données en vue d'analyses avancées, permettant ainsi d'extraire des informations utiles.

Le projet couvre l'ensemble des wilayas du pays, ce qui permet une analyse globale et nationale. Cela aide à identifier des tendances et des modèles qui reflètent les spécificités environnementales de l'Algérie. Les jeux de données incluent **Soil-DATA**, qui détaille les propriétés des sols à travers le pays, et **Climate-DATA**, qui fournit des informations climatiques pour l'année 2019. Des fichiers géospatiaux contenant les limites des pays sont également utilisés pour intégrer et segmenter les données de manière géographique.

La première étape de ce projet consiste à réaliser une analyse exploratoire des données (AED) et à les prétraiter. Cela inclut l'identification des caractéristiques des données, le traitement des valeurs manquantes, la gestion des valeurs aberrantes, ainsi que la transformation des données pour obtenir un format propre, normalisé et cohérent. Ce travail garantit que les données sont prêtes pour des analyses plus complexes, comme le regroupement, la classification ou la recherche de motifs.

Ce rapport détaille les méthodes, outils et techniques utilisés pour analyser et préparer les données. Il présente les résultats principaux de l'AED et du prétraitement.

Description des Données

I. Country Data :

Le dossier contient les fichiers suivants :

- a. **Le fichier projection (prj) :** est un fichier texte utilisé dans les systèmes d'information géographique (SIG) pour décrire le système de coordonnées associé à un jeu de données géospatiales. [1]

Il contient les informations relatives au système de référence de coordonnées (CRS : Coordinate Reference System) qui permet de positionner avec précision des points sur la surface de la Terre. On y retrouve :

- 1- <Geographic 2D CRS: EPSG:4326>
- 2- Name: WGS 84
- 3- Axis Info [ellipsoidal]:
 - lon[east]: Longitude (Degree)
 - lat[north]: Latitude (Degree)
- 4- Area of Use:
 - undefined
- 5- Datum: World Geodetic System 1984
 - Ellipsoid: WGS 84
 - Prime Meridian: Greenwich

1- Le type de CRS :

C'est un système de coordonnées géographiques en deux dimensions (Geographic 2D CRS), qui utilise la latitude et la longitude pour localiser des points sur la surface de la Terre.

2- Le système de référence :

Le CRS spécifié est l'EPSG : 4326, communément appelé WGS 84 (World Geodetic System 1984). Ce système est un standard global largement utilisé pour la cartographie, le GPS, et d'autres applications géospatiales.

3- Les axes de référence :

Deux axes ellipsoïdaux sont définis :

- Longitude (lon) : Mesurée en degrés vers l'est à partir du méridien de Greenwich (0°).
- Latitude (lat) : Mesurée en degrés vers le nord à partir de l'équateur (0°).

4- La zone d'application :

La zone d'utilisation est marquée comme "undefined", indiquant que ce système est applicable à l'ensemble de la Terre.

5- Le Datum¹ :

Il est basé sur le modèle géodésique mondial WGS 84, qui sert de référence pour définir les positions sur la Terre.

- Ellipsoïde : La forme de la Terre est approximée par l'ellipsoïde WGS 84, un modèle mathématique défini par son aplatissement et son rayon équatorial.
- Méridien de référence : Le méridien de Greenwich (longitude 0°) est utilisé comme point de départ pour mesurer les longitudes.

b. Le Fichier Database (dbf) : est un format utilisé pour stocker des données tabulaires. Couramment utilisé dans les systèmes d'information géographique (SIG), il est associé à des fichiers tels que .shp (géométries des entités) et .shx (index des géométries). [2]

Le fichier **.dbf** contient les **attributs** et les **informations descriptives** des entités géographiques. On y retrouve les informations suivantes :

Attribut	Description
Area	Aire de l'entité géographique, probablement en kilomètres carrés ou dans une unité spécifique.
Perimetre	Périmètre de l'entité géographique, probablement dans une unité spécifique (souvent des kilomètres).
CNT1M_1_	Identifiant ou numéro unique attribué à l'entité géographique (interne à la base de données).
CNT1M_1_ID	Autre identifiant unique, parfois utilisé pour regrouper ou différencier des entités similaires.
FAO_NAME	Nom de l'entité géographique selon la FAO (Organisation des Nations Unies pour l'alimentation et l'agriculture).
FAO_CODE	Code FAO associé à l'entité, utilisé pour l'identification dans les bases de données internationales.
UN_CODE	Code numérique des Nations Unies pour l'entité géographique.

¹ Une modélisation de la Terre afin d'exprimer des coordonnées géographiques

ISO_CODE	Code ISO 2 lettres pour identifier le pays ou territoire géographique (par exemple, "GL" pour le Groenland).
CNTRY_NAME	Nom complet du pays ou du territoire en anglais.
ISO3_CODE	Code ISO 3 lettres pour identifier le pays ou territoire géographique (par exemple, "GRL" pour le Groenland).

c. **Le Fichier Index (shx)** : est un fichier utilisé dans le cadre du format Shapefile, qui est largement utilisé pour stocker des données géographiques vectorielles. Il fonctionne en complément des fichiers .shp (qui contiennent les données géométriques) et .dbf (qui contiennent les données attributaires). [3]

d. **Le Fichier Shape (shp)** : composant clé du format Shapefile conçu par ESRI (Environmental Systems Research Institute), est utilisé pour stocker les données géométriques vectorielles des entités géographiques. [4]

Il permet de représenter des formes telles que des points, des lignes ou des polygones, qui décrivent la géométrie des entités spatiales.

Contrairement au fichier .dbf, qui contient les attributs descriptifs, le fichier .shp se concentre exclusivement sur les données géométriques :

Attribut	Description
Geometry	Données géométriques qui définissent la forme spatiale des entités géographiques sous forme de polygones.

Ensemble, ces fichiers permettent de représenter des entités géographiques, leurs formes et leurs attributs, formant ainsi un ensemble cohérent indispensable pour les analyses dans les systèmes d'information géographique (SIG).

II. Soil Data

Le Dataset contient des données relatives aux sols pour les régions géographiques de l'Algérie. Chaque enregistrement est associé à une géométrie représentée par une colonne géospatiale (Polygone). [5]

Propriété	Attribut	Description
Composition texturale du sol	sand % topsoil	Pourcentage de sable dans la couche supérieure du sol.
	sand % subsoil	Pourcentage de sable dans la couche inférieure du sol.
	silt % topsoil	Pourcentage de limon dans la couche supérieure du sol.
	silt % subsoil	Pourcentage de limon dans la couche inférieure du sol.
	clay % topsoil	Pourcentage d'argile dans la couche supérieure du sol.
	clay % subsoil	Pourcentage d'argile dans la couche inférieure du sol.
Propriétés chimiques du sol	pH water topsoil	Niveau de pH mesuré dans l'eau pour la couche supérieure.
	pH water subsoil	Niveau de pH mesuré dans l'eau pour la couche inférieure.
	OC % topsoil	Pourcentage de carbone organique dans la couche supérieure.
	OC % subsoil	Pourcentage de carbone organique dans la couche inférieure.
	N % topsoil	Pourcentage d'azote dans la couche supérieure.
	N % subsoil	Pourcentage d'azote dans la couche inférieure.
	BS % topsoil	Pourcentage de saturation en bases dans la couche supérieure.
	BS % subsoil	Pourcentage de saturation en bases dans la couche inférieure.

	CEC topsoil	Capacité d'échange cationique dans la couche supérieure.
	CEC subsoil	Capacité d'échange cationique dans la couche inférieure.
	CEC clay topsoil	Capacité d'échange cationique spécifique à l'argile dans la couche supérieure.
	CEC clay subsoil	Capacité d'échange cationique spécifique à l'argile dans la couche inférieure
	CaCO ₃ % topsoil	Pourcentage de carbonate de calcium dans la couche supérieure.
	CaCO ₃ % subsoil	Pourcentage de carbonate de calcium dans la couche inférieure.
Propriétés physiques du sol	BD topsoil	Densité apparente (Bulk Density) dans la couche supérieure du sol.
	BD subsoil	Densité apparente dans la couche inférieure du sol.
	C/N topsoil	Ratio carbone/azote dans la couche supérieure.
	C/N subsoil	Ratio carbone/azote dans la couche inférieure.
Géométrie	Geometry	Géométrie des zones de sol, représentée sous forme de polygones décrivant leur emplacement géographique.

III. Climate Data

Un fichier NC (NetCDF : Network Common Data Form) est un format standard utilisé pour stocker des données scientifiques multidimensionnelles. Il est particulièrement adapté pour les domaines comme la météorologie, la climatologie, et d'autres sciences environnementales. [6]

Les données sur le climat de l'année 2019 sont réparties en 72 fichiers avec le format NC. On retrouve :

- 6 observations météorologiques.
- 12 fichiers pour chaque observation représentant les 12 mois de l'année (De Janvier à Décembre).
- Chaque enregistrement représente une observation météorologique associée à :
 - Une latitude,
 - Une longitude
 - Et une heure d'une certaine journée du mois correspondant.

Attribut	Description
Lat	Latitude de l'emplacement observé (en degrés décimaux, où les valeurs positives indiquent l'hémisphère nord et les valeurs négatives, l'hémisphère sud).
Lon	Longitude de l'emplacement observé (en degrés décimaux, où les valeurs positives indiquent l'est et les valeurs négatives, l'ouest).
Time	Timestamp représentant le moment de l'observation
Spatial_ref	Système de référence spatial, utilisé pour définir les coordonnées géographiques de manière précise.
Psurf	Pression de surface. Indique la pression atmosphérique mesurée à la surface terrestre.
Qair	Humidité spécifique. Représente la masse de vapeur d'eau par unité de masse d'air.
Rainf	Flux de précipitations. Mesure l'intensité des pluies.
Snowf	Flux de précipitations solides. Représentant les chutes de neige.
Tair	Température de l'air.
Wind	Vitesse du vent.

Manipulation des données

Environnement de développement

- Le langage de programmation utilisé est Python sous la version 3.9.18
- Les bibliothèques python suivantes ont été utilisées pour la manipulation des différentes données :
 - **Pandas** [7] : est utilisée pour la manipulation et l'analyse des données. Elle offre différents outils de nettoyage, filtrage, transformation et stockage d'ensembles de données.
 - **Dbfread** [8] : Cette bibliothèque lit les fichiers DBF et renvoie les données sous forme de types de données Python natifs pour un traitement ultérieur.
 - **Pyproj** [9] : Effectue des transformations cartographiques. Convertit les coordonnées de longitude et de latitude en coordonnées x,y de la projection cartographique native et vice versa.
 - **Xarray** [10] : est un outil Python conçu pour la manipulation et l'analyse de données multidimensionnelles.
 - **Goepandas** [11] : comme son nom l'indique, étend la célèbre bibliothèque de science des données pandas en y ajoutant la prise en charge des données géospatiales.
 - **Numpy** [12] : est une bibliothèque utilisée pour le calcul numérique. Elle fournit plusieurs fonctions pour les opérations mathématiques sur différentes structures de données notamment sur les tableaux et les matrices.
 - **Scikit-learn** [13] : offre des outils pour la normalisation, la préparation des données et la construction de modèles d'apprentissage automatique.
 - **Matplotlib** [14] : cette bibliothèque est très utile pour la visualisation des données dans les projets d'apprentissage automatique car elle permet de représenter graphiquement les résultats des modèles.
 - **Shapely** [15] : permet de visualiser l'objet au format WKT ou d'importer un objet défini en WKT

Traitement des données :

1. Les données des pays (Country)

- Pour le fichier projection aucun traitement n'est nécessaire.
- Pour les 3 fichiers (database, index et shape) appliquer un filtrage sur l'attribut CNTRY_NAME afin de ne garder que les instances en relation avec l'Algérie.
- Les polygones permettront d'afficher la carte de l'Algérie suivante :

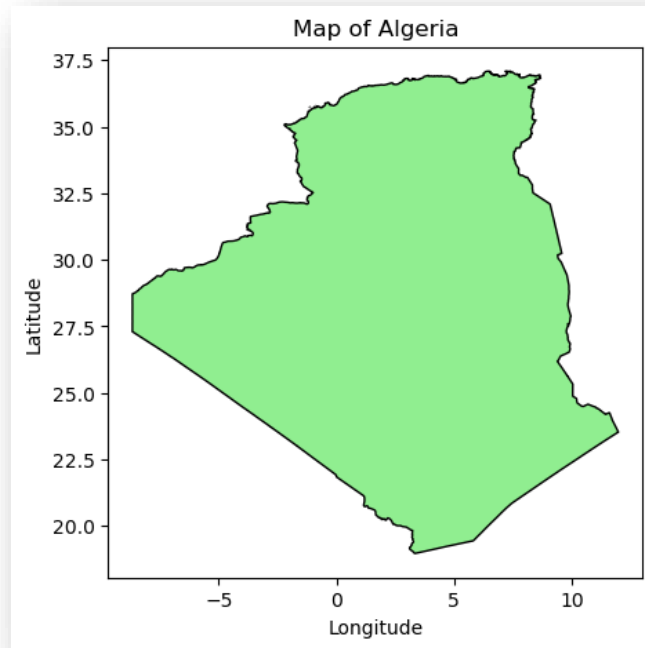


Figure 1 : La carte de l'Algérie

Pour la suite des traitements les fichiers database et index ne seront pas utilisés, car grâce à la bibliothèque Geopandas il est possible d'accéder directement et facilement à toutes les informations à partir du fichier Shape.

2. Les données du climat

Etant données la taille assez volumineuse des données, les étapes suivantes ont été réalisées afin de réduire les données :

Etape 1 : Filtrage

- Filtrer chaque fichier pour ne garder que les données en relation avec l'Algérie. Ceci en faisant correspondre la latitude et la longitude avec le polygone de l'Algérie récupéré à partir du fichier Shape.
- La sortie sera 72 fichiers contenant des informations sur l'Algérie uniquement.

Etape 2 : Conversion

- Pour une manipulation et une visualisation plus facile, les fichiers ont été convertis en un Dataframe et sauvegardés dans un format csv.
- L'affichage des données de l'observation PSurf (Pression de surface) sur la carte de l'Algérie donne le résultat suivant :

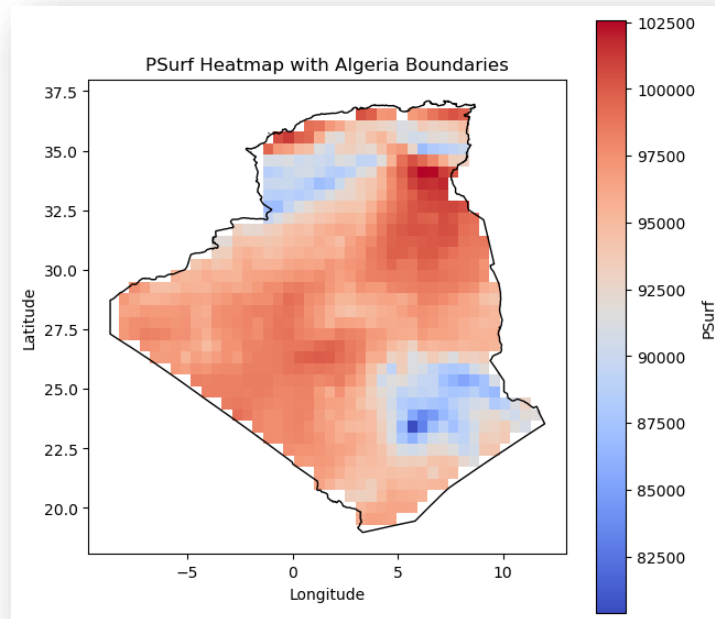


Figure 2 : La pression de surface sur le territoire Algérien

Etape 3 : Fusion et Réduction des données en appliquant l'agrégation par saison.

Cette étape consiste à :

- Fusionner, pour chaque observation météorologique, les 12 fichiers qui la concernent en un seul Dataframe.
- Remplacer la colonne « time » par la saison correspondante :
 - Été : Du 22 Juin au 21 Septembre.
 - Automne : Du 22 Septembre au 21 Décembre.
 - Hiver : Du 22 Décembre au 21 Mars.
 - Printemps : Du 22 Mars au 21 Juin.
- Ne garder que le minimum, le maximum et la moyenne de chaque observation météorologiques associée à chaque latitude, longitude et saison.

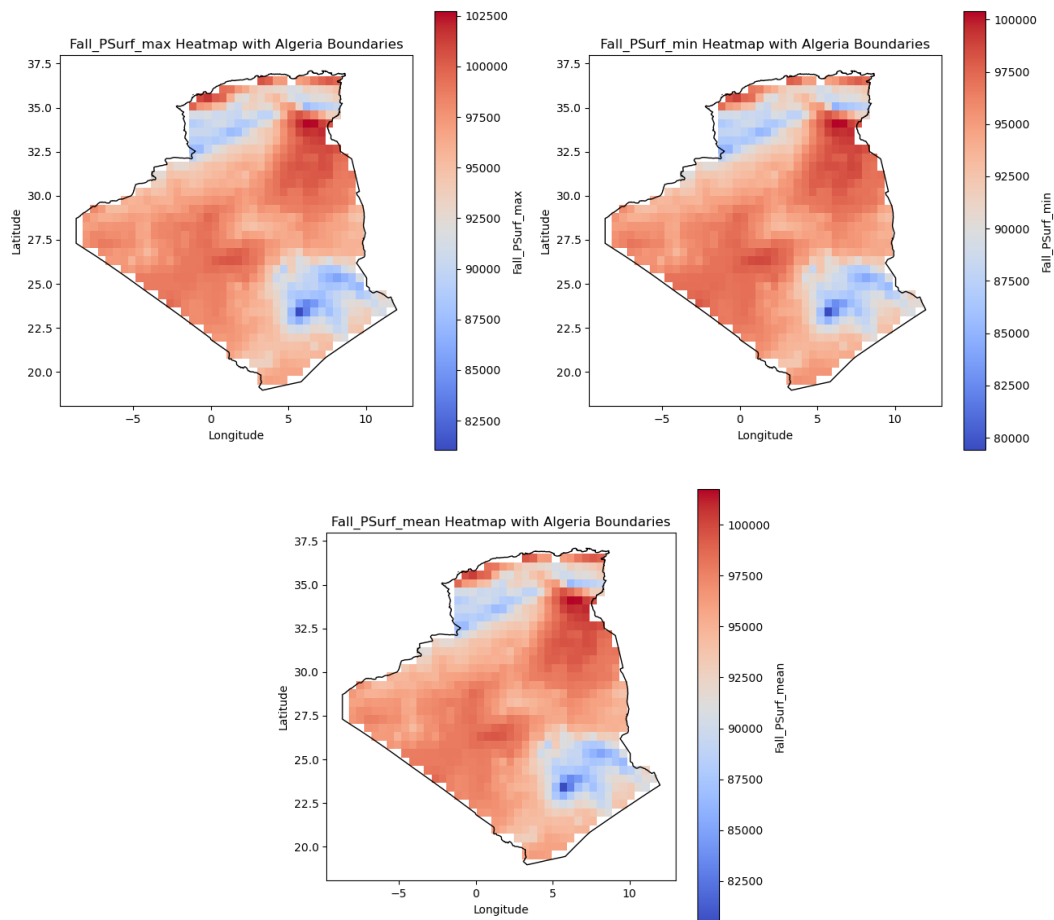
La Sortie sera 6 fichiers représentant chaque observation météorologique (PSurf, Qair, Rainf, Snowf, Tair et Wind) contenant chacun le minimum, le maximum et la moyenne associée à chaque latitude, longitude et saison.

Etape 4 : Intégration et réduction de données.

- D'abord fusionner les 6 fichiers en 1 seul dataframe, par rapport aux attributs latitude, longitude et saison. Chaque observation météorologique sera représentée par 3 colonnes (Exp : PSurf_min, PSurf_max, PSurf_mean).
- Ensuite, pour éviter la répétition des longitudes et latitudes pour chaque saison, répandre cette dernière sur toutes les colonnes des observations.
(Exemple : Summer_PSurf_max, Fall_PSurf_max, Spring_PSurf_max, Winter_PSurf_max ...)

Le Dataset final contient 856 lignes x 74 Colonnes.

L'affichage des données de Fall_PSurf sur la carte de l'Algérie donne le résultat suivant :



Figures 3 : La pression de surface maximum, minimum et moyenne sur le territoire Algérien

3. Les données du sol

Les données du sol sont rassemblées en un seul Dataset au format csv. Ces dernières sont propres et ne nécessitent pas de réduction ou de fusion entre elles. Les prétraitements en commun avec les données du climat seront évoqués par la suite.

Maintenant que chaque jeu de données a le format souhaité, les 2 Datasets Sol et Climat sont fusionnés sur la base de la longitude, latitude pour le climat et la géométrie (Polygone) pour le sol. Le résultat est un Dataset de 856 lignes x 100 colonnes.

Il faut ensuite nettoyer ce nouveau Dataset, en réalisant les tâches suivantes :

1. Le traitement des valeurs manquantes :

Lorsqu'un dataset contient des valeurs manquantes, il est crucial de les traiter pour garantir l'intégrité des analyses. Voici les trois méthodes proposées :

- **Suppression des instances :** Cette technique consiste à éliminer les instances contenant des valeurs manquantes. Elle garantit un dataset sans valeurs nulles et simplifie l'analyse. Cependant, si le pourcentage de données manquantes est élevé, cette approche peut considérablement réduire la densité des données, entraînant une perte d'information critique et affectant potentiellement la représentativité globale du dataset.

- **Remplacement par la moyenne :** Ici, les valeurs manquantes sont remplacées par la moyenne des valeurs observées pour l'attribut concerné. Cela permet de conserver toutes les instances du dataset, minimisant ainsi la perte d'information. Toutefois, cette méthode peut introduire un biais, en particulier si la distribution des données est asymétrique ou contient des valeurs aberrantes.
- **Remplacement par la médiane :** Cette approche substitue les valeurs manquantes par la médiane de l'attribut concerné. Elle est plus robuste face aux valeurs aberrantes que la moyenne, car la médiane n'est pas influencée par les extrêmes. Cependant, elle peut également déformer légèrement les données si la distribution est fortement asymétrique, car la médiane ne reflète pas toujours la tendance globale du dataset.

Après la fusion des datasets Climat et Sol, le dataset résultant contient deux instances (lignes) avec des valeurs manquantes dans toutes les colonnes liées aux données du sol. Cela est dû à l'absence de polygones correspondant aux coordonnées de latitude et de longitude de ces instances.

Étant donné que ces lignes ne représentent que 0,23% des données totales, leur suppression est une solution plus appropriée. Cette approche permet non seulement de maintenir la cohérence des données, mais également de simplifier le traitement ultérieur. En effet, elle évite le calcul de la moyenne sur 25 attributs, ce qui aurait été plus complexe et plus coûteux en termes de temps de calcul.

Ainsi, la suppression des valeurs manquantes est effectuée avant le traitement des outliers, permettant de diminuer le nombre d'instances à traiter.

2. La suppression des Outliers :

La manière de traiter les valeurs aberrantes dépend du but final :

- **Découverte de patterns ou de fonctions prédictives :** Si l'objectif est d'identifier un pattern ou de construire une fonction de prédiction, conserver les valeurs aberrantes risque de biaiser le modèle et d'empêcher d'atteindre des performances optimales. Dans ce cas, leur suppression est recommandée.
- **Détection de comportements inhabituels :** À l'inverse, si l'objectif est de détecter des comportements anormaux, comme des anomalies ou des fraudes, les outliers deviennent des éléments précieux. Ils sont essentiels pour permettre à un algorithme d'apprendre à reconnaître ces comportements.

Dans le dataset climatique, les valeurs de **Snowf** sont en majorité nulles, et les quelques valeurs supérieures à 0 sont considérées comme aberrantes. Cependant, ces valeurs pourraient refléter des événements météorologiques réels et significatifs, tels que des chutes de neige rares, des épisodes exceptionnels ou des conditions spécifiques liées à des altitudes élevées. Leur suppression risque donc de conduire à une perte d'informations sur la variabilité climatique.

Cela dit, dans ce projet, l'objectif est de mesurer les corrélations entre les attributs climatiques et ceux du sol, afin d'identifier des patterns généraux et, potentiellement, de réaliser des prédictions. Dans ce contexte, la présence de valeurs aberrantes peut

altérer la précision des métriques statistiques, notamment en biaisant les estimations des relations entre variables.

Ainsi, la suppression des outliers est une étape indispensable pour garantir la fiabilité des analyses et s'assurer que les résultats obtenus reflètent des tendances globales cohérentes.

3. La suppression des redondances horizontales et verticales :

Cette étape vise à réduire la taille du Dataset en éliminant les données redondantes ou inutiles, ce qui contribue à une analyse plus efficace.

- **La suppression horizontale** (au niveau des lignes) : retire les instances dupliquées ou non significatives, ce qui allège le Dataset et améliore l'efficacité du traitement.
- **La suppression verticale** (au niveau des colonnes) : permet d'éliminer les attributs qui fournissent des informations similaires ou non pertinentes, ce qui simplifie le modèle et réduit la complexité des calculs.

Appliquer cette opération à ce stade permet de diminuer le nombre d'instances et d'attributs à traiter, optimisant ainsi le temps de traitement tout en préservant la qualité et la pertinence des données.

4. La normalisation des données :

La normalisation est une étape clé permettant d'uniformiser les échelles des différentes variables afin de garantir leur cohérence et leur contribution équitable dans les modèles d'apprentissage.

Les deux méthodes présentées sont :

- **Min-max normalisation :**
Cette méthode transforme les données pour qu'elles soient comprises dans une plage définie, souvent entre 0 et 1.
 - **Avantage :** Elle conserve la distribution des données et convient particulièrement aux modèles sensibles aux intervalles des valeurs.
 - **Inconvénient :** Elle est sensible aux valeurs aberrantes, qui peuvent fortement influencer les bornes. Pour cette raison, il est crucial de traiter les valeurs aberrantes avant d'appliquer cette méthode.
- **Z-score normalisation :**
Cette technique centre les données autour de 0 avec un écart-type de 1, en les transformant selon leur écart par rapport à la moyenne.
 - **Avantage :** Elle est moins affectée par les valeurs aberrantes et est idéale pour les modèles où des distributions gaussiennes sont attendues.
 - **Inconvénient :** Elle peut être moins intuitive à interpréter lorsque les données ne suivent pas une distribution normale.

Dans ce projet, étant donné que les valeurs aberrantes ont déjà été supprimées du dataset, la méthode **Min-max normalisation** constitue une approche plus adaptée et pratique pour normaliser les données climatiques et pédologiques.

5. La discrétisation des valeurs continue :

La discrétisation consiste à transformer des variables continues en variables discrètes. Cette étape est particulièrement utile pour simplifier l'analyse et faciliter l'interprétation des données.

Deux méthodes ont été implémentées pour ce projet :

- **Equal-width (ou equal-amplitude) :**

Cette méthode divise la plage des valeurs d'un attribut en intervalles de même largeur. Elle est Simple à implémenter et efficace lorsque les données sont réparties de manière homogène. Cela dit, elle est sensible aux valeurs aberrantes, et peut produire des intervalles avec des densités très inégales, ce qui limite son efficacité pour des distributions déséquilibrées.

- **Equal-frequency :**

Cette méthode répartit les données en intervalles contenant un nombre égal d'observations. Particulièrement adaptée pour des distributions asymétriques ou des données hétérogènes, elle garantit une répartition équilibrée des observations, même en présence de valeurs extrêmes. Cependant, les intervalles obtenus peuvent varier en largeur, ce qui peut compliquer l'interprétation.

Compte tenu des caractéristiques des données climatiques et pédologiques, comme la présence de distributions asymétriques, la méthode equal-frequency est plus adaptée.

Exploration des données

I. Dataset Climat

a. Valeurs uniques et manquantes :

Attributs	Valeurs manquantes	Valeurs uniques
lat	0	36
Lon	0	41
Fall_PSurf_max	0	856
...
Spring_Wind_min	0	856
Summer_Wind_min	0	856
Winter_Wind_min	0	856

Le dataset climat ne contient pas de valeurs manquantes ni de redondances.

b. Box Plots

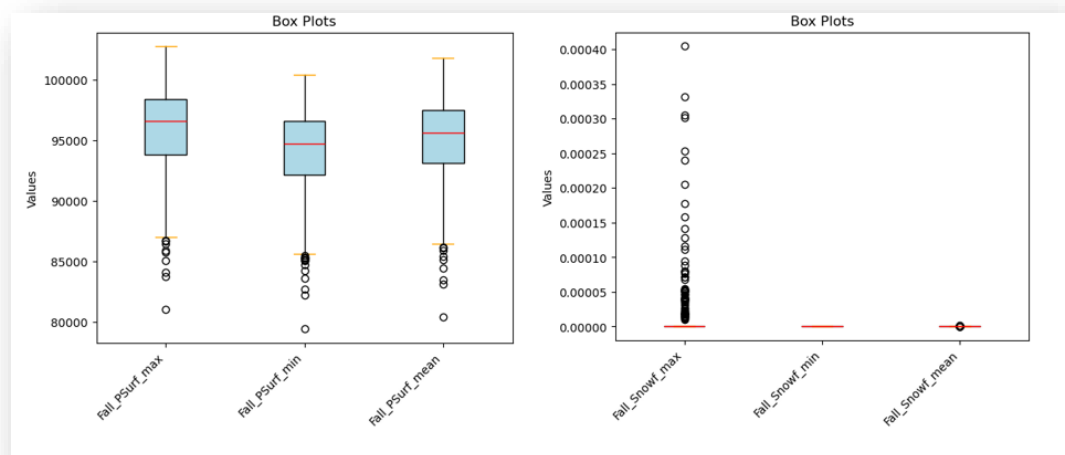


Figure 4 : Box plots des observations PSurf et snowf pour la saison d'automne

Analyse des Box plots de PSurf (pression de surface) :

- Valeurs générales : La majorité des valeurs (environ 75%) se situent entre 90 000 et 100 000, correspondant à des pressions de surface typiques pour la saison d'automne.
- Les valeurs inférieures à 90 000, identifiées comme des outliers, représentent environ 5 à 10% des données. Ces anomalies pourraient refléter des phénomènes spécifiques comme des systèmes dépressionnaires ou des régions géographiques particulières.

- Les médianes sont proches les unes des autres pour les trois statistiques (max, min, mean), oscillant entre **95,000** et **97,000** hPa, ce qui reflète une pression relativement stable.
- Les données sont modérément dispersées. L'étendue interquartile (IQR) est relativement similaire pour toutes les statistiques.

Analyse des Box plots de Snowf (précipitations neigeuses) :

- Valeurs générales : Les précipitations neigeuses de la saison d'automne (Fall_Snowf) sont **extrêmement faibles**, avec des médianes égales à 0 pour toutes les statistiques (max, min, mean).
- Quelques outliers sont visibles pour Fall_Snowf_max, dépassant **0.00035** m/jour, probablement liés à des épisodes isolés de chutes de neige.
- La dispersion est faible, surtout pour les valeurs minimales et moyennes, qui restent constamment à 0. Cette absence de variabilité confirme que les chutes de neige en automne sont rares ou limitées à des zones spécifiques (ex. : hautes altitudes).

c. Histogrammes

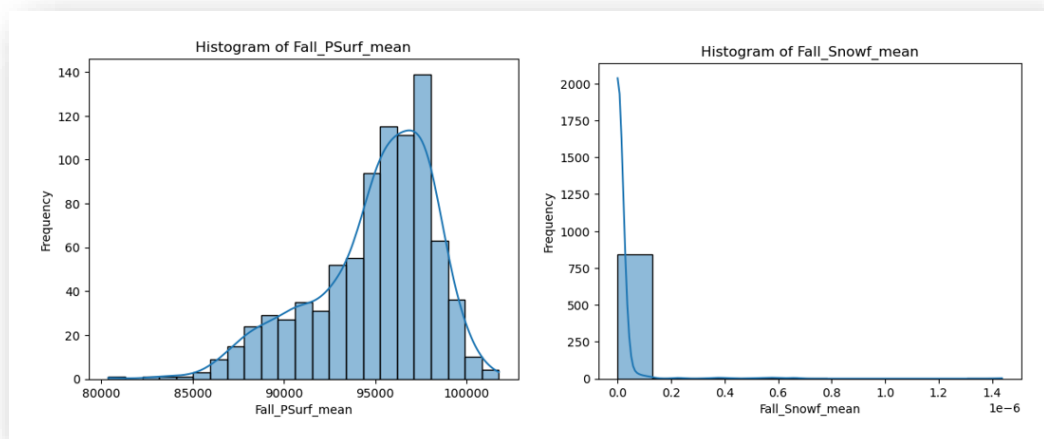


Figure 5 : Histogrammes des observations PSurf et Snowf pour la saison d'automne

Analyse de l'histogramme de PSurf :

- Distribution générale : Les valeurs moyennes de pression de surface suivent une **distribution presque normale** (presque symétriques), légèrement biaisée à droite.
- La plage des valeurs s'étend de 85,000 à 100,000, et la majorité des valeurs se situent entre 93,000 et 98,000.
- Mode et fréquence : Le pic (mode) se trouve autour de 95,000 unités, regroupant environ 140 observations, indiquant une pression moyenne stable dans la plupart des régions durant l'automne.

- Cette distribution montre une **variabilité modérée**, avec quelques zones ayant des pressions moyennes inférieures à **90,000** ou supérieures à **98,000**, mais celles-ci sont rares. Ces valeurs extrêmes pourraient correspondre à des régions spécifiques ou à des conditions climatiques exceptionnelles.

Analyse de l'histogramme de Snowf :

- Distribution générale : La variable Snowf moyenne est **fortement asymétrique**, avec un biais marqué vers la gauche.
- La majorité des valeurs sont **proches de 0**, regroupant plus de **2,000 observations** (soit environ 95% des données).
- La distribution présente une queue droite s'étendant jusqu'à 1.4×10^{-6} , reflétant des épisodes isolés de neige.
- La distribution confirme que la neige est **quasiment absente** en automne dans la majorité des régions. Les très faibles valeurs non nulles pourraient correspondre à des altitudes élevées où des chutes de neige isolées sont possibles.

d. Scatter & correlation

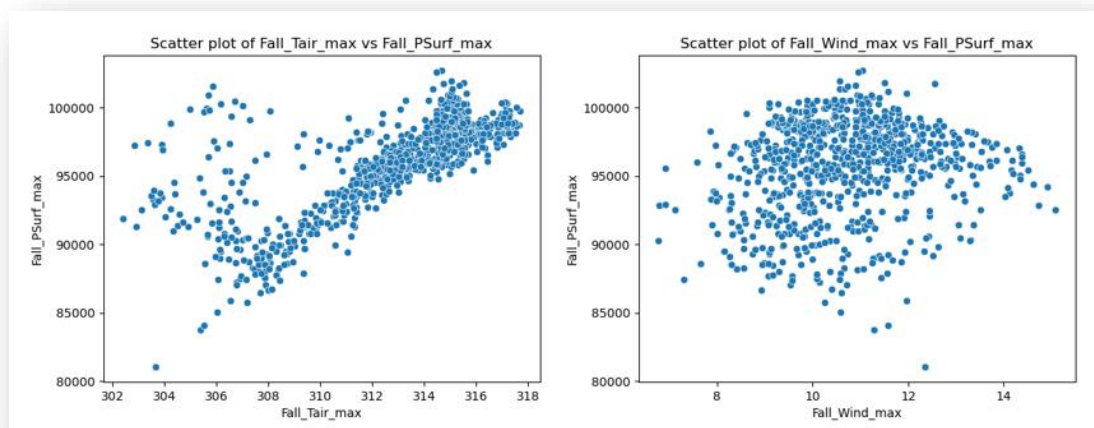


Figure 6 : Scatter Plot des observations (PSurf, Tair) et (PSurf, Wind) pour la saison d'automne

Analyse du scatter plot entre PSurf et Tair :

- Le degré de Corrélation entre Fall_Tair_max et Fall_PSurf_max est de **0.7355**.
- La corrélation indique une relation positive forte. Lorsque la température maximale augmente, la pression maximale augmente également.
- Cette relation est typique dans des conditions stables, où une augmentation de la température correspond souvent à une masse d'air plus chaude, entraînant une pression plus élevée.

Analyse du scatter plot entre PSurf et Wind :

- Le degré de Corrélation entre Fall_PSurf_max et Fall_Wind_max est égal à **0.1508**, révélant une relation très faible ou inexistante.
- Les points sont uniformément dispersés, montrant que la pression de surface reste stable indépendamment des variations de la vitesse du vent.
- Cela suggère que la vitesse du vent n'influence pas significativement la pression de surface, ce qui est cohérent avec les interactions atmosphériques typiques.

II. Dataset Sol

a. Valeurs uniques et manquantes :

Attributs	Valeurs manquantes	Valeurs uniques
Sand % topsoil	0	97
Sand % subsoil	0	96
Silt % topsoil	0	98
...
C/N topsoil	0	56
C/N subsoil	0	65
geometry	0	295

Le dataset Soil ne contient pas de valeurs manquantes.

b. Box Plots

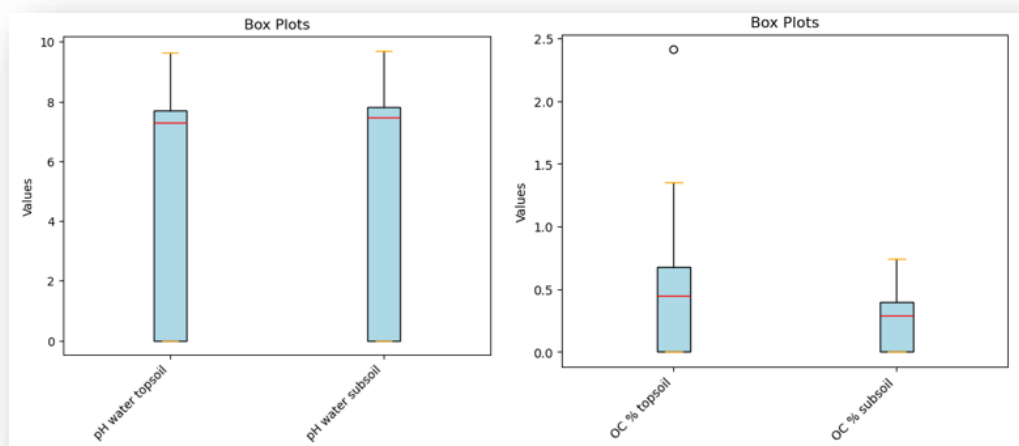


Figure 7 : Box plots du niveau du pH de l'eau et du pourcentage de carbone organique

Analyse des Box plots du niveau du pH de l'eau :

- Valeurs générales : Les distributions des valeurs de pH pour le **topsoil** et le **subsoil** sont **très similaires**, avec des médianes proches de 7, indiquant une alcalinité modérée.
- Il n'y a pas de points aberrants visibles, suggérant une distribution assez homogène des données.

Analyse des Box plots du pourcentage de carbone organique :

TopSoil :

- Valeurs générales : La médiane est proche de 0,5 %, indiquant une teneur relativement faible en carbone organique.
- Une valeur aberrante est visible au-dessus de **2,5 %**, signalant une observation avec une teneur nettement plus élevée, potentiellement liée à des zones riches en matière organique ou des pratiques agricoles spécifiques.

Subsoil :

- Valeurs générales : La médiane est encore plus basse, autour de 0,2 %, indiquant que la teneur en carbone organique diminue généralement avec la profondeur.
- Les valeurs extrêmes sont plus rapprochées et aucune aberration n'est visible.

c. Histogrammes

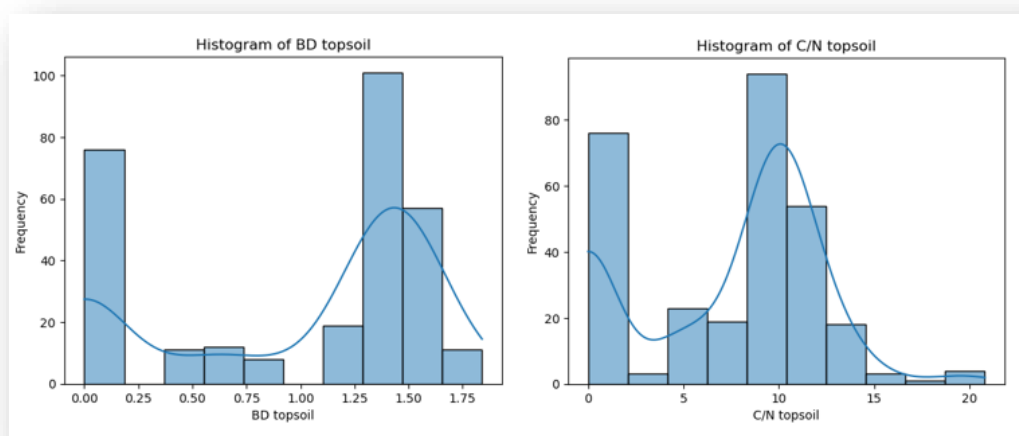


Figure 8 : Histogrammes de la Densité apparente (Bulk Density) et du Ratio carbone/azote

Analyse de l'histogramme de la Densité apparente (Bulk Density) :

- Distribution générale : Les valeurs se concentrent principalement entre **1,25 et 1,5 g/cm³**, correspondant à des sols compacts ou pauvres en matière organique.
- La densité apparente élevée (1,25-1,5) est typique des sols compacts ou pauvres en matière organique.
- Les faibles densités (< 0,5) sont probablement associées à des sols riches en matière organique ou peu compactés.

Analyse de l'histogramme du Ratio carbone/azote :

- La distribution est **asymétrique**, avec une forte concentration autour de **10**. La courbe de densité est **unimodale**, montrant un pic principal à cette valeur.
- Quelques valeurs extrêmes (supérieures à **15**) sont présentes, pouvant indiquer des données rares ou des outliers.
- Un rapport C/N de **10 à 12** est typique des sols cultivables, indiquant un bon équilibre entre la matière organique et les nutriments disponibles. Les valeurs très faibles ou élevées pourraient refléter des conditions de sol défavorables, comme une faible activité biologique ou une accumulation excessive de carbone.

d. Scatter & correlation

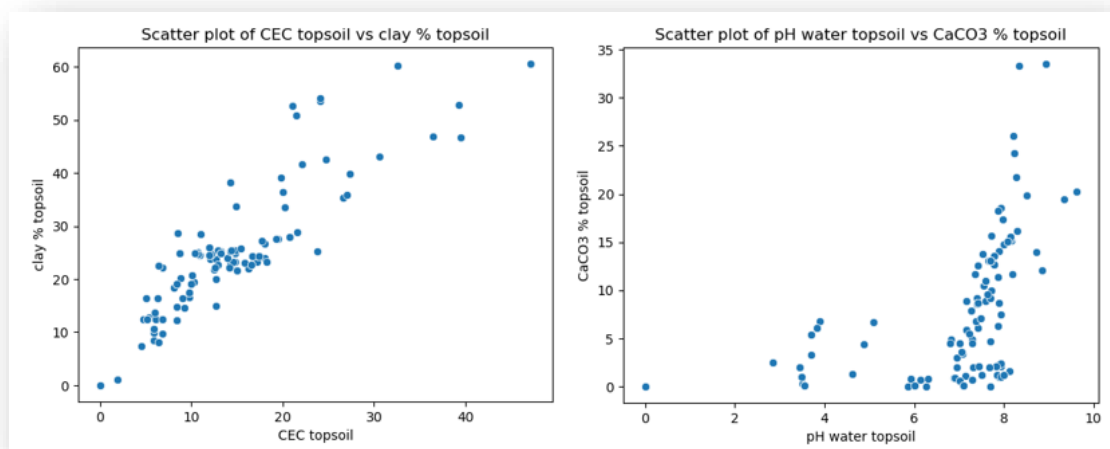


Figure 9 : Scatter Plot des attributs (CEC, Clay %) et (CaCo3 %, PH water) du dataset Sol

Analyse du scatter plot entre Clay % topsoil et CEC topsoil :

- Le degré de Corrélation entre Clay % topsoil et CEC topsoil est de **0.9170**, indiquant une relation positive très forte.
- À mesure que le pourcentage d'argile augmente, la CEC augmente également, reflétant l'influence de l'argile sur la rétention des nutriments dans le sol.

- L'argile, avec sa structure fine et sa surface chargée, contribue de manière significative à la fertilité du sol en augmentant la CEC. Les sols riches en argile sont souvent plus fertiles et adaptés à l'agriculture.

Analyse du scatter plot entre CaCO_3 topsoil et pH water topsoil :

- Le degré de Corrélation entre CaCO_3 % topsoil et pH water topsoil est de **0.6717**, indiquant une relation positive modérée à forte.
- Une augmentation du pourcentage de CaCO_3 est associée à une augmentation du pH, indiquant une alcalinité accrue dans les sols riches en carbonate de calcium.

III. Les deux dataset sol et climat fusionné

a. Valeurs uniques et manquantes :

Attributs	Valeurs manquantes	Valeurs uniques
lat	0	36
Lon	0	41
Fall_PSurf_max	0	856
...
BD subsoil	2	36
C/N topsoil	2	44
C/N subsoil	2	44

Le dataset contient 2 instances avec les attributs de sol à nul.

b. Scatter & correlation

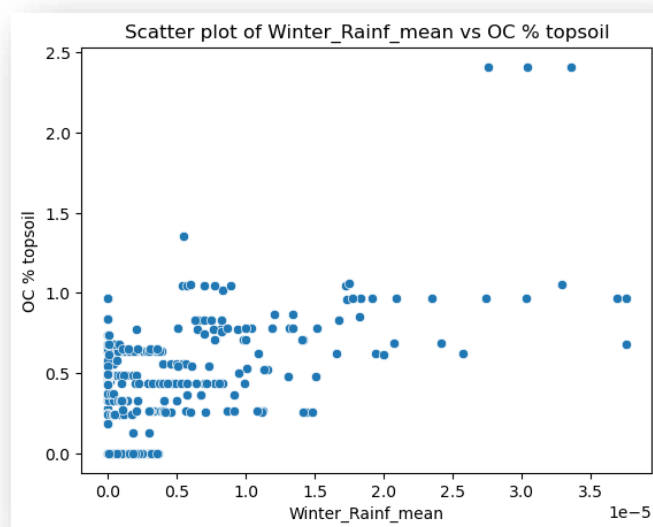


Figure 10 : Scatter Plot des attributs Fall_Rainf_mean et OC % topsoil du dataset fusionné climat-sol

Analyse du scatter plot entre Fall_Rainf_mean et OC % topsoil :

- Le coefficient de corrélation est de **0,4721**, indiquant une **relation modérément positive** entre les précipitations automnales moyennes et le pourcentage de carbone organique dans le sol de surface.
- Cela indique que des précipitations automnales plus importantes ont tendance à être associées à des niveaux plus élevés de carbone organique dans le sol.
- Cela pourrait refléter l'importance de l'humidité du sol dans la décomposition de la matière organique ou son intégration dans le sol au cours de l'automne.

Conclusion

Le projet présenté dans ce rapport avait pour objectif d'explorer et d'analyser les interactions entre les propriétés des sols et les conditions climatiques en Algérie en s'appuyant sur des techniques avancées de data mining. Grâce à l'utilisation de données géospatiales et climatiques, combinée à un traitement rigoureux incluant la réduction des données, la gestion des valeurs manquantes et des outliers, ainsi que des analyses approfondies, plusieurs résultats significatifs ont été obtenus.

Des analyses descriptives et exploratoires ont été menées, incluant des visualisations, des histogrammes, des scatter plots et des études corrélatives, permettant de dégager des tendances clés.

Ces résultats ouvrent des perspectives prometteuses pour l'agriculture durable et la gestion environnementale en Algérie. Ils permettent d'identifier des zones prioritaires pour des pratiques agricoles adaptées, de mieux comprendre l'impact des conditions climatiques sur les sols, et de formuler des recommandations stratégiques pertinentes à destination des décideurs.

En conclusion, ce travail constitue une base solide pour des recherches futures. Il pourra être enrichi par le développement de modèles prédictifs et l'intégration de nouvelles variables environnementales, afin de renforcer encore la compréhension des interactions sols-climat et de soutenir les initiatives en faveur de la durabilité agricole et environnementale en Algérie.

Références :

- [1] <https://support.esri.com/fr-fr/gis-dictionary/prj#:~:text=Le%20fichier%20PRJ%20contient%20les,1984%20UTM%20zone%2015%20nord%20%C2%BB>.
- [2] <https://www.lifewire.com/dbf-file-4144695>
- [3] <https://www.zwsoft.com/support/zwcad-base-faq/550#:~:text=shx%20file%20is%20a%20kind,and%20Zwcad%20and%20so%20on>.
- [4] <https://www.zwsoft.com/support/zwcad-base-faq/550#:~:text=shx%20file%20is%20a%20kind,and%20Zwcad%20and%20so%20on>.
- [5] <https://openknowledge.fao.org/server/api/core/bitstreams/149f1562-bf6a-439f-9d3a-eb93940f39cf/content> (page 14)
- [6] <https://cds.climate.copernicus.eu/datasets/derived-near-surface-meteorological-variables?tab=overview>
- [7] <https://pandas.pydata.org/>
- [8] <https://dbfread.readthedocs.io/en/latest/>
- [9] <https://pyproj4.github.io/pyproj/stable/api/proj.html>
- [10] https://docs.digitalearthfrance.org/en/latest/sandbox/notebooks/Beginners_guide/07_Intro_to_xarray.html
- [11] https://geopandas.org/en/stable/getting_started/introduction.html
- [12] <https://numpy.org/>
- [13] <https://scikit-learn.org/>
- [14] <https://matplotlib.org/>
- [15] <https://shapely.readthedocs.io/en/stable/>